ENHANCEMENT TO THE DISCRIMINATORY POWER OF STR TYPING THROUGH THE USE OF HAPLOTYPES

By

Catharine Worthen

Bachelor of Science in Biology

University of West Florida

Pensacola, Florida

2008

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May 2011

ENHANCEMENT TO THE DISCRIMINATORY
POWER OF STR TYPING THROUGH THE USE
OF HAPLOTYPES

Thesis Approved:

Dr. Robert Allen

Thesis Adviser

Dr. David Wallace

Committee Member

Dr. Jarrad Wagner

Committee Member

Jane Pritchard

Committee Member

Dr. Mark E. Payton

Dean of the Graduate College

ACKNOWLEDGMENTS

First I would like to thank my advisor, Dr. Robert Allen, for his guidance and support during this project. This project would also not have been possible without my committee, Dr. David Wallace, Dr. Jarrad Wagner, and Ms. Jane Pritchard for all their help that went into this project, especially the writing and statistic portions. The faculty and staff of the Forensic Science Department, Penelope Carr, Cathy Newsome, and Phylis Emery have helped me push through the rough patches.

I would like to give a special thank you to Dr. Jun Fu, for all her guidance and assistance in the DNA laboratory, and Dr. George Maha, who allowed me use his data for my project. I would also like to give a special thank you to Trang X. Pham, without her my samples would never have been labeled.

Lastly, I would like to thank my family for all their love and support. My parents and brother, Roger, who never let me give up, always believed in me, and made me the person I am today. My friends, Brandi Wahl and Victoria Comer, who always pushed me to be the best I could, thank you for your support. I would like to dedicate this thesis to my parents, Rick and Tina Worthen, who taught me that life is full of choices and you will never know what you’re capable of until you try.

TABLE OF CONTENTS

Topic Page

LIST OF TABLES

LIST OF FIGURES

Chapter I. Introduction

The California Department of Justice (2010) currently houses unidentified remains from over 2,100 individuals, some dating as far back as 1972. The backlog of unidentified human remains in California and the rest of the United States is small in comparison to the number of unidentified remains in other countries around the world. For example, in Colombia, it has been estimated that unidentified remains from about 30,000 individuals await identification and the situation in Colombia exists in many countries in Central and South America (Garcia, Martinez, Stephenson, Crews, & Peccerelli, 2009; Welsh, 2010). Identification of human remains can be achieved using DNA typing procedures. DNA is a very stable molecule and can be successfully extracted from the skeletal remains allowing short tandem repeat (STR) profiles, single nucleotide polymorphisms (SNPs), or mitochondrial DNA sequences (mtDNA) to be obtained and compared with those from reference samples representing surviving family members. While the success rate for identification is high when first order family relationships are investigated using DNA typing (e.g., parent:child or full siblings), surviving family members who are first order relatives of the deceased are often not available as a source of DNA. In such cases, to achieve a compelling result regarding the identity of the remains, either additional or specialized genetic markers must be tested to raise the discriminatory power of the test battery, or modified test methods must be used.

DNA results are interpreted by analyzing the alleles in the unidentified person's DNA profile and comparing those to alleles in a known relative's DNA profile. Allele frequencies, obtained from research done on the alleles and the number of times they appear in a population, are used to produce a likelihood ratio (LR) for each locus in the test battery which compares two hypotheses of relatedness: either the two individuals are related in some proposed way, or the two individuals are unrelated and random in the population and share alleles by chance. The likelihood ratios thus calculated can be multiplied together to create a cumulative ratio as long as the loci are independent of one another. Sometimes, though not often, LR values produced in an identification case are not convincingly high. This can occur when shared alleles are common in the population, or if the reference family member is not closely related to the missing person. In general, LR values produced in relationship testing cases in which the alleged relationship is second order (e.g., aunts/uncles or half-siblings) are lower for the same shared alleles than cases in which the alleged relationship is first order (e.g., parent/child or full siblings)

Approaches to raise the discriminatory power of a test battery (and hence raise the LR produced from testing) include adding DNA markers to the test battery. Another approach recently described by Lewis et al. (2010) is to use genetic markers linked together into haplotypes in place of single allele testing. A haplotype consists of two or more polymorphic genetic markers that are linked together, generally on the same chromosome. The population frequency of the haplotype is lower than the individual frequencies for the alleles and therefore the LR values produced from haplotype sharing within families is higher. Thus, the use of haplotypes may help alleviate the identification challenges when using distant family members as reference samples for the identification of remains.

Disasters happen everywhere; they can be natural (earthquakes, tsunamis) or unnatural (airplane crashes, acts of terrorism), and cause trouble for the teams who attempt to identify victims. Traditional methods of identification using dental records or through simple visual identification can fail, leaving questions and doubt as to the identity of victims. One such disaster was the Indonesian tsunami. Over 200,000 people lost their lives, their loved-ones, and/or their possessions, leaving very little for forensic practitioners to use to aid in the identification process (National Geographic News, 2005). The Swissair Flight 111 disaster is another example where traditional methods of identification failed (Leclair, Fregeau, Bowen, & Fourney, 2004). The bodies of the victims of the Swissair flight were badly fragmented and strewn about the crash site, which created problems for identifying tissue fragments in order to reunite the fragments into a single body.

Traditional methods of identification of victims include forensic odontology, anthropology, and fingerprints. The use of DNA as an identification tool began because these traditional identification methods could not always provide enough proof of identity. For forensic odontology to be useful, the alleged victim must have ante-mortem dental records that fit within the timeline when the person went missing and the remains were found. The same is true for fingerprints. There must be a set of known prints to match to the unknown prints. Also, for the most part, remains must be intact, which is very rare in mass fatalities (Graham, 2006). In the identification of remains from mass graves in Croatia and Bosnia-Herzegovina, standard techniques did not provide adequate support for identification of 30% of the victims (Alonso et al., 2001).

When major damage is done to bodies, the only way to identify the bodies is to perform DNA typing on body parts to re-associate the parts with one another (Graham, 2006). One of the first uses of DNA for identification of victims was in Waco, Texas in 1993. The Branch Davidians' compound caught on fire, killing everyone inside. An STR quadruplex was amplified and used to identify remains (Clayton et al., 1995). However, only 26 positive identifications could be made out of the 70 victims of the fire (Butler, 2005). Even so, without the DNA testing performed, all victims would have remained unidentified. Whole families were killed in the fire, leaving distant relatives as the only reference samples available for testing (Graham, 2006). This case first brought out the useful potential of DNA typing. Another champion case for DNA analysis in identifying mass disaster victims was the 1996 Spitsbergen crash. This disaster killed 141 people and badly fragmented the bodies (Olaisen, Sternersen, & Mevag, 1997). Olaisen et al. (1997) had an identification success rate of 98.6% for the victims by using STR typing and samples from at least one, or up to three references. These success stories paved the way for DNA analysis to be used with confidence in identifying victims of mass disasters and to new discoveries, such as new marker systems and better technology (Graham, 2006).

For all the power associated with DNA typing when first order family relationships are questioned, the technology can fail to produce a compelling result when more distant relatives must be used as the sole source of reference samples. In all of the cases mentioned above, parents or close relations were available for testing. When distant family members, such as aunts, cousins, or half-siblings are used as reference samples, the discriminatory power of DNA typing can be greatly reduced. The use of genetic marker haplotypes could help alleviate this problem because haplotypes are generally less common in the population than the alleles that compose them. Because haplotypes are less common than alleles, any sharing of the haplotype between remains and a reference sample will contribute more to the magnitude of the LR thereby increasing the discriminatory power of the test battery.

There is a lack of studies on the use of haplotypes for identification as haplotypes have typically been studied in association with genetically inherited diseases. A shift towards using haplotypes for identification is beginning and more research is being done in this area. Two haplotype-based genetic systems used extensively for identification are STR loci located on the Y chromosome and SNP type polymorphisms residing in mitochondrial DNA (Butler, 2005). Y-STR haplotypes can determine with high probability whether two men are related (Corach, Risso, Marino, Penacino, & Sala, 2001). Mitochondrial DNA can tell if individuals are related through their maternal line. In a novel investigation on the use of autosomal haplotypes, Lewis et al. (2010) analyzed haplotypes of genetic markers located on multiple chromosomes to confirm a familial relationship between the remains of a World War II pilot and a woman in Australia claiming to be his daughter. The pilot had died many years ago and thus distant family members had to be used as reference samples for DNA typing. Ultimately, the study of Lewis et al. (2010) demonstrated the power of haplotypes in providing compelling evidence of relatedness even when distant family members must be used as reference samples. More research on the use of haplotypes in such questioned relatedness scenarios would likely establish how generally useful genetic marker haplotypes can be. Methods already in use by relationship testing labs can be applied to develop haplotypes so no new methods (or additional loci) need to be developed.

DNA samples obtained from relationship testing cases submitted to the Human Identity Laboratory of Oklahoma State University were used for the study. In each case, the relationship between parents and child(ren) had been established with high probability using traditional DNA typing methods. The cases were randomized and given new case numbers to keep the families anonymous and the research was approved through the Institutional Review Board (IRB). The only information collected from each case was the self-identified racial background of the mother and father. The FFFL (F13A, FESFPS, F13B, and LPL) and Penta E STR loci were amplified from genomic DNA that had been extracted using DNA IQ, using regents supplied with the STR typing kits (available from Promega Corp,, Madison, WI). Amplified STR products were then analyzed using an ABI 310 genetic analyzer (Applied Biosystems, Foster City, CA) in most cases. In some cases an ABI 3130XL (Applied Biosystems, Foster City, CA) was used for DNA analysis. One of the loci in the FFFL quadriplex, FESFPS, is linked to the Penta E marker (AABB, 2010), both are located on chromosome 15 (National Institute of

Standards and Technology (NIST), 2010). Thus the individual LR values produced using FESFPS and Penta E cannot be multiplied together to produce a cumulative result in relationship tests because the loci are not independent (AABB, 2010). However, alleles at the FESFPS and Penta E loci represent a haplotype, and if a haplotype database were available, the frequency of a particular haplotype could be used in LR calculations, possibly with an enhancement in the discriminatory power of the overall STR typing performed in the case. This was the rationale underlying this study.

Ultimately, the importance of researching haplotypes for family relationship testing is to try to increase the discriminatory power of a test battery and consequently the level of certainty of the result produced when a suspected family relationship is subjected to DNA testing. Specific goals of this study were to develop FESFPS-Penta E haplotype frequencies for two major ethnic groups, Caucasians and Blacks, and to determine the degree of linkage of the two markers through counting the number of recombinations occurring between FESFPS and Penta E alleles in families with multiple children. The questioned enhancement to the discriminatory power of the overall test battery (consisting of 20 loci, including FESFPS and Penta E) was investigated through comparison of final LR values produced with the 19 locus panel (using either FESFPS or Penta E, whichever gave a higher LR) versus the 20 locus panel using the frequency of the FESFPS-Penta E haplotype to calculate a LR incorporated into the final result by using the product rule. The higher LR value is used to obtain the highest probability.

Results of this study have shown that within the Caucasian and Black ethnic groups, there are at least 102 different FESFPS-Penta E haplotypes with the most common haplotype existing in the Caucasian population with frequency of 0.107. In addition, it appears that there is more haplotype diversity within the Black ethnic group wherein 45 haplotypes seen in Blacks have not been seen in Caucasians but only 6 Caucasian haplotypes have not been seen in Blacks. Results have also shown that the use of FESFPS-Penta E haplotypes increases the statistical power of the STR test battery approximately 7.05 fold in paternal multation calculations. There was also a 2.43 fold increase in paternal relationship tests in Blacks and a 1.84 fold increase in Caucasian parentage calculations.

Chapter II. Review of Literature

2.1. STR Typing for remains identification

DNA analysis for the identification of human remains has come a long way since its first uses in the Spitsbergen incident or Waco, Texas (Butler, 2005). STR typing has become the gold standard in forensic and family relationship testing because the technology has proven reliable and extremely discriminatory. Moreover STR typing, as opposed to RFLP (Restriction Fragment Length Polymorphism) analysis, can be automated. DNA typing has become the primary technique used to identify disaster victims with traditional methods, like odontology or fingerprints, now used to confirm the identification, whereas before the mid-1990's DNA was used as a last resort (Graham, 2006). Commercial kits, made by Applied Biosystems (Foster City, CA) and Promega Corp (Madison, WI), are readily available; these kits contain all of the reagents needed to produce a DNA profile essentially unique to an individual, and the kits are continually being modified to include more genetic systems, further increasing the overall discriminatory power of the test battery.

