An Assessment of the Forensic Aspects of Genetic Genealogy

Sarah Hester
seh00011@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Forensic Science and Technology Commons

Recommended Citation
https://researchrepository.wvu.edu/etd/11774

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
An Assessment of the Forensic Aspects of Genetic Genealogy

Sarah Hester

Thesis submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in
Forensic & Investigative Science

Keith Morris, Ph.D., Chair
Tina Moroose, M.S.
Jamie Spaulding, Ph.D.

Department of Forensic & Investigative Science

Morgantown, West Virginia
2023

Keywords: single nucleotide polymorphism, investigative genetic genealogy, forensic genetic genealogy

Copyright 2023 Sarah Hester
ABSTRACT

An Assessment of the Forensic Aspects of Genetic Genealogy

Sarah Hester

Forensic genetic genealogy has grown in both popularity and controversy in recent years. The genetic information comes from large numbers of single nucleotide polymorphism (SNP) markers. Polymorphism variations can be traced through familial lineage, determining how closely individuals are related to each other based on the similarities and differences in SNPs within their DNA. Polymorphisms account for approximately 85% of human variation. Genealogy and ancestry determination companies trace the lineage of a person back through multiple generations using polymorphisms. As more individuals submit their DNA for analysis the genealogical profile databases continue to grow in size. Law enforcement uses genealogical determination companies to compare DNA evidence from an unknown source to the thousands of profiles readily available on such websites. Analysts can then determine possible relatives, hopefully leading to an identification of the unknown DNA samples. Companies such as GEDmatch have made identification possible. Law enforcement used the genealogy database, GEDmatch, to identify the Golden State Killer in 2018. Most of these companies explicitly stated that consumer information could only be used in limited situations by law enforcement. However, many companies allowed agencies access in additional situations, as well as subtly changed their terms and conditions to allow law enforcement increased access.

A DNA sample was submitted to AncestryDNA™[2], Family Tree DNA™[10], MyHeritage™[11], and 23andMe™[3] for SNP analysis. Although the profile was from a known individual, for the purpose of this research it served as an unknown DNA profile found at a crime scene. Resulting SNP profiles were uploaded to GEDmatch to identify law enforcement accessible familial matches and assign numerical values to determine relationship strength. AncestryDNA™ Library Edition was used to search public records to build family tree clusters for these matches. From there, a family tree was constructed through which two of these clusters could be linked. This allowed for a union couple to be identified, and, in turn, the individual who committed the crime.

This research streamlines the investigative process with transparency, a practice not typically employed by private genealogy companies or law enforcement. This process was completed using only those samples approved by the individuals for law enforcement use, minimizing any ethical dilemmas, while addressing and evaluating controversial aspects like investigatory privilege, the use of discarded DNA, and other highly debated topics of forensic science within the legal system. This research aimed to create a reference for which those interested in investigative genetic genealogy could use as a guideline, supported via figure summaries, current regulations, possible setbacks, prospective budget and timeline, and clear examples of what resources are available and how they can be utilized.
# Contents

1 Introduction 1

2 Literature Review 3

   2.1 Genetic Genealogy 3
       2.1.1 Single Nucleotide Polymorphisms (SNPs) 3
       2.1.2 Relevant Terminology 5
       2.1.3 SNP Nomenclature 7
       2.1.4 Analysis in GEDmatch 7

   2.2 Current Investigative Processes 9
       2.2.1 Official Guidelines in the United States 10
       2.2.2 Errors 10
       2.2.3 Mixtures 11
       2.2.4 Case Studies 11

   2.3 Ethics 12

   2.4 Legal 13

3 Methodology and Results 15

4 Discussion 37

5 Conclusions and Future Research 41

6 Bibliography 43
1. Introduction

In forensic science, DNA evidence is generally considered to be the most highly associated type of evidence – in the mind of the public, it is the poster child of news stories and entertainment media centered on the discipline. The utility of DNA can be broken down into three main forensic applications – identity testing, where DNA found at a crime scene is used to identify a suspect, relationship testing, where DNA is used to determine a biological relationship, and missing person identification, where DNA is used to identify an individual in situations such as a natural disaster or the discovery of long-abandoned human remains.[1]

Genetic genealogy, also known as investigative genetic genealogy, has the ability to incorporate all three of these scenarios into one succinct forensic application that can answer questions forensic scientists have previously been unable to address. It also includes traditional genealogical research.

Forensic genetic genealogy has grown in both popularity and controversy in recent years. The genetic information comes from large numbers of single nucleotide polymorphism (SNP) markers. Polymorphism variations can be traced through familial lineage, determining how closely individuals are related to each other based on the similarities and differences in SNPs within their DNA. Polymorphisms account for approximately 85% of human variation. Genealogy and ancestry determination companies trace the lineage of a person back through multiple generations using polymorphisms. As more individuals submit their DNA for analysis the genealogical profile databases continue to grow in size. Law enforcement uses genealogical determination companies to compare DNA evidence from an unknown source to the thousands of profiles readily available on such websites. Analysts can then determine possible relatives, hopefully leading to an identification of the unknown DNA samples. Companies such as GEDmatch have made identification possible. Law enforcement used the genealogy database, GEDmatch, to identify the Golden State Killer in 2018. Most of these companies explicitly stated that consumer information could only be used in limited situations by law enforcement. However, many companies allowed agencies access in additional situations, as well as subtly changed their terms and conditions to allow law enforcement increased access.

A DNA sample was submitted to AncestryDNA™[2], Family Tree DNA™[10], MyHeritage™[11], and 23andMe™[3] for SNP analysis. Although the profile was from a known individual, for the purpose of this research it served as an unknown DNA profile found at a crime scene. Resulting SNP profiles were uploaded to GEDmatch to identify law enforcement accessible familial matches and assign numerical values to determine relationship strength. AncestryDNATM Library Edition was used to search public records to build family tree clusters for these matches. From there, a family tree was constructed through which two of these clusters could be linked. This allowed for a union couple to be identified, and, in turn, the individual who committed the crime.

This research streamlines the investigative process with transparency, a practice not typically employed by private genealogy companies or law enforcement. This process was completed using only
those samples approved by the individuals for law enforcement use, minimizing any ethical dilemmas, while addressing and evaluating controversial aspects like investigatory privilege, the use of discarded DNA, and other highly debated topics of forensic science within the legal system. This research aimed to create a reference for which those interested in investigative genetic genealogy could use as a guideline, supported via figure summaries, current regulations, possible setbacks, prospective budget and timeline, and clear examples of what resources are available and how they can be utilized.
2. Literature Review

2.1 Genetic Genealogy

2.1.1 Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms, or SNPs, are a specific sub-category of variations in a DNA sequence. In order to be named a polymorphism, the variant must be present in at least one in 100 individuals ($\geq 1\%$). Generally, polymorphisms can vary in size and complexity, with some variations as complex as long chains of DNA, or a simple change in one base pair i.e. a SNP. A base pair refers to the binding of two complementary nucleotide bases, adenine to thymine or guanine to cytosine, to help comprise a DNA strand. Another example would be a short tandem repeat (STR) which is commonly used in forensic DNA analysis. STR polymorphisms utilize a 2 – 5 base pair region that is repeated a number of times. The presence of SNPs within an individual genome affect the way the body responds to diseases, medicine, and drugs. In forensic science, the primary areas of interest are, firstly, changes in phenotype, and secondly, the way in which SNPs are passed down through generations of family lineage.[4]

Typical forensic approaches to DNA testing primarily utilize STR loci, but the recent turn to SNPs as an analysis method offer both advantages and disadvantages. The main advantage of using SNPs is its usefulness in analyzing degraded DNA samples. Due to its small size, the polymerase chain reaction (PCR) required to amplify the sample can significantly reduce the required fragment length (around 60 – 80 base pairs (bp)) when compared to the required fragment length for STR amplification. Another advantage lies in the low mutation rates of SNPs. These limited rates make SNPs more stable genetic markers compared to larger polymorphisms that tend to have higher probabilities of mutation. In addition, the increased use of SNP technology has led to more interest in SNP applications. Current research centers on the analytical capabilities of the method, aiming to increase productivity with the use of multiplex assays and automation while maintaining quality and efficacy. Although SNPs do have this promising potential, there are certain limitations that must be considered. Current SNP testing technology is not as informative for identity testing compared to STR due to the biallelic property of SNPs. Therefore, more SNPs would be necessary for use in order to equal the discriminating power of STR testing.[4] This would involve using anywhere from 50 – 100 SNP loci compared to the 20 core STR loci. The use of SNPs in DNA mixtures has proven to be ineffective.[4] Finally, although the low mutation rate in SNPs does have positive implications, it also leads to population substructure effects, requiring a more selective sampling of SNPs when performing identity testing.[4]

There are currently four main applications in which SNPs are categorized for genealogical analysis: identity testing/individual identification SNPs (IISNPs), lineage informative SNPs (LISNPs), phenotype informative SNPs (PISNPs), and ancestry informative SNPs (AISNPs). IISNPs are primarily used to differentiate individuals in an effort to include or exclude them as an evidentiary...
source and thus have the same purpose as the commonly used STR loci. Selecting SNPs for identity testing purposes focuses on those with a high heterozygosity (>50%) and a low inbreeding coefficient, or Fst value. These criteria ensure a high discriminating power and require less reference population datasets. Previous studies attempted to determine a select panel of SNPs that would best work for identity testing – in a fashion, similar to the 20 core STR loci. The most notable of these studies was performed by Kidd et al. and published in 2006.[5] Researchers have narrowed down over 90,000 potential SNPs to a panel of 19 based on screening in a wide number of nationalities for the aforementioned optimal criteria of high heterozygosity and low Fst. These SNPs are able to provide a match probability range of $10^{-7}$ to $10^{-6}$ depending on population characteristics such as isolation and inbreeding. Kidd et al. also hypothesized that further development of the panel to approximately 50 SNPs would lower the match probability to approximately $10^{-15}$.[4,5,6]

Lineage informative SNPs (LISNPs) are located on both the mitochondrial DNA (mtDNA) genome and the Y chromosome. LISNPs are typically used in a forensic setting for missing person cases and mass disaster identifications due to their utility in assessing familial linkage. In these instances, a family member of a missing person or a suspected family member of an unidentified individual provides a DNA sample which is then compared to the unknown DNA. The low mutation rates and tendency not to recombine of LISNPs provide ideal characteristics for this kind of casework. They are, however, limited to factors such as the amount of DNA and the low discriminating power of the mtDNA and Y chromosome loci.[4]