2.1.1. Family Relatedness Testing

Genetic testing of questioned family relationships began with serological testing methods (blood and tissue typing) and has continued to evolve with DNA typing technology. Most testing performed utilizes first order relationships, such as parent and child as reference samples for comparison to remains. Sometimes, however, when parents are not available for testing, other family members can be used to establish or refute the suspected relationship. Relationships other than parent-child, present a greater challenge to the discriminatory power of short tandem repeat (STR) typing.

STR genetic markers used in relationship and forensic DNA testing are inherited independently of one another and are considered to be in linkage equilibrium. Haplotypes, on the other hand, consist of genetic markers that are inherited together either due to being physically linked on a chromosome or because when inherited together, they confer a selective advantage to the host. Markers that are not inherited independently of one another within a population are linked, and exist together as haplotypes. Since physically linked markers are inherited together on a single chromosome, their inheritance through multiple generations can be very stable.

Family relatedness testing is straightforward when both parents are available for testing. Often in forensics and sometimes in relationship testing, one parent is considered unquestioned (normally the mother) and the other is alleged. For example, in a paternity case, the mother is assumed to be the true mother of the child; thus the alleles she transmits can be subtracted from the child's profile. The remaining alleles are attributed to the father. A likelihood ratio (LR) calculation is performed, which incorporates the population frequencies attributed to alleles detected in the alleged father into a number that reflects the statistical weight supporting the claim of paternity for the child (Lee, Lee, Han, & Hwang, 2000). Paternity testing becomes more complicated when the mother of the child is not available for testing. One consequence of the lack of a known parent is a possible reduction in the discriminatory power of the test battery with a concomitant reduction in the magnitude of the likelihood ratio produced for an alleged parent who cannot be excluded. The mother's profile can no longer be used to define the paternal obligate alleles in the child's profile, thus introducing ambiguity into the analysis of the DNA test results. Lee et al. (2000) analyzed motherless paternity cases in Korea and concluded that the "mean exclusion chance in trio cases (with a known parent) is higher than that produced when the mother is not tested (i.e. motherless cases)" and that there is a significant difference between the two exclusion calculations. Lee et al. (2000) also concluded that in motherless or deficient cases, likelihood ratios are lower, making compelling probabilities of paternity more difficult to produce.

2.1.2. Statistical Analysis of Family Relatedness Testing

The International Society of Forensic Genetics, or ISFG, (Gjertson et al., 2007) recommends that all paternity calculations are done using the likelihood ratio method. This method compares two alternate hypotheses to come to a conclusion of parentage. The first hypothesis in a paternity case for example, H_0, is the probability that the alleged father is the true father of the child. The second hypothesis, H_1, is the probability that the alleged father in not the true father. In such cases of exclusion of the tested man, the true father of the child is someone unrelated to the alleged father and random in the population. A ratio of the two probabilities produces a value reflecting how likely H_0 is relative to H_1. The ratio calculated for each locus can be combined into a cumulative value using the product rule (i.e., multiplying the individual LR values together) because the autosomal STR loci widely tested are independent of one another and in linkage equilibrium (Gjertson, et al., 2007).

It is also possible to use haplotypes of linked markers to produce a likelihood ratio. In cases where haplotypes are used, such as Y-STR and mitochondrial DNA typing, calculated frequencies for those haplotypes must be used in place of allele frequencies and all frequencies must be validated before use (Gjertson, et al., 2007). Haplotype frequency databases exist for Y-STR and mtDNA markers. However for newly developed haplotype markers, validated databases must be created.

2.1.3. Mutation and Recombination

One complication that occurs in relationship analysis is mutation (Allen, 2010; Calafell, Shuster, Speed, Kidd, & Kidd, 1998). For STR loci, mutations in which repeats are either added to an allele or deleted from it occur during meiosis with an average frequency of about one in every 500 cases (AABB, 2010; Allen, 2010; Myers, Bottolo, Freeman, Mcvean, & Donnelly, 2005). Thus mutant STR alleles can become incorporated into gametes that differ in repeat number from those in the parental reference sample. If this gamete contributes to the conception of a child, non-parentage will be suggested through STR testing for one who is in fact the true parent. Similarly, when using haplotypes, recombination between the linked markers can occur and create a haplotype in an offspring that differs from the haplotype of the true parent. Recombination occurs when two homologous chromosomes cross-over and exchange genetic information during meiosis. The closer together the two markers are on the chromosome, the less likely it is that recombination will occur, whereas if the markers are far apart, recombination is more likely to occur. The probability of recombination is expressed in centiMorgans (cM). Since recombination varies throughout the genome, there is no standard process for converting centiMorgans of recombination frequency into basepairs of DNA length (Fearnhead & Donnelly, 2002). However, Fearnhead and Donnelly (2002) contend that one cM corresponds to about 1.2 megabases of human DNA length. Rates of recombination are lower in areas rich in either TA or GC repeats (Myers, et al., 2005).

When recombination occurs within a haplotype, the haplotype inherited by the child has a chance to differ from the parent’s true haplotype. Recombination rates are generally low because the linked markers are typically close to one another on the chromosome The lower the recombination rate between linked markers, the lower the general haplotype diversity exhibited within the population (Hellmann, Ebersberger, Ptak, Paabo, & Przewroski, 2003). A recombination rate of less than 50% indicates that two or more genetic loci are linked whereas markers on the same chromosome but with a recombination rate of 50% or higher are considered statistically unlinked and therefore in equilibrium.

2.2. The Need for More effective Methods of Identification

2.2.1. Identification of Victims from Genocide and Wars

Identification of victims of war or genocide is important for many cultures and peoples (Huffine, Crews, & Davoren, 2007). In 1992, Finland undertook a project to identify Finnish soldiers who were considered Missing In Action (MIA) or fallen and left on the battlefield in World War II, even though it had been over 60 years since the end of the war. Palo et al. (2007) found that all relatives contacted and asked to donate DNA samples for identification purposes did submit samples. Identification brings closure to families. The identification of the fallen will let families know what happened to their loved ones and let the families give the victims proper burials. It can also lead to criminal prosecution for crimes against humanity, as is the case for the former Yugoslavia and Rwanda.

Genocide is defined by the 1948 *Convention on Prevention and Punishment of the Crime of Genocide* (CPPCG) as "intent to destroy , in whole or part, a national, ethnical, racial, or religious group" (United Nations, 2008) and was created in response to the Holocaust. Since the CPPCG's inception, it has held the Nuremburg trials and tribunals for Rwanda and the former Yugoslavia. More recently, mass graves have been discovered in Iraq that are being investigated with the possibility of opening criminal investigations and trials in that area as well.

The conflict in former Yugoslavia produced a politically charged climate as the republics fought for independence from one another and the number of missing persons kept climbing (ICTY-TPIY, 2011). Serbian aggression against Croatia in 1991 and in 1992 for Bosnia-Herzegovina left over 11,000 people missing and many more displaced (Andelinovic et al., 2005). In 1995, Srebrenica, a Bosnian town and an U.N. declared “safe area”, was attacked by the Serbian army and became the second largest systematic killing in Europe, surpassed only by the Holocaust during World War II (Huffine, et al., 2007). An effort began to identify the remains found in 135 mass graves representing the victims killed during this aggression or genocide.

The identification of the victims from Srebrenica and surrounding areas served two purposes. The first was to bring closure to family members of the fallen. The second, and perhaps most novel purpose, was to help establish the accountability of those responsible for committing genocide (Huffine, et al., 2007). The Serbian army had steadfastly denied any wrongdoing in Srebrenica. As more mass graves of genocide victims were unearthed, the evidence that genocide had indeed taken place became overwhelming. DNA analysis was used on thousands of bodies that were unearthed in mass graves in the former Yugoslavia. The analysis later determined that some bodies, especially those from the Srebrenica massacre, had been exhumed, dismembered, and re-buried in one or more mass graves that were spread across the country (Huffine, et al., 2007). The identification of remains linked to victims from Srebrenica forced Serbia to recognize its role in the genocide of Bosnian Muslims (Weaver, 2003).

During the conflicts in the former Yugoslavia, Rwanda was fighting a bloody civil war ignited by the death of the Hutu president (Geltman & Stover, 1997). In 1994, explosive fighting between the Hutu army and Tutsi guerillas led to massacres of both Tutsis and moderate Hutus, many of whom were refugees trying to flee the conflict. The largest problem faced by the Rwandan genocide was trying to reunite orphaned or lost children with family members. 94,000 children were registered as unclaimed and only 10,500 were successfully reunited with family (Geltman & Stover, 1997). Parents had also registered and, if DNA testing had been available the number of reunited families could possibly have increased.

War causes problems when identifying victims. After a thirty-six year civil war, the Guatemalan Forensic Department started the daunting task of identifying the victims (Garcia, et al., 2009). This was complicated by the fact that many family members were either missing, dead, or unavailable for testing. The Department analysts used statistical software and family trees to discern relationships between the victims, the missing, and people available for testing (Garcia, et al., 2009).

Wars that have been long over can still have unidentified remains that pose special identification problems because of the age of the remains and the unavailability of reference samples from surviving family members. In 1992 in Finland, after a growing public outcry, the decision was made to attempt the identification of soldiers from World War II (Palo, et al., 2007). Sixty years after the war, mitochondrial DNA (mtDNA) samples were collected from surviving family members for comparison to mtDNA recovered from the bones of the missing and unidentified. Palo et al. (2007) observed that all the suspected relatives donated samples leading them to conclude that identifying

MIA soldiers is important no matter the length of the passing time since the disappearances. Even though remains could only be linked by maternal lineage, the response of the relatives of the unidentified reinforced the importance of identifying the fallen.

While trying to identify victims of the conflicts in Croatia and Bosnia and Herzegovina, forensic teams came across two mass graves that witnesses said were created around 1945. In the effort to identify the remains of those interred in the graves, Marjoanovic et al. (2007) were able to obtain DNA profiles with a range of 13 to 16 detectable loci out of the 16 loci used in the Power Plex 16 kit (Promega Corp., Madison, WI) for all 27 samples. War can cause problems when trying to positively identify remains. Many times when traditional identification methods cannot be used for identification, DNA becomes the last resort.

When the United States invaded Iraq in 2003, many mass graves were uncovered, most of which were assumed to be filled with rebels from the 1991 uprising against Saddam Hussein (Dareini, 2003). It was speculated that skeletal remains from over 300,000 bodies were buried in mass graves spread all over the Iraqi desert (Roberts, 2005). Anthropological teams from the United States were dispatched to the area to recover the remains after the mass graves were found (Burns, 2006). Thus, Iraq has become a large source of unidentified remains. Wars and genocide are two types of unnatural disasters that require identification of remains. Other types of unnatural disasters and natural disasters can be just as devastating.

2.2.2. Airplane Crashes and Natural Disasters

Another leading cause for victims remaining unidentified is disasters not caused by wars and genocide. These disasters can be natural or man-made. Special teams in the United States that consist of special forensic disciplines, known as Disaster Mortuary Operational Response Teams (DMORT), are dispatched to help recover the remains of victims (Alonso et al., 2005). In mass disasters, like the Swissair Flight 111 crash, the bodies can be fragmented, burned, or otherwise unidentifiable (Leclair, et al., 2004). The Swissair disaster posed a particular problem because the airplane crashed into the ocean, leaving remains 70 meters under the surface (Leclair, et al., 2004). The methods used to identify victims of the crash were dental records, medical records, and fingerprints. These traditional methods used to identify victims are not always available and records may not be up to date. For example, when dental records are used to identify a victim, the identification team must find the most recent set of dental records for the remains being analyzed. Finding the correct records can be a daunting task, and they are not always available.