Phenotype informative SNPs (PISNPs) are able to provide information about specific phenotypical descriptions for an individual. Although PISNPs are able to be used to determine a wide range of characteristics such as pigmentation, height, and facial features, the majority of the research performed on these SNPs are pigmentation-based, therefore revealing color of the hair, skin, and eyes. The *melanocortin 1 receptor gene* (*MC1R*), the first of the human genome to be attributed to pigment variation, is associated with individuals with red hair and fair skin, going so far back as to finding a variant of this gene (R307G) in the Neanderthal population. An assay of 12 *MC1R* variants was created and tested based on this information, resulting in an 84 percent success rate for detecting individuals with red hair. Other important pigmentation phenotype genes include *SLC45A2* (*MATP*), *SLC24A5*, and *SLC24A4*. SNP F374L in *SLC45A2* is attributed to dark hair, skin, and eyes, while other SNPs on the gene that do not display linkage disequilibrium with F374L are associated with a more olive skin color. *SLC45A2* has a variant that displays lighter skin color in certain European populations. *SLC24A4* is still relatively unexplored due to its very recent correlation with hair and eye pigmentation. The *P* (*OCA2*) gene and alleles on the *TYRP1* and *DCT* genes are associated with iris color. Blue and brown eye color have been attributed to two PISNPs located on the *HERC2* gene. While the comprehension of pigmentation genetics is still relatively narrow, current findings show that the aforementioned genes and their variants are part of the limited population of the human genome with an effect on hair, skin, and eye color.[4]

Lastly, ancestry informative SNPs (AISNPs) are often used – when no suspect is associated with a developed DNA profile – to determine biogeographic ancestry. SNPs are selected for this determination instead of STRs due to the high frequency of allele-sharing which takes place among populations. These SNPs are found throughout the human genome, distributed in different frequencies depending on the ancestry of an individual. It is important to note that unlike PISNPs, AISNPs do not give individual-specific phenotypic predictions. Instead, based on the proportional genetic ancestry information, one can associate typical elements of a particular ancestry with the individual. One example of this would be associating a lighter skin color with a sample that is deemed to be Northern European, as regional and global reference databases have statistically
shown individuals from this region tend to have a light skin color. Kits for this type of analysis have been developed by DNAPrint Genomics, Inc., with their first kit specific to using AISNPs to assign and quantify via percentage an individual’s ancestry to either Sub-Saharan African, European, East Asian, or Native American. Additional kits include a panel of 320 AISNPs specific to sub-categories of Eurasian ancestry and a panel of 1476 AISNPs specific to sub-categories of European ancestry. One of the first applications of AISNPs in forensic casework took place in 2003 in Louisiana. A panel of 73 AISNPs were used to determine the ancestry of a suspected serial killer’s DNA profile – analysis resulted in a breakdown of 85% African ancestry and 15% Native American ancestry, allowing investigators to determine they were looking for an individual with a dark complexion.[4]

2.1.2 Relevant Terminology

**Exonic** present in the exon area of a gene sequence. The exon region is a coding region, meaning that this portion of the DNA codes for amino acids. These areas are not generally used in genetic genealogy.[7]

**Intronic** present in the intron area of a gene sequence. The intron region is a non-coding region, meaning that this portion of the DNA, found between the coding regions, does not have any effect on what amino acids are produced. These areas are typically used in genetic genealogy.[7]

**Minor Allele Frequency** the frequency, expressed as a percentage, of the less frequent allele of a biallelic SNP within a given population.

**Hardy-Weinberg Equilibrium** this theorem states that the amount of genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors. It is defined by the equation:

\[ p^2 + 2pq + q^2 = 1 \]  \hspace{1cm} (2.1)

where \( p \) refers to the frequency of the dominant homozygous genotype, \( q \) refers to the frequency of the recessive homozygous genotype, and \( pq \) refers to the frequency of the heterozygous genotype.

**Triangulation** while the method of triangulation has been around for decades, “genetic triangulation” was first named by Bill Hurst in 2004. It involves a three-way link between three individuals. If Person 1 and Person 2 match, the link between them would form the base of the triangle. Their genealogy goes back in time for each person, forming the two sides of the triangle. These sides converge at the Most Recent Common Ancestor, forming the point of the triangle. If the genetic match forming the base of the triangle is present but the genealogical links to the MRCA is absent, the individuals are genetic matches to an unknown MRCA. If the genetic match forming the base of the triangle is absent but the genealogical links to the MRCA are present, the individuals are genealogical matches but not genetic matches, which potentially alludes to a non-paternal event. In the case of distant relatives who have a MRCA, it is possible that there is no genetic match. A visual representation of triangulation can be seen below in Figure 2.1.[8]

**Most Recent Common Ancestor (MRCA)** when presented with a DNA match, the MRCA is the ancestor from which the two matching individuals received the shared DNA segment or segments.[8]
Non-Paternal Event (NPE) any instance which caused a disconnect between a hereditary surname and the Y-chromosome. Examples include situations such as adoption, formal name change, or infidelity.[8]

Pedigree similar in structure to a family tree, but generally displays genetic inheritance of a particular trait. In genealogy, pedigree is used synonymously with family tree.[7]

Recombination occurs during meiosis. Similar DNA sequences on paired chromosomes cross over each other and results in shuffled genetic material passed down to create genetic variation. A visual representation of recombination can be seen in Figure 2.2[7]

Phasing assigning alleles to paternal or maternal chromosomes. In other words, determining if a particular allele was inherited from an individual’s mother or father.[8]

Segment a continuous section of DNA which is transferred to a person from one parent. Each segment is bounded by a recombination. Such recombinations could have taken place many generations previously.