Identification through DNA may also pose challenges in mass disasters in terms of recovering DNA from the tissue and other samples and also from appropriate surviving family members. Destruction of the body and dispersal of remains, degradation of DNA, the number of victims, and availability of samples from closely related family members can all present challenges to the use of DNA for identification (Alonso, et al., 2005). In most cases, DNA can be obtained from a multitude of sources: such as muscle, skin, bone, and blood, most of which can be found in more than one part of a victim’s body, which still makes DNA the first choice for identification because newer and better methods have been developed to extract nuclear DNA from tissues not rich in DNA (like bone for example). Experts in this area recommend obtaining samples from tissues that are least affected by the disaster (Alonso, et al., 2005).

In mass fatality disasters, sometimes whole families perish, contributing to identification challenges, as closely related reference samples become harder to obtain (Leclair, et al., 2004). In the Swissair flight disaster, reference DNA samples were also recovered from personal items, such as toothbrushes and hair brushes, which were undisputed as belonging to the victim. Alonso et al. (2005) note that personal items can be destroyed or altered, as encountered with the tsunami disaster in Indonesia. For the Swissair crash, even if the DNA obtained from a personal item resulted in a positive match with the victim, DNA from a second personal item was needed to confirm the match to the victim and make a positive identification (Alonso, et al., 2005) . Leclair et al. (2004) pointed out that parentage analysis is still the most preferred method of identifying victims when the parents are available. Thus, disasters that kill whole families pose a special challenge to DNA used for identification of remains.

2.2.3. Human remains discovered in clandestine graves

Unidentified remains not resulting from wars or mass disasters can be found all around the world. In Mexico in August of 2010, two mass graves containing over 50 bodies each were discovered, one near Monterrey and the other near the town of Taxco (Hawley, 2010). The bodies in the graves are thought to be victims of Mexico’s drug cartels. The Mexican government is having difficulties identifying all the remains. Thus far, nine sets of the remains from the mass graves have been positively identified (Hawley, 2010). In another area of the world, a Nazi mass grave was found located under an Austrian military installation that many believe is from World War II (Associated Press, 2010). Identification of these remains could prove challenging, since they are over 70 years old (Associated Press, 2010). Close family members who's DNA could aid in the identification of the remains are generally not available for various reasons.

Unidentified remains are not only found in Europe and Mexico, but a backlog of unidentified remains exists in the United States. In California, over 2,100 sets of remains, some dating as far back as 1972, exist (California Department of Justice, 2010). Applied Biosystems, in an effort to help reduce the backlog of unidentified remains, donated genetic analyzers and materials to California labs; however, the backlog has not been reduced by a significant amount (Applied Biosystems, 2003). Another state that has a backlog of cases is New York, which has a missing- persons database containing over 3,500 people (Caher, 2009). To try and help identify some of the unidentified remains, in 1999 the Doe Network has strived to help identify missing persons that could number among the unidentified. Currently, unidentified victims from the United States and other parts of the world are among those the Doe Network is trying to identify (Wahlstrom, 2001).

Integrating mass fatality identification within daily work in a forensic lab could help alleviate the backlog of unidentified remains. Budowle et al (2005) recommends that DNA labs use the commercial kits that are also used in routine forensic work for identification. Haplotypes consisting of Y-STR or X-STR markers produced using commercially available kits would be useful. The use of these two haplotypes, from the X and Y chromosomes, from widely used DNA typing kits would save time and money, being readily and widely available commercially and in routine use by forensic laboratories. The FFFL and Penta E kits are also commercially available and are easy to use and implement into casework, as their use is much the same as typing kits already in use in forensic and relationship testing laboratories.

2.3. Current Methods and Uses of Haplotypes for Relationship Testing

Haplotype systems useful for identification through family relatedness testing will be stably inherited within a pedigree because of a relatively low rate of recombination between linked markers. In addition, useful haplotype systems will consist of linked genetic markers that individually consist of multiple alleles that are evenly distributed through the population.

The frequencies of different haplotypes within a candidate system and the rate of recombination within the system can be determined through family studies. Haplotypes are established and counted by following the transmission of alleles from linked loci (i.e. haplotypes) within a known family consisting of mother, father, and child. The individual alleles for each of the loci known to be linked that are transmitted from parent to child thus constitute a haplotype (Lathrop, Lalouel, Julier, & Ott, 1984). Four distinct haplotypes can generally be discerned for linked genetic markers that are highly polymorphic in this approach, two from the mother and two from the father. Recombinations within haplotypes during meiosis are most often identified through similar family analysis, by examining families with multiple children in which a recombination event can be detected in one of the children where others harbor haplotypes that match the phase of markers seen in the parent. Phase is when the pattern of inheritance of alleles from parent to child is determined. An example of the process is summarized in Figure 1.

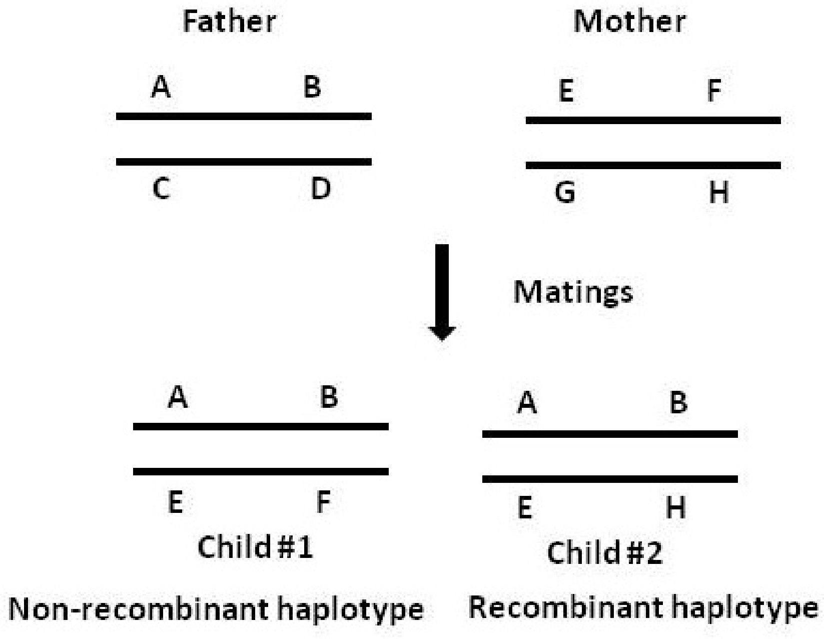

Figure 1. Recombination Haplotype Study
Figure 1 shows a family study used to identify the different haplotypes within a system with Child 1 having the true haplotype (based upon the shown haplotypes in the parents). Child 2 has a recombinant haplotype inherited from the Mother.

As was stated above, combining likelihood ratio calculations into a cumulative result depends upon all of the markers used in the calculation being independent of one another. Lathrop et al. (1984) recommend knowing the rate of recombination and the relative chromosomal location of the two markers to help determine if the markers are linked.

Haplotype markers have been used for determining family relatedness. STR loci residing on the Y chromosome are considered a haplotype because there is no recombination between them due to the fact there is only a single Y chromosome in males; that is to say there is no homologue with which to recombine during meiosis. Mitochondrial DNA is also considered a haplotype, as is the single X chromosome donated from the father to his daughter.

2.3.1. Y Chromosome and Mitochondrial DNA Haplotype Markers

The most extensively used of these haplotype systems is the Y chromosome. The Y chromosome is inherited through multiple generations within the paternal lineage and can be a good indicator of relatedness. A male child will demonstrate relatedness through his Y-STR haplotype to his father and any of the father’s male relatives because they will all share a common Y chromosome in the paternal lineage (Corach, Risso, Marino, Penacino, & Sala, 2001; Roewer et al., 2001). Y-STR typing is also useful when the mother is not available for testing. The alleles on the Y chromosome are tightly linked and therefore every male child in the lineage will exhibit the same alleles.

Several Y-STR haplotype databases have been created. The Europeans for example have created a database for Y chromosome haplotypes that is available for population research and statistical calculations in relatedness cases (Roewer, et al., 2001). Databases in the U.S. and other countries have also been created and are available on the web. The European Y chromosome database, YHRD, is available for use all over the world; Roewer et al. (2001) plans to expand the database with frequency data from populations all over the world.

Y chromosome DNA markers are useful in determining relatedness through the paternal lineage (Corach, et al., 2001) and are especially useful if an alleged father is not available for testing since another male relative can be substituted in his place to provide the Y-haplotype for the male lineage within the family. A drawback of Y chromosome markers however, is that all males in the lineage will have the same markers. Thus, a child can be shown to be a member of the male lineage within a family, but the exact relationship between the child and the untested alleged father remains unknown. Another drawback associated with Y-STRs is the reduced discriminatory power associated with the use of these genetic systems for family relatedness testing. However, with the growth of Y-STR databases the discriminatory power may rise, but it will never likely surpass the power of autosomal markers.

2.3.2. Forensic and other Uses of Autosomal Haplotypes

Linkage among markers and genes is not a new concept. The medical field uses linkage analysis to find genes that are associated with diseases (Barrett, Fry, Maller, &

Daly, 2005). Only recently have haplotypes found use in the world of forensics and relationship testing. In a presentation to the American Academy of Forensic Sciences (AAFS), Lewis et al. (2010) pioneered the use of linked autosomal STR markers in an attempt to determine whether a World War II soldier was the father of a woman in Australia. The woman had been told that this dead soldier was her father. The soldier died during the war and his remains were missing until 2002 when they were discovered in New Guinea. In this case, Y-STR typing was not useful because the person questioning the relation was a female. Also, mitochondrial DNA testing would not have yielded compelling results, as mtDNA is inherited from the mother (Lewis, et al., 2010). The soldier's bones did not yield enough genomic DNA for testing, so testing of distant family members was necessary to investigate paternity. Lewis et al. (2010) used 9 linked STR markers on different autosomal chromosomes to form rare haplotypes that could be compared within the surviving family members for comparison to the corresponding haplotypes harbored by the woman claiming the decedent was her father. Because haplotypes are less common than that of corresponding alleles from the individual genetic markers that compose them, the statistical power of the DNA testing is increased and, in this case, produced a compelling level of certainty that the soldier was indeed the father of the woman from Australia (Lewis, et al., 2010).

2.3.3. Genetic Diversity in Black and Caucasian Populations

Genetic diversity among populations originating from different continents has been a highly discussed topic (Bamshad, Wooding, Salisbury, & Stephens, 2004;

Calafell, et al., 1998; Hellmann, et al., 2003; Myers, et al., 2005). Research has suggested that diversity is greatest in African populations followed by European and then Asian populations.

In a study by Jorde et al. (1997), researchers found that there was a statistically significant elevation in the genetic diversity in African populations. In one study, private alleles, which were alleles only observed in one population but not in others, were discovered in a single population group (Calafell, et al., 1998). The private alleles described by Calafell et al. (1998) were most prevalent in African populations, followed by European populations. It was discovered that alleles found in populations on other continents, such as Asia, would also be found in African populations. This suggests that African populations have greater genetic diversity (Calafell, et al., 1998).

2.4. Why use FESFPS and Penta E?

The STR markers, FESFPS and Penta E, are routinely used in the Human Identity Laboratory at Oklahoma State University to provide extended DNA profile information in cases of questioned relationships. Both loci are located on chromosome 15 and are only six million base pairs apart which amounts to about five centiMorgans of recombination frequency distance (AABB, 2010). Penta E has a five nucleotide repeat, AAAGA , and is a large locus consisting of 26 different alleles (National Institute of Standards and Technology (NIST), 2010), whereas FESFPS is a four nucleotide repeat and is a much smaller locus with only nine alleles (National Institute of Standards and Technology (NIST), 2010). Because these two loci are likely linked, transmission of alleles within families is not random (i.e., for unlinked loci) and therefore the likelihood ratios produced for each locus cannot be multiplied together to calculate a cumulative LR value using the product rule. Currently, labs produce LR values for each of the two loci and incorporate the LR with the higher value for the final cumulative LR calculation.

Given the published increase in discriminatory power associated with the use of haplotype systems (Corach, et al., 2001; Lewis, et al., 2010; Marjanovic, et al., 2007; Palo, et al., 2007) and the need for more powerful test batteries for the worldwide problem of remains identification, it makes sense to explore the potential of a haplotype consisting of FESFPS and Penta E alleles to increase the discriminatory power of the STR test battery used by laboratories in questioned family relatedness cases. Based upon the experience of others, it is likely the use of haplotypes will reduce the number of instances in which inconclusive results are produced and thus increase the effectiveness of a test battery used for identification purposes.