Centimorgan (cM) a unit of measure to describe probability of genetic linkage. One cM is equivalent to a 1% chance that two genetic markers will end up on different chromosomes
during recombination. The higher the cM value, the higher the probability of genetic linkage, and therefore the closer the familial relationship.

2.1.3 SNP Nomenclature

The following nomenclature has been produced and standardized by the Human Genome Variation Society. An example SNP is named rs3211371 (*5, c1459C >T, R487C). The rs3211371 refers to a reference SNP – a standard SNP given a unique numerical value for easy reference in articles and other research across authors and institutions. The *5 refers to the location on the chromosome where the SNP is located. The c in c1459 indicates the SNP is present within a coding region (exon), and the 1459 refers to the position of the nucleotide where the polymorphism occurs. The C > T shows that a cytosine has been replaced by a tyrosine at this location. Lastly, the R487C can be read as an amino acid substitution at position 487 where an arginine has been replaced by cysteine.

2.1.4 Analysis in GEDmatch

Humans have 46 chromosomes that are paired together, with pairs 1 – 22 being autosomal chromosomes and pair 23 being sex chromosomes. The autosomal chromosomes of an individual are analyzed for genetic genealogy purposes. For each chromosomal pair, one of the chromosomes comes from an individual’s mother and the other comes from the father. The singular chromosomes are some type of combination of their two original chromosomes created through recombination. This continues throughout generations.[9] A depiction of this can be seen in Figure 2.3.

![Figure 2.3: Simplistic Depiction of Chromosome Inheritance and Recombination. Colors used are meant solely as a visual aid to represent sections on the chromosome.[9]](image)

It is important to note that genetic information from each parent is not necessarily passed down in a 50:50 ratio. For example, looking at the chromosome composition for Mother shown in Figure 2.3, the chromosome passed down from Grandmother is approximately 25:75 pink to yellow, while the chromosome passed down from Grandfather is approximately 75:25 orange to light orange. The chromosome transfer seen between Mother and Child in Figure 2.3 displays non-recombination – no chromosome segments from Grandmother (maternal) were passed on to Child due to the lack of recombination. Another important factor is the increased complexity of the chromosomes as more generations are added. This can be seen in the increased number of colors throughout the generations seen in Figure 2.3. As more recombination events occur, more variation takes place that makes determining genetic linkage difficult. For example, 10% of an individual’s third cousins and 50% of an individual’s fourth cousins would not return any matching chromosome segments. Because of these complications, reliable autosomal DNA analysis is typically reserved to looking at
only six or seven generations.[9]

The program used for autosomal DNA analysis for this thesis is GEDmatch, which uses numerical data to report the strength of a relationship between two individuals based on matching chromosome segments. GEDmatch is able to combine data from a wide variety of genealogy testing companies, including Family Tree DNA™[10], AncestryDNA™[2], MyHeritage™[11], and 23andMe™. An example GEDmatch output is pictured below in Figure 2.4.

![Example GEDmatch Output](image)

**Figure 2.4: Example GEDmatch Output**[9]

As seen in Figure 2.4, the blue bars in the bottom visual depict the matching segments on Chromosome 3 between two individuals. There are two matching segments in this particular example: a smaller one on the far left and a larger one towards the right of the visual. The output table, shown in the top left of Figure 2.4 gives more quantitative information about the two matching segments. The start and end location of the segments are identified. The first segment begins at the 36,495th base pair and ends at the 5,168,135th base pair of the chromosome, while the second, longer segment begins at the 104,270,146th base pair and ends at the 168,695,458th base pair. The next value given is centimorgans (cM). The first segment only gives 15.8 cM of DNA information, while the second segment gives 57.2 cM of DNA information; therefore, the second segment is 3.6 times more informative than the first. The last column in the table depicts how many matching SNPs were found within these segments. The first segment had 2,114 matching SNPs and the second had 13,878 matching SNPs.

When analyzing GEDmatch output, matches between two individuals can be categorized into Identical by Descent (IBD) or Identical by State (IBS). IBD matches come from SNPs shared due to a common ancestor, while IBS matches either occur by chance or represent a relationship that cannot be proven. The greater the match between chromosome segments, meaning the higher the cM value and matching SNP counts, the greater the chance the match is IBD. Any segments notably low in these values (i.e. <7 cM or <700 matching SNPs) are likely IBS matches.[9]

Identical by Chance (IBC) is a relatively recent category of match terminology made in an effort to narrow down the number of matches that can be classified as IBS. IBC matches would refer to only those that are proven false positives and are typically found with relatively small cM values. IBS match terminology would then refer to those matches that cannot be proven either way, such as an extremely distant individual that could potentially be a legitimate relative.

The relationship between individuals can be determined based on the overall cM value for all matching chromosome segments. Table 2.1, seen below, lists standard cM values expected from known familial relationships. The general trend with this measure finds that as the distance between the relationship of two individuals increases, the cM value decreases. Figure 2.5 displays a pedigree that demonstrates the variations in degree of relationship in regards to shared DNA segments. The degree of relationship is reported in relation to the self/twin box. Other notable features include that for the types of relationships grouped together for the same Measure of Genetic Linkage in
Table 2.1: Expected Genetic Linkage Values for Known Relationships. While the actual measure of genetic linkage for identical twins is 6800 cM, this value is actually reported as 3400 cM in GEDmatch output.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Measure of Genetic Linkage (cM)</th>
<th>Percent Matching Chromosomes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical Twin</td>
<td>6800</td>
<td>100</td>
</tr>
<tr>
<td>Parents</td>
<td>3400</td>
<td>50</td>
</tr>
<tr>
<td>Siblings</td>
<td>2500</td>
<td>37.5</td>
</tr>
<tr>
<td>Grandparents</td>
<td>1700</td>
<td>25</td>
</tr>
<tr>
<td>Aunts/Uncles</td>
<td>1700</td>
<td>25</td>
</tr>
<tr>
<td>Great-Grandparents</td>
<td>850</td>
<td>12.5</td>
</tr>
<tr>
<td>First Cousins</td>
<td>850</td>
<td>12.5</td>
</tr>
<tr>
<td>Second Cousins</td>
<td>212.5</td>
<td>3.125</td>
</tr>
<tr>
<td>First Cousins Twice Removed</td>
<td>212.5</td>
<td>3.125</td>
</tr>
</tbody>
</table>

![Pedigree Diagram](image)

Figure 2.5: Pedigree demonstrating various degrees of relationship. Note that the “self/twin” box for which the relationship degree is determined is located below the color key in red.[12]

### 2.2 Current Investigative Processes

The current investigative process for forensic genetic genealogy is not one that is well-known to the public, even with its use in almost 200 cold cases and of the occasional active investigations. This ambiguity can be explained from both a business and a law enforcement angle. Commercial interests in these techniques have caused these companies to be tight-lipped about their analytical processes, and law enforcement agencies also avoid being too forthcoming with their protocol, especially considering that these analytical techniques have yet to be reviewed in court.[13,14]

The basic description of the investigative process involves uploading the unknown DNA sample into CODIS and determining if a match can be found. If no match is found and there is sufficient
genetic material available, only in this instance is forensic genetic genealogy utilized. The crime scene sample then undergoes a SNP analysis. The unknown profile thus generated is uploaded, for example, to GEDmatch to identify potential matches and to establish the degree of relatedness between these matches and the suspect profile. From here, a family tree is constructed using both genetic and classical genealogical information to identify the individual who deposited the DNA at the crime scene. This is done by either finding an exact possibility in the family tree or finding a known relative with a closer degree of relatedness and performing kinship testing or having that relative volunteer a name. A DNA analysis is performed on a sample taken from the suspected individual or a sample discarded by the suspected individual and a STR profile is developed. This STR profile is then compared to the original crime scene profile to determine a match. The level of success of an investigator is directly proportional to the closeness of the matches that are found in public databases – the closer the degree of relatedness, the greater the chance a suspected individual can be identified.

2.2.1 Official Guidelines in the United States

In late 2019, the United States Department of Justice (DoJ) released an Interim Policy for Forensic Genetic Genealogical DNA Analysis and Searching that discusses basic guidelines for forensic laboratories completing this type of analysis. The Scientific Working Group on DNA Analysis Methods (SWGDAM) released a similar statement in early 2020. Both organizations stressed the use of CODIS in conjunction with forensic genetic genealogy. An agency must first upload an unknown DNA profile to CODIS and receive no matches before they can analyze the profile through a genealogical approach. If this analysis returns a potential match for the unknown DNA profile and a suspect is announced, the agency must perform STR DNA analysis to create a known DNA profile. This profile must then be compared to the original one uploaded to CODIS. Agencies must also identify themselves as law enforcement to the genetic genealogy companies they approach, and they should also only access databases for companies which have notified their users that their profiles could be utilized by law enforcement. In addition, most circumstances require the agency to receive informed consent from third parties – exclusions to this guideline occur when the integrity of the investigation may be affected if this request takes place.[13,15]

2.2.2 Errors

Two main categories of errors have been noted during previous investigations: technological errors and induced errors. Technological errors can take place throughout amplification and sequencing and can result in incorrect genotype calls. Induced errors can take place during imputation, in which the analyst infers missing data, or during phasing, where chromosomes are accidentally switched when being categorized as maternal or paternal in origin. In 2014, Durand et al. showed 23andMe™ averaged a genotyping error rate of <1% and a phasing error rate of <0.2%.[16] This same study found that AncestryDNA™ had a phasing error rate of 0.64% but suggested that this would decrease as the size of the SNP panels increased. Other suggestions include using a haplotype score after processing to screen for erroneous IBD calls. Parabon[28], another notable name in SNP analysis, has accurately deconvoluted data from mixtures of two individuals when the suspect DNA contributes to at least 40% of the mixture and a profile of the other contributor can be referenced. These analyses have been performed using an Illumina[29] Infinium CytoSNP-850K array that contains almost all of the same SNPs that AncestryDNA™, Family Tree DNA™, and MyHeritage™ use.[13]