In the study reported here, the linked FESFPS and Penta E STR markers were investigated as a haplotype system that could be useful for identification purposes. Haplotype frequencies in two major ethnic groups (Caucasians and Blacks) were produced and the magnitude of the use of haplotype frequencies in likelihood calculations was evaluated. Results showed that the FESFPS-Penta E haplotypes can increase overall discriminatory power of an existing STR test battery which will contribute to the effectiveness of the battery in cases of questioned family relatedness.

Chapter III. Methodology

3.1. Sample Selection and Preparation

To identify and count FESFPS-Penta E haplotypes in the different population groups, DNA samples from previously tested parentage cases in the archives of the Human Identity Laboratory (HIT) at Oklahoma State University Center for Health Sciences (OSU-CHS) in Tulsa, OK were used. Study protocols were approved by the IRB. Some of the cases consist of DNA from three family members, a mother, a child, and a father whereas in other cases, samples from multiple children and their parents were available for testing. All of the cases used were inclusions of the alleged father. DNA from cases was extracted by the Human Identity Lab using the DNA-IQ extraction method available as a kit from Promega Corp (Madison, WI). Samples were chosen from archived cases based upon the ethnic background self-identified by the father and mother of the child(ren). Both parents had to be of the same ethnic background to qualify for the study. The two ethnic backgrounds initially selected for the study were Caucasian, and Black. Other considerations for selecting samples were whether there were multiple children within a family. Families with multiple children were used to estimate the recombination rate within the FESFPS-Penta E haplotypes.

Anonymity of the selected cases was preserved by assigning new case and sample identities by an uninvolved researcher outside of the DNA lab. This ensured that no personal information associated with any of the samples was obtained. The only additional information provided with each case was the ethnic background of the parents, which was obtained from the consent forms initially completed by the clients prior to the parentage testing performed by the HIT laboratory.

3.2. Sample Amplification

Amplification of FESFPS and Penta E loci was performed using primer sets supplied with the FFFL and Penta E STR typing kits available from Promega Corp. (Madison, WI). During PCR set-up, these primers were added to a master mix containing other reactants needed for PCR. The master mix, as shown in table 1, consists of 4.83 µl (microliters) of water, 0.83 µl of Gold STAR 10X Buffer (Promega Corp., Madison, WI), 0.83 µl of FFFL primers, 0.83 µl of Penta E primers (Promega Corp., Madison, WI), and 0.17 µl of AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA) for each sample. After the master mix was prepared, 7.5 µl was pipetted into each sample PCR tube. Studies determined the optimum amount of DNA for amplification to be 0.25 µl (approximately 250 pg of genomic DNA) and that amount was added to each PCR reaction tube containing master mix. The yield range for DNA from buccal swabs extracted with the DNAIQ extraction kit from Promega Corp (Madison, WI) is 2-5 µg of DNA. A negative and a positive control were also amplified.

Table 1. Reagents used in PCR set up for FFFL and Penta E

Reagents	**Amount (µL) per reaction**
DIWater	4.83
10X Buffer	0.83
FFFL Primer Set	0.83
Penta E Primer Set	0.83
AmpliTaq Gold	0.17

Amplification occurred in a GeneAmp 9700 thermal cycler (Applied Biosystems, Foster City, CA). Thermal cycling recommended by Promega Corp. (Madison, WI) consisted of two cycle systems, one with 10 cycles and the other with 22 cycles, as outlined in Table 2.

Table 2. Amplification Cycle for FFFL and Penta E

Incubation	96 Degrees for 11 minutes
10 Cycles	94 Degrees for 30 seconds 60 Degrees for 30 seconds 70 Degrees for 30 seconds
22 Cycles	90 Degrees for 30 seconds 60 Degrees for 30 seconds 70 Degrees for 30 seconds
Elongation	65 Degrees for 45 Minutes
Hold	4 Degrees

3.3. Sample Analysis

Sample analysis was performed using ABI 310 genetic or ABI 3130XL genetic analyzers (Applied Biosystems, Foster City, CA). A mixture of Hi-Di Formamide and GeneScan Liz 500 size standard (Applied Biosystems. Foster City, CA) was added to the PCR amplification product. The amount of amplicon added to the Hi-Di/Liz mixture was determined to be 0.3 µl for the best resolution of peaks. The amplicon and Hi-Di/Liz were mixed by vortexing or by pipetting gently up and down.

Once samples were loaded onto the genetic analyzer, the analysis began by using a three second sample injection time and each electrophoretic separation of amplicons occurred during a 27 minute run. A variety of sample injection times was tested along with the amount of amplicon to be added for analysis. The three seconds injection time with 0.3 µl of amplicon gave the best resolution and balance of peak heights. Along with the samples, the negative control, positive control, and an allelic ladder were also analyzed for quality control purposes and sizing information. The allelic ladder for running the FFFL-PE panel contains two separate ladders, FFFL and Penta E available from Promega (Madison, WI) that were combined.

Allele peaks for the FESFPS and Penta E Loci were analyzed using GeneMapper v 3.2 software (Applied Biosystems, Foster City, CA). Figure 1 shows results from a family trio that was analyzed using GeneMapper software.

Analysis began with observing the FESFPS and Penta E alleles in the mother's and father's profile, which made up the possible haplotypes inherited by the child. The child's profile was then analyzed to identify the particular haplotypes inherited from the parents. In table 3, the haplotypes inherited by the child from the mother and father are shown in the example results from one case.

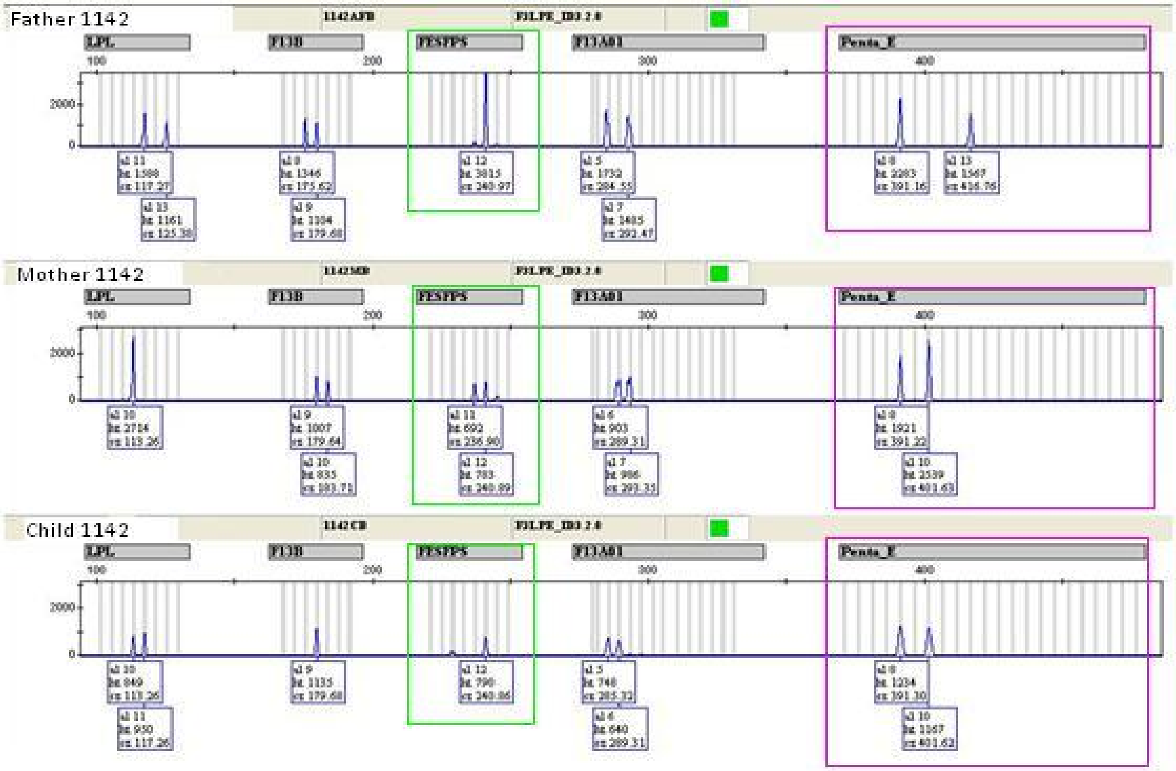

Figure 2. FESFPS and Penta E GeneMapper results from Family Trio 1142

STR typing results obtained from DNA of a family trio is shown. Loci amplified in the multiplex include LPL (first on the left), F13B (second from the left), FESFPS (green box, third from the left), F13A01 (fourth from the left) and Penta E (purple box, far right). The alleles at each locus are identified by the number of repeats in the amplicon (labeled "al #" in the box below each peak in the histogram). Other information provided by the genotyping is the peak height in relative fluorescent units (RFU) and peak area.

Table 3. Inherited Haplotypes from Trio 1142.

The First number represents the FESFPS allele and the second the Penta E allele in each haplotype.

Haplotype from Mother	12/10
Haplotype from Father	12/8

A total of 100 cases were subjected to FESFPS and Penta E typing for each of the two ethnic groups studied, giving a total of up to 400 haplotypes entered into the database. The number of haplotypes changed depending on whether a haplotype underwent recombination or if a parent is homozygous or heterozygous. An excel spreadsheet containing the inherited haplotypes served as the database. The frequency of each haplotype in each population group was calculated by dividing the number of observations of a specific haplotype by the total number of observed haplotypes in the population. Observed frequencies were converted to the upper 95% confidence interval before being used in family relatedness calculations. Table 4 shows examples of relationship testing using the upper confidence interval.

Table 4. Example of a parentage calculation using the 95% confidence interval frequency in three different cases

Haplotype inherited by the child from the father	Upper 95% confidence interval frequency	Likelihood ratio calculation	LR value
10/12	0.0714	$(0.5_{M})(0.5_{AF})/(0.5_{M})/(0.0714_{RM})$	7.007
11/16	0.0278	$(0.5_{AM})(0.5_{F})/(0.0278_{RF})(0.5_{F})$	11.373
12/9	0.0198	$(0.5_{M})(0.5_{AF})/(0.5_{M})(0.0198_{RM})$	25.315

3.4. Statistical Analysis Methods

3.4.1. Calculation of haplotype frequencies and frequency of recombination between FESFPS-Penta E markers

Haplotype frequency was determined by using the counting method, n/N, where n is the observed number of a specific haplotype and N is the total number of haplotypes in the database. A total of 400 Black haplotypes and 396 Caucasian haplotypes were added to the database

Since linked genetic marker systems usually have smaller databases than unlinked genetic systems, the upper 95% confidence interval was used for calculating likelihood ratios for parentage indexes. If a haplotype had only been encountered once before in the database then the 95% CI value was determined with the formula $1 - \alpha^{1/N}$ where N is the total number of haplotypes in the database and α is the 0.95 (the confidence interval desired). When a haplotype was observed more than once a second formula, $p + 1.96\sqrt{}\{\frac{p(1-p)}{N}\}$, where p is the observed frequency of the haplotype, was used.

The FESFPS-Penta E recombination rate was estimated through the use of multi-children families. The parents and children were typed to determine the phase, or the order of FESFPS and Penta E alleles in the parents. If all children had the same haplotype phase inherited from the parents, then recombination did not occur. If more than one possible phase for the haplotypes inherited from the parents were observed, recombination occurred and the original haplotype phase must be determined. Determination of the original phase was done by observing all the haplotypes present in the children. Since haplotypes do not undergo recombination frequently, the predominant haplotype observed was considered to be the un-recombined haplotype inherited from the parents.

3.4.2. Statistical Analysis using Two-Way ANOVA

A frequency distribution table was created to show the range of haplotypes and their frequencies for both Blacks and Caucasians. The distribution of haplotypes in the two ethnic groups was examined statistically in an attempt to detect any significant association of haplotypes, gender, and ethnicity. Two-way ANOVA was the chosen method to evaluate the statistical significance of ethnicity versus haplotype frequency and also gender and ethnicity versus haplotype using Graphpad Prism software (GraphPad Software, Inc., La Jolla, CA). The mean and standard error of the mean for both ethnicity versus haplotype frequency and ethnicity with gender versus haplotype frequency were investigated using Graphpad software. Results of the analysis identify any interaction or association between variables that were significant and whether the variables themselves were significant.