As mentioned previously, any IBD matches less than 7 cM are typically deemed unreliable for
genealogical analysis. One study has even reported that known IBD matches of 2-4 cM segments came back with a false-positive rate of over 67%. In an effort to decrease this false-positive rate and increase the reliability and accuracy of these weaker segments, the same study suggests the use of HaploScore, a computational metric that reports IBD segments in relation to the amount of haplotype switch errors they contain. This is a relatively newer area to be explored that has high potential in increasing analytical capability.[16]

2.2.3 Mixtures

The analysis of DNA mixtures for forensic genetic genealogy is still a relatively unexplored field. Two studies have published results stating that they had successfully separated mixed profiles but did not provide specifics on their process. The basic understanding is that for forensic genetic genealogical analysis, a single individual’s DNA profile is needed, making it necessary to separate a mixed profile based on conditioning via known contributors or the use of a statistical model. This, however, leads to uncertainty in the integrity of the profile that should be addressed when discussing the results of analysis. At present, use of mixtures does not seem to be a viable undertaking in genetic genealogy.[13]

2.2.4 Case Studies

One example of a case where forensic genetic genealogy was implemented is a 1987 double homicide that took place in Snohomish County, WA. Tanya Van Cuylenborg and Jay Cook were traveling from British Columbia to Washington State when the couple went missing—Tanya’s body, Jay’s body, and their vehicle were all recovered in separate locations days later. DNA evidence was found on the scene, but no matches were returned in CODIS. After running the profile through GEDmatch, two 5th degree matches were found that had no shared DNA between them. After family tree construction and descendancy research, a triangulation marriage was found between a son of the first match’s great-grandmother and a granddaughter of the second match’s great-grandparents. This makes the resulting children half first cousins once-removed to the first match and second cousins to the second match—both 5th degree relationships. Only one son came from this marriage: William Earl Talbott II. An STR profile was constructed for Talbott from a discarded cup and a match was found with the DNA from the crime scene. Talbott was convicted in 2018.[12]

![Figure 2.6: Simple family tree representing the above case study.]

Another example can be found in the 1986 sexual assault and murder of 12-year-old Michella Welch in Tacoma, WA. Investigators believed that her homicide was connected to that of Jennifer
Bastian, another young girl, but DNA samples left at the scene did not return any results in CODIS. Genetic ancestry predictions suggested the suspect to be predominantly Northern European with approximately 10% Northern Native American, and GEDmatch returned two matches with no shared DNA, meaning that the two matches were on separate branches of the suspect’s family tree. Further analysis identified a pair of brothers living in the area, and investigators were able to build a STR profile based on a DNA sample left on a discarded napkin from one of the brothers, Gary Charles Hartman. This profile matched the DNA found at Michella Welch’s scene – Hartman was arrested and was convicted in 2022.

In instances where not enough individual information is able to be learned from forensic genetic genealogy, analysis can be combined with other methods such as kinship testing. A DNA sample from a 40-year-old homicide case returned two matches in the 6th-8th degree range in GEDmatch, but family tree construction was unable to find an intersection. It was believed that the match could be a great- or great-great-grandson to one of the first match’s great-great-grandparents, but this was all the information that could be determined at the time. Investigators then obtained a buccal swab from a cousin on the first match’s paternal side, which came back as unrelated to the suspect, but a buccal swab from a cousin on the first match’s maternal side came back as a 3rd degree relative to the suspect. Based on this information, family trees were constructed, and the suspect was determined to be a son of Match #1’s maternal uncle. After determining one of the sons to have a DNA profile matching that found at the crime scene, the son was arrested in 2019 and is awaiting trial.

2.3 Ethics

While terms and regulations of AncestryDNA™ and 23andMe™ do not allow for law enforcement access and MyHeritage™ and Family Tree DNA™ require company permission and appropriate legal documentation, GEDmatch has always marketed itself as a public database since its creation in 2010. Any data uploaded that has been set to “public” by the user can be accessed by anyone who uses the site, including law enforcement. After data on GEDmatch was used in an effort to identify the Golden State Killer, terms and conditions were exchanged to explicitly inform users that this situation was a possibility, and any new or recurring users of the site were required to consent to this information before starting or continuing use. In theory, any data available for public use has been approved by the individual who uploaded it from a private genetic genealogy company after reading the terms and conditions and chose to make it “public” rather than “private” on the GEDmatch database.

One potential ethical concern to be addressed is the revealing of sensitive genetic information, such as health issues caused by genetic predispositions, to law enforcement during the course of the investigation. This unease can be put to rest due to the manner in which the data is communicated to law enforcement. While information regarding health issues can be found in the raw DNA data submitted to GEDmatch, no raw genotypes are output by the program. Output consists solely of comparisons to determine relationships between individuals and never includes the raw data.

Another ethical concern with forensic genetic genealogy stems from whether or not an individual has an ability to truly “opt out” of participating if a family member has chosen to submit their DNA sample. Take, for example, a family with four siblings. Three of the four siblings have decided to upload their DNA profile to GEDmatch and make their information public and searchable, but the fourth wants nothing to do with it and refuses to participate. In this instance, a large portion of that fourth sibling’s DNA is now accessible through his three siblings, even without his consent.
and active participation.

Post-mortem privacy is also an ethical factor that must be considered. Although post-mortem privacy is not currently protected by law, it is possible to extract DNA samples from the deceased and their belongings. While these profiles would be extremely helpful in filling out family trees and other genealogical research, there is a question as to its use for law enforcement. The deceased individual cannot consent and their descendants may refuse to give law enforcement such permissions.[13]

One public concern is the risk of reidentification from a DNA profile used over the course of an investigation. In response, studies have been conducted on the likelihood of reidentification based on different scenarios and profile characteristics. Based on potential surname information and physical characteristic data obtained through these sites, it is entirely possible to determine more individual-specific information by cross-checking theseis data with public data records, such as voter registration records. In clinical research, straightforward identifiable information is kept private under the Health Insurance Portability and Accountability Act (HIPAA), but this document does not apply to GEDmatch. Possible solutions to this problem lie in genomic encryption or cryptographic data sharing models.[18]

Another factor to consider is the ability of minors to submit DNA to genealogy databases. Family Tree DNA™ in particular allows users as young as 13 to submit samples, but most companies state that a minor can submit as long as parental permission is obtained. In reality, there is no way to police these sites to screen out contributions of minors, and profiles located through GEDmatch do not include any personal identifying information, such as age. This raises an ethical issue due to the inability of said minors to properly consent to their profiles being used from an investigative standpoint, as well as the inability of investigators to determine whether or not the profile they are utilizing is that of a minor.[19]

Despite potential ethical complications, the general image of forensic genetic genealogy seems to be positive. A 2018 survey resulted in significant public backing of the technique, and GEDmatch reported a substantial rise in public participation once the Golden State Killer case was resolved. However, many participants admitted to having ethical concerns with the method that they would like to see addressed before regular implementation in law enforcement practices.[13,20]

### 2.4 Legal

A potential legal issue that arises from forensic genetic genealogy is that of investigatory privilege. As it stands, federal law prevents law enforcements’ release of case-pertinent information in a manner that would harm an investigation. An example of this could include specific details that would make the identities of involved parties public knowledge, causing suspects to flee or evidence to be tampered with. This privilege could result in some complications with the forensic genetic genealogical investigative process. In instances where kinship testing or asking a relative to share a name is necessary to identify a suspect, one must consider if revealing to an individual that a family member is potentially involved in a crime would fall under investigatory privilege.[21]

Investigatory privilege also goes hand-in-hand with another popular legal concept, “fruit of the poisonous tree.” At face value, this doctrine refers to the inadmissibility of evidence if it originated as a result of illegally obtained evidence. Originally established in 1920 as a result of Silverthorne Lumber Co. v United States, there are three exceptions to this concept. Evidence is still admissible if:
its source was not part of the illegal activity
its discovery was inevitable, or
attenuation occurs between the illegal activity and the evidence discovery.

The “fruit of the poisonous tree” becomes problematic if, especially under investigatory privilege, the act of obtaining a DNA profile or legal name is technically prohibited. If that information is, in theory, illegal for investigators to have received, the DNA profile of the suspect obtained through forensic genetic genealogy could be inadmissible, leaving investigators without a critical aspect in securing a conviction.[22]

It is worth noting that there are no published standards on forensic genetic genealogy, nor has the method ever been scrutinized independently in court as a forensic analysis method. As the method gains popularity, it will be interesting to see if it becomes generally accepted.
3. Methodology and Results

DNA samples in the form of buccal swabs were submitted to AncestryDNA™, 23andMe™, MyHeritage™, and Family Tree DNA™ (FTDNA™) for processing by each company’s respective SNP genotyping assay. AncestryDNA™ and 23andMe™ samples were submitted as saliva samples, while samples sent to MyHeritage™ and FTDNA™ were buccal swabs. Each company’s DNA submission kit contained step-by-step instructions to ensure samples were collected correctly. The samples, whether saliva or buccal swab, were encapsulated in a company-specific preservation additive to slow bacteria growth and general sample degradation during transport. Excluding AncestryDNA™, which uses a customized Illumina OmniExpress Assay, the DTC companies in this project each use a customized Illumina Global Screening Array to develop a profile. This array is comprised of a minimum of 654,027 SNPs with the ability to add up to an additional 100,000 custom SNP markers. Both of these assays are designed to genotype a sample using high-density oligonucleotide SNP microarrays, where a small chip contains repeating hybridization probes in order to determine specific alleles. A genotype profile for the suspect was completed and accessible approximately four to six months after initial submission from each company. This turnaround time was reported to be above average from the four companies, with typical expected turnaround time to be closer to two to three months. In total, over 37,000 DNA matches were returned to the unknown profile: 13,905 from MyHeritage™, 17,058 from AncestryDNA™, 4,815 from FTDNA™, and 1,484 from 23andMe™.