3.4.3. Likelihood Ratio Calculations

For the calculation of likelihood ratios, the effect of using haplotypes versus either the FESFPS or Penta E alleles (whichever resulted in the higher LR value) on the calculated likelihood ratio (LR) was calculated for 256 first order relationship tests in Blacks and 184 in Caucasians ± the standard error of the mean. Additional types of relationship tests were also evaluated (i.e., sibships and half-sibships), using haplotypes versus allele frequencies. Depending on the test performed and the availability of family members, likelihood ratios were produced using either the counting method as described by Dr. Myrna Traver (Allen, 2010) or the exact method as described by Dr. Robert Allen

(Allen, 2010). For cases of half-sibships, the exact method was used. In a case where a parent was known and the relationship being tested was full sibship, the counting method was employed.

Chapter IV. Results

4.1. Haplotype Results

4.1.1. FESFPS-Penta E Haplotypes

102 distinct FESFPS-Penta E haplotypes were observed among 200 parentage trios consisting of a mother, a father, and a child compared to 234 theoretically possible haplotypes, which was calculated by taking the number of all FESFPS alleles and multiplying it by the number of all the Penta E alleles. 100 of the trios were Caucasian and 100 were Black. Ninety-six different haplotypes were observed among parents in the 100 Black families whereas 57 different haplotypes were seen in Caucasians, 51 of these haplotypes were observed in both ethnic groups. Of the total number of haplotypes detected, six haplotypes in Caucasians were not observed in the Black population while 45 of the haplotypes seen in Blacks were not observed in Caucasians (Table 5).

Table 5. Observed Haplotypes and Differences

Ethnicity	Black	Caucasian	Total Number of Haplotypes Observed in both ethnic groups
Number of different haplotypes observed	96	57	153
Number of haplotypes not observed in the other population	45	6	
Number of haplotypes seen in both ethnic groups	51		

4.1.2. Recombination

Among the parentage cases from the Black population, there were 18 families with multiple children. A family with multiple children is especially useful when studying haplotypes since recombination events between the linked markers can only be detected when the event occurs and is seen in one child but not others. Out of 18 multi-child, Black families, with a total of 51 children, six families had a child in which the FESFPS-Penta E haplotype differed from the other child(ren), in the family, 8 children in total, and is most logically explained through recombination. There was insufficient number of Caucasian families with multiple children to detect any recombination events.

Table 6. Recombination in the Black Population

Total Number of Children	Number of Children with Observed Recombination	Observed Rate of Recombination
51	8	16%

In families with only two children, recombination could be detected. However, distinguishing which haplotype was the recombinant and which was the non-recombinant was not possible when only two children were available for testing. Therefore, families with more than two children that exhibited recombination were the only ones in which it was possible to determine which haplotype was the non-recombinant haplotype (see Figure 3 for example).

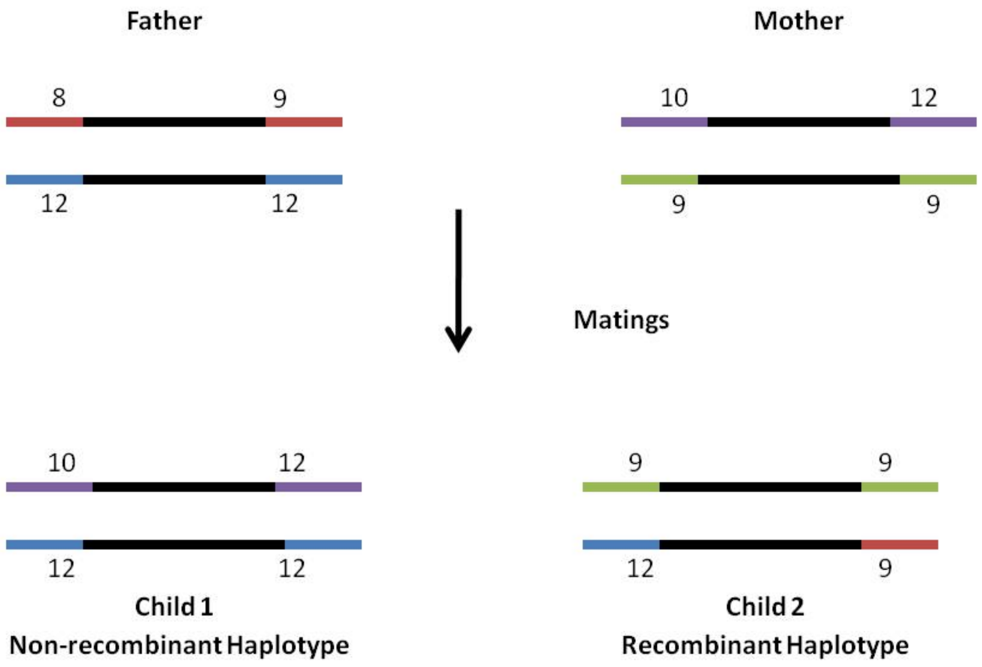

Figure 3. Results of a recombination in a multi-child family
Figure 3 shows results obtained from one family with five children that underwent recombination in the father's haplotype. Results for one of the children (child 2) exhibiting a recombinant haplotype are shown as are results for a child (child 1) exhibiting a non-recombinant haplotype representative of the remaining children (not shown).

4.1.3. Haplotype Frequency Database

In order to assess the increase in the discriminatory power associated with the use of the FESFPS-Penta E haplotype as opposed to individual allele frequencies, a haplotype frequency database was needed. The database produced is listed in Appendix A and lists each observed haplotype, the number of times each haplotype was observed in the different ethnic groups, the absolute haplotype frequency, and the haplotype frequency corrected to the 95% confidence interval (CI). The haplotype frequency database shown below (Table 7) contains the haplotype and the corrected 95% confidence interval frequency that was used in likelihood ratio calculations. Like allele frequency databases, the haplotype frequency database takes into account ethnicity/racial status of the sample donor.

Table 7. Haplotype Frequency Database

Haplotype	Corrected Black Frequency	Corrected Caucasian Frequency
7/7	0.0075	0.012*
8/5	0.0075	0.0075
8/6	0.0075	0.012*
8/7	0.0075	0.0075
8/8	0.016	0.012*
8/9	0.0075	0.012*
8/10	0.0075	0.012*
8/11	0.0234	0.012*
8/12	0.0198	0.012
8/13	0.0198	0.012*
8/14	0.016	0.0075
8/15	0.0234	0.012*
8/16	0.016	0.0075
8/17	0.016	0.0075
8/18	0.0198	0.012*
8/19	0.0075	0.012*
9/5	0.026	0.0075

9/8	0.0119	0.012*
9/9	0.0075	0.012*
9/13	0.0075	0.012*
9/15	0.0075	0.012*
9/16	0.0075	0.012*
9/17	0.0075	0.012*
9.3/12	0.0119	0.012*
10/5	0.0119	0.0472
10/7	0.0198	0.066
10/8	0.0198	0.0199
10/9	0.0119	0.0161
10/10	0.0198	0.0374
10/11	0.0467	0.0504
10/12	0.0714	0.0781
10/13	0.0269	0.0567
10/14	0.0403	0.0272
10/15	0.053	0.0307
10/16	0.0337	0.0272
10/16.4	0.0075	0.012*
10/17	0.0269	0.0199
10/18	0.0198	0.0199
10/19	0.0119	0.012
10/20	0.016	0.0075
10.2/7	0.016	0.012*
10.2/10	0.012*	0.0075
10.2/11	0.0075	0.012*
10.2/12	0.016	0.012*
10.2/13	0.0075	0.012*
10.2/14	0.0198	0.012*
10.2/16	0.0075	0.012*
10.2/17	0.0119	0.012*
10.3/7	0.0075	0.012*
10.3/12	0.0075	0.012*
11/5	0.0269	0.0374
11/7	0.0304	0.0721
11/8	0.0435	0.012*
11/9	0.0304	0.0075
11/10	0.016	0.069
11/11	0.037	0.0567
11/12	0.037	0.107
11/13	0.0435	0.0535
11/14	0.0269	0.0535
11/15	0.0337	0.0272
11/15.4	0.012*	0.0075

11/16	0.053	0.044
11/16.4	0.0075	0.012*
11/17	0.0234	0.0472
11/18	0.0337	0.0236
11/19	0.0198	0.012
11/20	0.0119	0.0075
11/21	0.012*	0.0075
11/22	0.0075	0.012*
11/23	0.012*	0.0075
12/5	0.0119	0.0199
12/7	0.0198	0.0535
12/8	0.0435	0.012
12/9	0.0198	0.012*
12/10	0.0304	0.0374
12/11	0.0269	0.0472
12/12	0.0304	0.0407
12/13	0.037	0.0407
12/14	0.0337	0.0341
12/15	0.0435	0.0236
12/16	0.0269	0.0272
12/17	0.0234	0.0236
12/18	0.0269	0.012*
12/19	0.0119	0.012*
12/20	0.016	0.0075
12/21	0.0075	0.012
12/22	0.0075	0.012
13/5	0.012*	0.0161
13/7	0.016	0.012
13/8	0.0119	0.012*
13/10	0.0075	0.012*
13/11	0.0075	0.012
13/12	0.0337	0.012*
13/13	0.0075	0.0199
13/14	0.0075	0.012*
13/15	0.0075	0.012*
13/16	0.0075	0.0075
13/17	0.012*	0.0075
13/19	0.0075	0.012*
13/20	0.0075	0.012*
13/22	0.0075	0.012*
14/8	0.0075	0.012*

* denotes the minimum frequency of an unobserved haplotype calculated using the equation 5/N+1, where N is the number of haplotypes in the database, suggested in the second report of the NRC.

The haplotype with the highest frequency observed for Caucasians was 11/12 at a frequency of 0.107 and the haplotype with the highest frequency for Blacks was 10/12 at a frequency of 0.07. The lowest frequency of an observed haplotype was 0.0075 in either population. The National Research Council suggested using a minimum frequency calculated by 5/N+ 1 where N is the number of haplotypes in the database being used, in this case 400. This formula was used for haplotypes not observed in either population.

4.2. Statistical Analysis of Haplotypes

4.2.1. Statistical Analysis of Haplotype Frequencies

To ensure likelihood ratio calculations were sufficiently conservative, given the number of haplotypes collected in the population study, the upper 95% confidence limit of the frequency estimate was calculated. For haplotypes that were observed only once the formula used to calculate the 95% CI frequency was $1 - \alpha^{1/N}$. For haplotypes observed more than once the formula used was $p + 1.96\sqrt{}\{\frac{p(1-p)}{N}\}$. In the case of one observation of a haplotype, the counted frequency would be 0.0025 while, after correction to the 95% CI, the frequency rises 0.0075. This effect was observed for both Black and Caucasian populations. For haplotypes observed more than once in a population, the difference between the absolute frequency and the frequency corrected to the upper 95% confidence interval differed among Blacks and Caucasians due to the different haplotype counts in the two populations. For example, in Blacks haplotype 11/12 was observed 9 different times and in Caucasians is was observed 32 times. In the

Black population, the mean difference between the counted and upper 95% CI frequency of a given haplotype was 0.0091. In the Caucasian population, the mean difference was higher at 0.0115. Using an unpaired t test, the means between the two populations was determined to be significant (p=0.0065) and the variance was also significant (p<0.0001). These results suggest there is a difference in haplotype frequency between the two ethnic groups.

4.2.2. Analysis between Populations

Two-way ANOVA was performed to examine possible correlations or significant differences between the diversity of haplotypes observed in Blacks and Caucasians and their relative frequencies in the two populations. Ethnicity and gender showed no statistically significant relationship in determining haplotype diversity (p=0.9918 for gender and p=0.9709 for ethnicity). These results indicate that knowing ethnicity and/or gender will be of no predictive value in determining the haplotypes exhibited by an individual. The only significant predictive indices in a person's haplotype makeup is the haplotype frequency in the particular ethnic population (p <0.0001).

Table 8. Number of Haplotypes Observed Specific Between Gender

	Black Males	Caucasian Males	Black Females	Caucasian Females
Number of apparent "restricted haplotypes"	**32**	**7**	**28**	**8**

4.2.3. Effect of using haplotypes rather than allele frequencies for likelihood ratio calculations

One of the goals of this study was to assess the possible increase in the discriminatory of the STR test battery used by the OSU-HIT laboratory through the use of FESFPS-Penta E haplotypes as opposed to using the allele frequencies in LR calculations for one of the two loci (whichever produced a higher value for the LR). Among the cohort of archived relationship cases tested, the use of a FESFPS-Penta E haplotype in calculations resulted in an increase in the likelihood ratio (LR) produced when compared with the use of either FESFPS or Penta E locus (whichever produced the higher LR result). Thus, the use of haplotypes increases the discriminatory power of the test battery. The average increase in combined LR in Blacks was 1.84 fold whereas in Caucasians, the average increase was 2.43 fold. In 38 of 256 Black parentage cases (14%) in which LR values were calculated, there was no significant increase in the LR accompanying the use of haplotypes. In Caucasians, 12 out of 184 LR values (6%) did not benefit from the use of haplotypes rather than alleles.

Five out of 10 analyzed cases where low probability values were obtained were for relationship testing of half-sibship calculations; three were in paternity cases involving STR locus mutations, one case involved a full sibship with a known parent, and one case of full sibship with no known parents. In the paternity cases with mutations, an average increase of 7.05 fold in LR value power was observed. In the case of full sibship with a known parent, the haplotype calculation did not improve the probability calculated. Table 9 shows the calculations from a paternity mutation calculation using the haplotype from case 1. In this case, the alleged father showed a mutation at one STR locus which greatly reduced the overall combined LR for the test battery. The mother was not available for testing and the father shared two alleles at the Penta E locus with the child, introducing ambiguity into the analysis. Thus both 8/8 and 8/9 haplotype frequencies were used in the calculation.

Table 9. Haplotype versus allele LR calculations from case 1, paternal mutation, mother not tested

LR calculation using the FESFPS-Penta E haplotype: 8/8 or 8/9 $((0.016_{RW-8/9}*0.5_{AF-8/8})+(0.0075_{RW-8/8}*0.5AF\text{-}_{8/9}))/(2*(0.016*0.0075))= 48.96$
LR calculation using FESFPS allele: 8 $(0.109_{RW-8}*0.5_{AF-8})/[(0.109^{2}+0.109(0.891)(0.01)]= 4.24$
LR calculation using Penta E allele: 8 or 9 $[(0.18_{RW-8}*0.5_{AF-9})+ (0.045_{RW-9}*0.5_{AF-8})]/(2(0.18*0.045))= 6.94$

Table 10 presents the calculations from a case of questioned full sibship, designated as case 4, in which the use of haplotype frequencies did not improve the cumulative likelihood value. The known parent had 10/15 and 10.3/18 as their haplotypes. The reference sibling inherited 10/15 and 8/11 as the haplotypes, the obligate haplotype from the second parent must be 8/11. The alleged sibling had inherited 10/15 from the known parent and 8/8 from the unknown parent.

Table 10. Haplotype versus allele LR calculations from case 4, sibling relationship, one known parent

LR calculation using FESFPS-Penta E haplotype: 8/8 P(8/8>P2)=[(0+0.016)/2]= 0.008 $(0.5_{\text{KP1-10/15}}*0.008_{\text{P2-8/8}})/(0.5_{\text{KP1-10/15}}*0.016_{\text{RP-8/8}})= 0.5$
LR calculation using FESFPS allele: 8 P(8>P2)= [(1+0.109)/2]= 0.507 $(0.5_{\text{KP1-10}}*0.507_{\text{P2-8}})/(0.5_{\text{KP1-10}}*0.109_{\text{RM-8}})=5.08$
LR calculation using Penta E allele: 8 P(8>P2)= [(0+0.18)/2]=0.09 $(0.5_{\text{KP1-15}}*0.09_{\text{P2-8}})/(0.5_{\text{KP1-15}}*0.18_{\text{RM-8}})=0.5$

Chapter V. Discussion

5.1. Collection of Haplotypes

5.1.1. Construction of an FESFPS-Penta E Haplotype Frequency Database

Allele and haplotype frequency databases are generally constructed along ethnic/racial lines using samples from individuals who self-identify their ethnic/racial status. Thus, the databases in widespread use in identification laboratories in the U.S. probably contain significant numbers of samples from individuals who represent themselves as belonging to one group, but are actually admixtures of more than one racial/ethnic group. Such admixtures are probably also represented in the haplotype frequency database constructed in this study for Caucasians and Blacks, due to individuals actually belonging to more than one ethnic/racial group. Nonetheless, the self-identification method for sample designation is useful, and realistically, it is the only method available to categorize individuals.

The same concepts were employed to construct the FESFPS-Penta E haplotype frequency database as those used for the construction of an allele frequency database. The individual haplotypes were counted from profiles produced from collected samples, and when divided by the total number of haplotypes observed, produced a haplotype frequency value. However, because of the low number of samples collected (i.e., 400 potential haplotypes from 100 parentage cases for each ethnic/racial group), the upper 95% confidence interval was calculated for each haplotype frequency to assure that any calculations using haplotype frequency were conservative.

The FESFPS-Penta E haplotype database was constructed in the same way as a Y-STR database. Y-STRs are considered haplotypes because there is no other chromosome for the Y chromosome to recombine with during meiosis. Therefore the STRs on the Y chromosome are tightly linked. Y-STR databases use population data collected by researchers working in various regions of the world and the largest Y haplotype database currently available is the Y chromosome Haplotype Reference Database (YHRD, http://www.yhrd.org). The YHRD contains Y-STR haplotypes from Europe, Asia, and the United States and consists of 93,290 haplotypes at this time (Willuweit & Roewer, 2007). The data used is compiled in the same way that was used for this study, using commercially available kits and a standardized method. The ethnic/racial identity of a DNA sample donor is important in collecting Y haplotypes because those markers are highly conserved within different ethnic groups (Willuweit & Roewer, 2007). However, identification of the ethnic origins of sample donors for the YHRD database was accomplished by self identification by the donor, as was done here.

The haplotype databases produced in this study showed differences both in the numbers of different haplotypes observed between Caucasians and Blacks, and in the relative frequencies of haplotypes observed in both ethnic populations. An examination of Figure 4 reveals that there were more distinct haplotypes detected in Blacks than Caucasians. A total of 96 different haplotypes were observed in the Black population whereas in Caucasians, 57 different haplotypes were observed. Moreover, some of the observed haplotypes were seen only in the Black population (Figure 4).

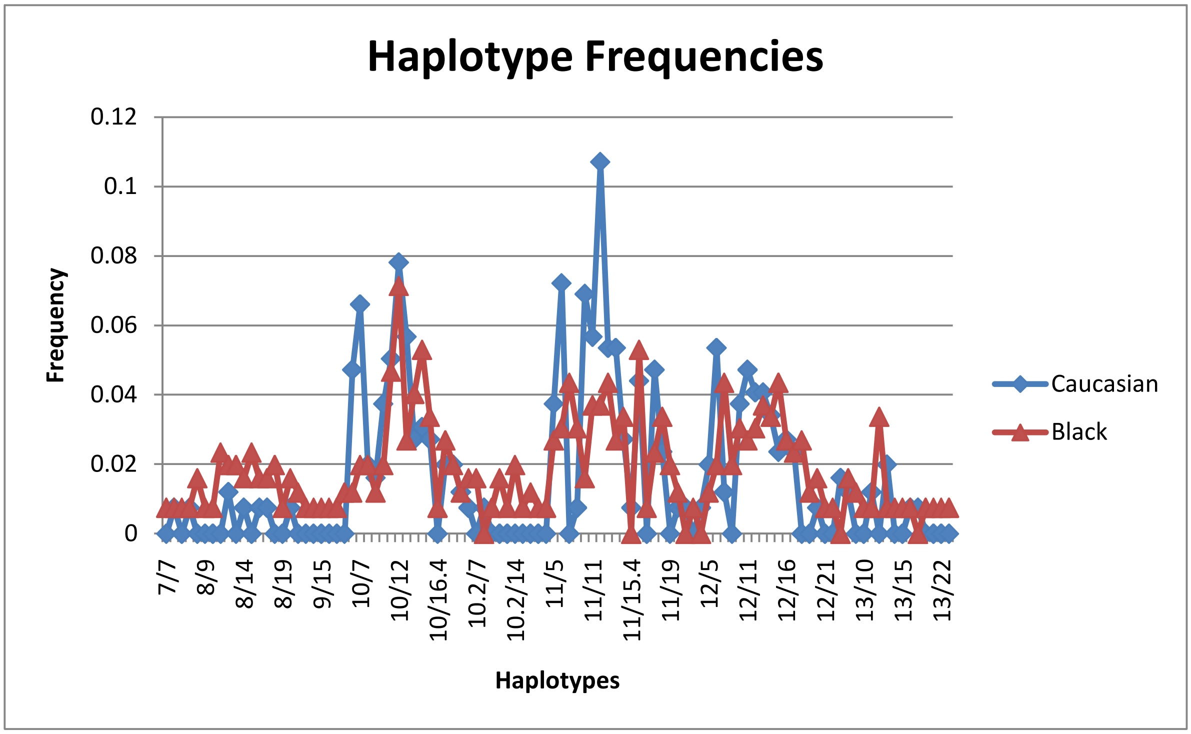

Figure 4. Frequency Distribution for Black and Caucasian Populations

Figure 4 shows the frequency distribution FESFPS-Penta E haplotypes in the Black (red) and Caucasian (blue) populations. The frequency values shown represent the 95% CI values.

A relationship between haplotype diversity and ethnicity is suggested from our data due to the rather large number of haplotypes observed in only one population group, much like the private alleles for the D9S164 genetic marker observed in the study done by Calafell et al. (1998). The difference in haplotype diversity observed between Blacks and Caucasians is not restricted to one sex or the other. Even with the excess in haplotype diversity seen in the Black ethnic/racial group, there are 51 haplotypes that were observed in both populations (Table 4).

5.1.2. Possible relationship between haplotypes and gender

As might be expected for a haplotype located on an autosome, there did not appear to be any kind of relationship between gender and haplotype. However, 20 haplotypes were observed in Black males that were not seen in Black females and there were 24 haplotypes seen in Black females that were not observed in Black males. Although Caucasians exhibited fewer haplotypes that appeared to be restricted (8 in males and 11 in females), restricted haplotypes existed nonetheless in our population sampling (Table 8).

Finally, when both ethnicities/racial groups were examined, there were haplotypes that appeared to be restricted both by gender and ethnicity. For example, haplotype 8/6 was only observed in Black females and 13/7 was observed only in Caucasian males.

Black males had 16 haplotypes, and Black females had 15 haplotypes that appeared to be specific to their gender within their particular ethnic group, whereas Caucasian males had 1 specific haplotype and Caucasian females had 4 restricted haplotypes (Table 11). When ethnicity was taken into consideration along with gender, the number of restricted haplotypes in both males and females were further reduced past the number of specific haplotypes to each ethnicity. The number of haplotypes specific to Blacks is 45 haplotypes; however some males and females shared haplotypes and are not considered specific to ethnicity and gender. Although there were 51 haplotypes that overlapped in both population groups, overall, the Black population had a larger number of distinct haplotypes than did Caucasians. Similar work done by others (Bamshad, et al., 2004; Calafell, et al., 1998) suggests that Blacks exhibit greater genetic diversity for many DNA markers when compared to Caucasians.

Table 11. Haplotypes Specific to Ethnicity and Gender

	Black Female	**Black Male**	**Caucasian Female**	**Caucasian Male**
Number of Haplotypes	**15**	**16**	**4**	**1**

When separated by gender, there was a significant interaction between haplotypes and gender when two-way ANOVA was used to analyze the data. However, analysis with ANOVA proved that gender did not affect the haplotype seen. The significant interaction is likely due to the restricted haplotypes observed. When looking at haplotype diversity within each ethnic group subdivided by gender, Black females exhibit the greatest haplotype diversity, followed by Black males, Caucasian females, and lastly Caucasian males. Thus, Caucasians males are the least diverse in terms of FESFPS-Penta E haplotypes, having only 6 different haplotypes observed out of 100 DNA samples subjected to FESFPS and Penta E typing. It should also be noted that at least some of the apparent restriction in haplotype diversity between the sexes and between the different ethnicities could be due to the size of the population sampled. As stated in the results section, apparent frequencies of the different haplotypes in each population group is fairly low. Although the 95% confidence limit on haplotype frequencies was produced and used for calculations, there still could have been sampling bias that affected the result of the statistical analysis of the haplotype frequency databases, having just 200 individuals characterized from each ethnic group.

Genetic diversity is generally assessed through sampling populations that self-identify their ethnicity and comparing the number of different genotypes within the population or sub-populations sampled. Diversity can account for the significant interaction between ethnicity and the haplotype encountered. Bamshad et al. (2004) concluded that genetic analysis can establish and differentiate between groups. The frequencies of STR alleles at a locus are inversely related to the size of the population and STRs that arose from a single group have a higher chance of remaining solely within the group due to geographical restrictions (Bamshad, et al., 2004). Comparison of continental populations around the world, Bamshed et al. (2004) found that Africa had the highest genetic diversity. On the African continent, differences in genetic diversity were also found within sub-population groups of Africans (Bamshad, et al., 2004; Serre & Paabo, 2004). Differences in geographical origin and population admixture can also impact genetic diversity, giving rise to more alleles at a locus. More off-ladder STR alleles were seen in the Black population in this study than in Caucasians, which agrees with previous results reported in the study of Calafell (1998), who saw more off ladder alleles in populations originating in Africa than other areas of the world. Off ladder alleles are alleles that do not fit within the bins created by the sizing ladder of the loci. These alleles were specific and only appeared in the Caucasian population twice, also suggesting more genetic diversity exists within the Black population.

5.1.3. FESFPS-Penta E Recombination

Out of the 18 multiple child families typed in the Black population, 6 families exhibited evidence of a recombination between the FESFPS and Penta E loci, suggesting a recombination rate of about 33.3%. Since the recombination rate is below the 50% mark, FESFPS and Penta E are not independent of each other, and therefore must be considered to be linked. For two loci to be considered independent, the recombination rate between them must be at least 50%.

In a family with five children, the non-recombinant haplotype passed from the parents to the children was determined. Out of the five children, only one inherited the recombinant haplotype (shown in Figure 3), in which the FESFPS-Penta E haplotype in the father recombined to form a 12/9 haplotype instead of the precursor 12/12 haplotype.

The recombination rate, also generally reflects how close two markers exist on a chromosome. The more tightly linked two or more markers are, the lower the general recombination rate between them will be. The recombination rate for FESFPS/Penta E is below the threshold of marker independence, but it is not low enough for the FESFPS and Penta E markers to be considered to be tightly linked. FESFPS and Penta E are located on chromosome 15, on the very end of the q arm. More detailed mapping suggests the two loci are approximately 6 million basepairs apart on chromosome 15 (AABB, 2010). Whereas the recombination rate in Blacks was about 16%, there were no recombinations observed in Caucasians. However, there also were not as many multi-child families available for analysis in the Caucasian group. While we cannot estimate a recombination rate in Caucasians for these reasons, Dr. Maha of LabCorp, in his study that involved more multi-children families, found a recombination rate of about 16% in Caucasians (G.C. Maha, personal communication, April 25, 2010). Dr. Maha has been able to sample hundreds of recombinants, while this study only had 18 recombinants. However, both rates are still higher than the expected rate estimated by typical recombination rate in human chromosomes as reflected in the centiMorgan distance reported to be between the two loci. Thus the recombination rate would appear not to differ significantly between Blacks and Caucasians, especially if both studies were able to observe the same amount of samples, even though the genetic diversity of haplotypes in Blacks might suggest the rate would be higher.

The 6 million basepairs of separation between FESFPS and Penta E correlates to about 5 centiMorgans (cM), which is the unit of recombination distance of a genetic map (Yu et al., 2001). A cM is measured to be roughly equal to 1% recombination on a typical chromosome. The recombination rate between FESFPS and Penta E therefore should be about 5%. However, the recombination rate observed in Blacks and reported in Caucasians was about five times higher than this value. These results suggest that there is a recombination hotspot between the two markers, making the recombination rate higher than expected, but low enough that the two markers are still linked.

The moderate recombination rate observed for FESFPS-Penta E haplotypes in Blacks will undoubtedly cause complications in family studies in which FESFPS-Penta E haplotypes are used to assess claimed or suspected family relationships because a recombination that occurs during meiosis will produce a gamete that may contribute to conception that does not reflect genetically the haplotypes existing in the parent, perhaps providing false evidence of non-parentage. Caution will therefore need to be used in evaluating haplotype results when results that disagree with the totality of autosomal STR testing are encountered. One approach to alleviate this complication would be to identify a third polymorphic marker located between FESFPS and Penta E on chromosome 15 that would define more tightly linked haplotypes for the system and also further likely enhance the discriminatory power expected of the resulting three locus haplotypes. Additional linked loci on chromosome 15 would also help define the location of recombination events when they occur.

In order to use the FESFPS-Penta E haplotypes, one must know the phase of FESFPS and Penta E markers in the individual. This can be rather straightforward when comparing the mother and child since the mother's relationship to the child is unquestioned and thus the matching FESFPS and Penta E markers will be considered her haplotype transmitted to the child. However, in cases lacking a mother (i.e. known parent), establishing the identity of the haplotypes through establishing the phase of the individual STR markers may be much more difficult. Thus all possible haplotypes that can be produced from two FESFPS alleles and two Penta E alleles in an alleged parent may need to be considered in the likelihood ratio calculations. This effect was observed when calculating a relationship index for a case of suspected sibship. In this case there were no known parents to help determine the phase of the FESFPS-Penta E haplotypes inherited and all 86 possible parental combinations of haplotypes able to produce the reference sibling had to be considered in the calculation.

5.2. Statistical Analysis

5.2.1. Haplotype use in Likelihood Ratios

Using haplotypes in first order relationship testing showed an average increase of 2.43 fold for Blacks and 1.84 fold for Caucasians in the magnitude of the LR produced when compared with using either FESFPS or Penta E results (whichever gave the highest LR value). However, 14% of Black parentage cases and 6% of Caucasian parentage cases did not have a significant increase in the magnitude of the LR value. The number of LRs calculated that showed no improvement resulting from the use of haplotypes was small. The LRs calculated using haplotypes that showed no significant increase in magnitude were calculations involving either FESFPS or Penta E alleles that were very rare in the population, as witnessed by the low allele frequencies in the allele database, and thus produced large LR values using the single locus alone. Cases in which the use of haplotypes had the greatest effect were cases involving FESFPS or Penta E alleles that were relatively common in the population. Of course such cases will be the most often encountered since the allele frequencies are higher. In such cases, the use of haplotypes will have a more pronounced effect since the frequency of even common FESFPS-Penta E haplotypes is much lower than that of the corresponding alleles composing them.

The use of haplotypes in cases of questioned family relatedness in which second order relatives were all that were available for testing sometimes did not improve the calculated probability that two individuals were related. This most often occurred when the two tested individuals did not share a common haplotype. The result was the same as if two individuals did not share a common allele at the locus in question. No cases with shared haplotypes were encountered so it is not possible to estimate the effect of the use of haplotypes on the LR values produced. What does appear clear is that when the haplotype can be conclusively identified, the LR was increased using haplotypes compared with alleles for one of the two loci, but when the haplotype could not be conclusively identified, a LR close to or less than one often resulted. This observation is typical of LR calculations produced in pedigrees containing significant ambiguity in the genotypes of family members, who must have their genotypes reconstructed through the testing of other family members.

Appendix A

Table 12. Haplotype Frequency Database

Haplotype	Black Number Observed	Black Absolute Frequency	Corrected Black Frequency	Caucasian Number Observed	Caucasian Absolute Frequency	Corrected Caucasian Frequency
7/7	1	0.0025	0.0075	0		0.012*
8/5	1	0.0025	0.0075	1	0.0025	0.0075
8/6	1	0.0025	0.0075	0		0.012*
8/7	1	0.0025	0.0075	1	0.0025	0.0075
8/8	3	0.0075	0.016	0		0.012*
8/9	1	0.0025	0.0075	0		0.012*
8/10	1	0.0025	0.0075	0		0.012*
8/11	5	0.0125	0.0234	0		0.012*
8/12	4	0.01	0.0198	2	0.005	0.012
8/13	4	0.01	0.0198	0		0.012*
8/14	3	0.0075	0.016	1	0.0025	0.0075
8/15	5	0.0125	0.0234	0		0.012*
8/16	3	0.0075	0.016	1	0.0025	0.0075
8/17	3	0.0075	0.016	1	0.0025	0.0075
8/18	4	0.01	0.0198	0		0.012*
8/19	1	0.0025	0.0075	0		0.012*
9/5	3	0.0075	0.026	1	0.0025	0.0075
9/8	2	0.005	0.0119	0		0.012*
9/9	1	0.0025	0.0075	0		0.012*
9/13	1	0.0025	0.0075	0		0.012*
9/15	1	0.0025	0.0075	0		0.012*
9/16	1	0.0025	0.0075	0		0.012*
9/17	1	0.0025	0.0075	0		0.012*
9.3/12	2	0.005	0.0119	0		0.012
10/5	2	0.005	0.0119	12	0.0303	0.0472
10/7	4	0.01	0.0198	18	0.0455	0.066
10/8	4	0.01	0.0198	4	0.01	0.0199
10/9	2	0.005	0.0119	3	0.0076	0.0161
10/10	4	0.01	0.0198	9	0.0227	0.0374
10/11	12	0.03	0.0467	13	0.0328	0.0504
10/12	20	0.05	0.0714	22	0.0556	0.0781

10/13	6	0.015	0.0269	15	0.0379	0.0567
10/14	10	0.025	0.0403	6	0.0152	0.0272
10/15	14	0.035	0.053	7	0.0177	0.0307
10/16	8	0.02	0.0337	6	0.0152	0.0272
10/16.4	1	0.0025	0.0075	0		0.012*
10/17	6	0.015	0.0269	4	0.01	0.0199
10/18	4	0.01	0.0198	4	0.01	0.0199
10/19	2	0.005	0.0119	2	0.005	0.012
10/20	3	0.0075	0.016	1	0.0025	0.0075
10.2/7	3	0.0075	0.016	0		0.012*
10.2/10	0		0.012	1	0.0025	0.0075
10.2/11	1	0.0025	0.0075	0		0.012*
10.2/12	3	0.0075	0.016	0		0.012*
10.2/13	1	0.0025	0.0075	0		0.012*
10.2/14	4	0.01	0.0198	0		0.012*
10.2/16	1	0.0025	0.0075	0		0.012*
10.2/17	2	0.005	0.0119	0		0.012*
10.3/7	1	0.0025	0.0075	0		0.012*
10.3/12	1	0.0025	0.0075	0		0.012*
11/5	6	0.015	0.0269	9	0.0227	0.0374
11/7	7	0.0175	0.0304	20	0.0505	0.0721
11/8	11	0.0275	0.0435	0		0.012*
11/9	7	0.0175	0.0304	1	0.0025	0.0075
11/10	3	0.0075	0.016	19	0.0480	0.069
11/11	9	0.0225	0.037	15	0.0379	0.0567
11/12	9	0.0225	0.037	32	0.0808	0.107
11/13	11	0.0275	0.0435	14	0.0354	0.0535
11/14	6	0.015	0.0269	14	0.0354	0.0535
11/15	8	0.02	0.0337	6	0.0152	0.0272
11/15.4	0		0.012	1	0.0025	0.0075
11/16	14	0.035	0.053	11	0.0278	0.044
11/16.4	1	0.0025	0.0075	0		0.012*
11/17	5	0.0125	0.0234	12	0.0303	0.0472
11/18	8	0.02	0.0337	5	0.0126	0.0236
11/19	4	0.01	0.0198	0		0.012*
11/20	2	0.005	0.0119	1	0.0025	0.0075
11/21	0		0.012	1	0.0025	0.0075
11/22	1	0.0025	0.0075	0		0.012*
11/23	0		0.012	1	0.0025	0.0075
12/5	2	0.005	0.0119	4	0.01	0.0199
12/7	4	0.01	0.0198	14	0.0354	0.0535
12/8	11	0.0275	0.0435	2	0.005	0.012
12/9	4	0.01	0.0198	0		0.012*
12/10	7	0.0175	0.0304	9	0.0227	0.0374

12/11	6	0.015	0.0269	12	0.0303	0.0472
12/12	7	0.0175	0.0304	10	0.0253	0.0407
12/13	9	0.0225	0.037	10	0.0253	0.0407
12/14	8	0.02	0.0337	8	0.02	0.0341
12/15	11	0.0275	0.0435	5	0.0126	0.0236
12/16	6	0.015	0.0269	6	0.0152	0.0272
12/17	5	0.0125	0.0234	5	0.0126	0.0236
12/18	6	0.015	0.0269	0		0.012*
12/19	2	0.005	0.0119	0		0.012*
12/20	3	0.0075	0.016	1	0.0025	0.0075
12/21	1	0.0025	0.0075	0		0.012*
12/22	1	0.0025	0.0075	0		0.012*
13/5	0		0.012*	3	0.0076	0.0161
13/7	3	0.0075	0.016	2	0.005	0.012
13/8	2	0.005	0.0119	0		0.012*
13/10	1	0.0025	0.0075	0		0.012*
13/11	1	0.0025	0.0075	2	0.005	0.012
13/12	8	0.02	0.0337	0		0.012*
13/13	1	0.0025	0.0075	4	0.01	0.0199
13/14	1	0.0025	0.0075	0		0.012*
13/15	1	0.0025	0.0075	0		0.012*
13/16	1	0.0025	0.0075	1	0.0025	0.0075
13/17	0		0.012*	1	0.0025	0.0075
13/19	1	0.0025	0.0075	0		0.012*
13/20	1	0.0025	0.0075	0		0.012*
13/22	1	0.0025	0.0075	0		0.012*
14/8	1	0.0025	0.0075	0		0.012*

Chapter VI. References

AABB. (2010). Linked Loci: FESFPS and Penta E. *AABB Accreditation Relationship Testing News, 5*(1). Retrieved from

Allen, R. (2010). *Calculations used in assessing the strength of the evidence in forensic DNA typing applications*. Tulsa: Oklahoma State University, Center for Health Sciences.

Alonso, A., Andelinovic, S., Martin, P., Sutlovic, D., Erceg, I., Huffine, E., . . . Primorac, D. (2001). DNA typing from skeletal remains: Evaluation of multiplex and megaplex STR systems on DNA isolated from bone and teeth samples [ ]. *Croation Medical Journal, 42*(3), 260-266.

Alonso, A., Martin, P., Albarran, C., Garcia, P., Fernandez de Simon, L., Iturralde, M. J., . . . Sancho, M. (2005). Challenges of DNA Profiling in Mass Disaster Investigatons. *Croation Medical Journal, 46*(4), 540-548.

Andelinovic, S., Sutlovic, D., Ivkosic, I. E., Skaro, V., Ivkosic, A., Paic, F., . . . Primorac, D. (2005). Tweleve-year experience in identification of skeletal remains from mass graves. *Croation Medical Journal, 46*(4), 530-539.

Applied Biosystems. (2003). California Missing Persons DNA Program Challenged with Identifying Thousands of Human Remains.

Associated Press. (2010). Mass graves of Nazi victims are found, *The Boston Globe*. Retrieved from http://www.boston.com

Bamshad, M., Wooding, S., Salisbury, B., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. [review]. *Nature Genetics, 5*, 598-609.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265. doi: 10.1093/bioinformatics/bth457

Burns, J. (2006). Uncovering Iraq's horrors in desert graves, *The New York Times*. Retrieved from http://www.nytimes.com

Butler, J. (2005). *Forensic DNA Typing* (2nd ed.). Amsterdam: Elsevier.

Caher, J. (2009, 2009). relatives of missing loved ones to "bank" their DNA Retrieved September 17, 2010, from http://www.criminaljustice.state.ny.us

Calafell, F., Shuster, A., Speed, W., Kidd, J., & Kidd, K. (1998). Short tandem repeat polymorphism evolution in humans. [Original Paper]. *European Journal of Human Genetics, 6*, 38-49.

California Department of Justice. (2010). Cailifornia Missing Persons Retrieved September 15, 2010, from http://ag.ca.gov/missing

Clayton, T. M., Whitaker, J. P., Fisher, D. L., Lee, D. A., Holland, M. M., Weedn, V. W., . . . Gill, P. (1995). Further validation of a quadruplex STR DNA typing system: a collaborative effort to identify victims of a mass diseaster. *Forensic Science International, 76*, 17-25.

Corach, D., Risso, L. F., Marino, M., Penacino, G., & Sala, A. (2001). Routine Y-STR typing in forensic casework. *Forensic Science International, 118*, 131-135.

Dareini, A. A. (2003). more bodies unearthed in Iraq, from http://iraqfoundation.org

Fearnhead, P., & Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society Series B, 64*(4), 657-680.

Garcia, M., Martinez, L., Stephenson, M., Crews, J., & Peccerelli, F. (2009). analysis of complex kinship cases for human identification of civil wa victims using M-FISys software. *Forensic Science International: Genetics, Supplement Series 2*, 250-252. doi: 10.1016/j.fsigss.2009.08.128

Geltman, P., & Stover, E. (1997). Genocide and the plight of the children in Rwanda. *Journal of the American Medical Association, 277*(4), 289-294.

Gjertson, D., Brenner, C., Baur, M., Carracedo, A., Guidet, F., Luque, J., . . . Morling, N. (2007). ISFG: Recommendations on biostatistics in paternity testing. [review]. *Forensic Science International: Genetics, 1*(3-4), 223-231. doi: 10.1016/j.fsigen.2007.06.006

Graham, E. (2006). Disaster victim identification. *Forensic Science, Medicine, and Pathology, 2*(3), 203-207. doi: 10.1385/Forensic Sci. Med. Patjol.:2:3:203

Hawley, C. (2010). mass graves help solve mystery of Mexico's missing, *Arizona's Home Page*. Retrieved from http://azcentral.com

Hellmann, I., Ebersberger, I., Ptak, S., Paabo, S., & Przewroski, M. (2003). a neutral explanation for the correlation of diversity with recombination rates in humans *American Journal of Human Genetics, 72*, 1527-1535.

Huffine, E., Crews, J., & Davoren, J. (2007). Developing role of forensics in deterring violence and genocide. [Editorial]. *Croation Medical Journal, 48*, 431-436.

ICTY-TPIY. (2011, February 18, 2011). International Criminal Tribunal for the former Yugoslavia: the former Yugoslavia-conflicts Retrieved February 20, 2011, 2011, from http://www.icty.org/sid/322

Jorde, L., Rogers, A., Bamshad, M., Watkins, W. S., Krakowiak, P., Sung, S., . . . Harpending, H. (1997). Microsatellite diversity and the demographics history of modern humans. *94*, 3100-3103.

Lathrop, G. M., Lalouel, J. M., Julier, C., & Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc National Academy of Science, 81*, 3443-3446.

Leclair, B., Fregeau, C., Bowen, C., & Fourney, R. (2004). enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster. *Journal of Forensic Science, 49*(5), 1-15.

Lee, H.-S., Lee, J. W., Han, G.-R., & Hwang, J.-J. (2000). motherless case in paternity testing. *Forensic Science International, 114*, 57-65.

Lewis, K., Caserza, M., Walsh, T., & King, M.-C. (2010). *Identification of distant relatives using haplotypes constructed from multiple linked autosomal STRs*. Paper presented at the American Academy of Forensic Sciences Annual Conference, Seattle, Washington.

Marjanovic, D., Durmic-Pasic, A., Bakal, N., Haveric, S., Kalamujic, B., Kovacevic, L., . . . Primorac, D. (2007). DNA identification of skeletal remains from World War II mass graves uncovered in Slovenia. [Forensic Science]. *Croation Medical Journal, 48*, 513-519.

Myers, S., Bottolo, L., Freeman, C., Mcvean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. [Reports]. *Science, 310*(5746), 321-324.

National Geographic News. (2005, January 7, 2005). The Deadliest Tsunami in History? *National Geographic*.

National Institute of Standards and Technology (NIST). (2010). STR fact sheets (observed alleles and PCR product sizes) Retrieved September 2, 2010, from http://www.cstl.nist.gov/biotech/strbase/str_fact.htm

Olaisen, B., Sternersen, M., & Mevag, B. (1997). identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster *Nature Genetics, 15*, 402-405.

Palo, J., Hedman, M., Soderholm, N., & Sajantila, A. (2007). Repartriation and identification of Finnish World War II soldiers. *Croation Medical Journal, 48*, 528-535.

Roberts, J. (2005). mass graves found in Iraq, *CBS News*. Retrieved from http://www.cbsnews.com

Roewer, L., Krawczak, M., Willuweit, S., Nagy, M., Alves, C., Amorim, A., . . . Kayser, M. (2001). Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Science International, 118*, 106-113.

Serre, D., & Paabo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Research, 14,* 1679-1685.

United Nations. (2008). Convention of the Prevention and Punishment of the Crime of Genocide Retrieved February 20, 2011, 2011, from http://untreaty.un.org/cod/avl/ha/cppcg/cppcg.html

Wahlstrom, H. (2001). "The Doe Network": International center for Unidentified & Missing Persons Retrieved September 2010, from http://www.doenetwork.org

Weaver, K. (2003). Identifying the fallen. *British Medical Journal, 326*, 1110.

Welsh, T. (2010, Novermber 10, 2010). Government vows to identify dead, News, *Colombia reports*. Retrieved from http://colombiareports.com/colombia-news/news/12835-government-unidentified-bodies.html

Willuweit, S., & Roewer, L. (2007). Y chromosome haplotype reference database (YHRD): Update *Forensic Science International: Genetics, 1,* 83-87. doi: 10.1016/j.fsigen.2007.01.017

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungali, A., Deloukas, P., . . . Weber, J. (2001). Comparison of human genetic and sequence-based physical maps. *Nature 409*, 951-953.

VITA

Catharine Worthen

Candidate for the Degree of

Master of Science

Thesis: The enhancement to the discriminatory power of STR typing through the use of haplotypes

Major Field: Forensic Sciences

Education:

Completed the requirements for the Master of Science in Forensic Sciences at Oklahoma State University Center for Health Sciences, Tulsa, Oklahoma, in May, 2011.

Completed the requirements for the Bachelor of Science in Biology at the University of West Florida, Pensacola, Florida, in 2008.

Name: Catharine Worthen Date of Degree: May 2011

Institution: Oklahoma State University CHS Location: Tulsa, Oklahoma

Title of Study: The enhancement to the discriminatory power of STR typing through the use of haplotypes

Pages in Study: 67 Candidate for the Degree of Master of Science

Major Field: Forensic Sciences

Scope and Method of Study:

Identification of remains in the past has been accomplished by using DNA analysis, specifically STR typing. However, there are times when the DNA results do not produce a compelling result for identification. Genetic marker systems of haplotypes are starting to be considered to enhance the STR test battery. The goal of this study was to use two markers, FESFPS and Penta E, that were physically linked and commercially available in STR typing kits to develop a haplotype database and determine the level of enhancement, if any, to the already used STR test batteries. A haplotype frequency database needed to be constructed and the rate of recombination determined.

Findings and Conclusions:

Frequencies for haplotypes consisting of FESFPS and Penta E markers were calculated using a conservative frequency estimate of the 95% upper confidence interval. The FESFPS-Penta E haplotype system was used to calculate parentage test cases and was found to enhance the likelihood ratio values by 2.43 fold in Blacks and 1.84 fold in Caucasians. In paternal mutation cases where the full STR test battery was available, the haplotype frequencies increased the LR values by an average of 7.05 fold. When a case of a questioned sibling relationship was analyzed, using the haplotype frequency instead of the allele frequency favored a non-relationship. The recombination rate of the FESFPS-Penta E haplotype was higher than the expected rate, indicating a recombination hotspot somewhere between the two markers.

ADVISER’S APPROVAL: Dr. Robert Allen