Figure 3.1: Illumina Global Screening Array[25]
Figure 3.2: SNP Microarray. Measurement of relative intensity of excited nucleotide labels provides locus-specific ratios of alleles[26]

The resulting profiles from genotyping contained a wealth of information, such as ancestry reports and genetic predispositions, but the data of use to an investigative genetic genealogist is the location of SNPs on a given chromosome. This positioning can be compared to potential relatives to see what grouping of SNPs overlaps between the two individuals – in other words, if any “shared DNA” is present. The greater the amount of shared DNA, measured in centimorgans (cM), the closer the relationship between the individuals. For this reason, the profiles were uploaded to GEDmatch. All four profiles were set as public and allowed for law enforcement access.

This project primarily utilized GEDmatch’s One to Many tool. The tool allows the user to compare, in this case, a suspect profile to other uploaded profiles. In order to return as a search result, profiles must be both public and authorized for law enforcement use, a setting controlled by the profile’s owner. Search results highlight the categories of data seen in Figure 3.3 and can be sorted based on user-selected variables. Parameters for search results included a minimum value of 7 cM and a maximum overlap value of 45,000. For this research, results were filtered based on total cM for autosomal DNA. A candidate list with which to start the process was created. This list, seen below in Table 3.1, was comprised of DNA matches that were considered “searchable” – had real names associated with the profiles and were provided from one of the four DTC companies used in this research. Table 3.1 does not include the Kit Number, Name and Email of the matches. This information was removed to maintain anonymity. Instead, each match was given a two-letter alias to which the match will be referred to in this thesis. Matches were ranked starting with the closest relative and then in descending order. The closest possible relative had a comparable cM value to a second cousin, while the last match included had a comparable cM value to a sixth cousin. An additional GEDmatch application, People who match both, or 1 of 2 kits, was used to verify the candidate list further. In this instance, a match was consistent with a 10 cM threshold. This tool was used to determine if a candidate had at least two matches from two different kits to the suspect. The candidates pictured in Table 3.1 all met this criteria, excluding Match 10, JK. However, JK was kept in the Candidate List since JK and JW (Match 1) had matching DNA segments.
Figure 3.3: GEDmatch One to Many categories. Search parameters allow for the user to set cutoff values for cM and overlap results.

Table 3.1: Candidate List. Top ten searchable candidates and additional thirteenth candidate provided from GEDmatch. Aliases have been provided to ensure confidentiality.

<table>
<thead>
<tr>
<th>Match #</th>
<th>Name</th>
<th>Autosomal</th>
<th>X-DNA</th>
<th>Source</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total cM</td>
<td>Largest</td>
<td>Total cM</td>
<td>Largest</td>
</tr>
<tr>
<td>1</td>
<td>JW</td>
<td>182.3</td>
<td>37.5</td>
<td>3.15</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>MC</td>
<td>62.7</td>
<td>31.4</td>
<td>3.92</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>CX</td>
<td>51.9</td>
<td>51.9</td>
<td>4.06</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>RW</td>
<td>44.8</td>
<td>23.9</td>
<td>4.16</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>SM</td>
<td>44.6</td>
<td>11.5</td>
<td>4.17</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>AM</td>
<td>43.7</td>
<td>13.8</td>
<td>4.18</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>PA</td>
<td>43.6</td>
<td>10.2</td>
<td>4.18</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>PH</td>
<td>42.5</td>
<td>15.8</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>RB</td>
<td>42.4</td>
<td>10.3</td>
<td>4.20</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>JK</td>
<td>42.1</td>
<td>10.2</td>
<td>4.21</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>GM</td>
<td>37.7</td>
<td>30.2</td>
<td>4.29</td>
<td>0</td>
</tr>
</tbody>
</table>

The One to Many tool does not automatically filter out other profiles submitted by the user. In this instance, there were four different profiles submitted for this project – one from each company. When the search was performed, the first three closest matches were actually the suspect profiles returned from the other three companies, as seen in Table 3.2. It is interesting to note that there is a slight variation in the amount of shared DNA – the cM value is not consistent. This is due to the aforementioned unique customization of the arrays used by each company. Because the same genetic markers are not consistent, all four arrays cannot return the exact same genetic profile. There was very limited intra-variability between the suspect profiles, which is not expected when observing the wide variety in quantity of SNPs tested (Tables 3.3 – 3.6). The primary focus is the comparison of quantity of autosomal SNPs tested (Table 3.3). Differences in quantity range from approximately 7,500 to 54,000 SNPs. However, DNA segment data were relatively consistent, as seen in Table 3.2. Despite some small fluctuation in the total cM value, all returned values were still equivalent to a self-match when compared to the other three profiles. Every profile returned the same cM value for the largest segment (151.8). The profile returned from AncestryDNA™ had a notably high variation in the total cM value for X-DNA, likely due to the use of a different assay than the other three companies. However, since comparisons were performed with autosomal SNPs, this did not affect the investigative process.
Table 3.2: Top Three Initial GEDmatch Results Belonging to Suspect.

<table>
<thead>
<tr>
<th>Autosomal X-DNA</th>
<th>Total cM</th>
<th>Largest Gen</th>
<th>Total cM</th>
<th>Largest Source</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>3580</td>
<td>151.8</td>
<td>1.00</td>
<td>187.3</td>
<td>104.6 FTDNA™</td>
<td>337568</td>
</tr>
<tr>
<td>3577.9</td>
<td>151.8</td>
<td>1.00</td>
<td>187.3</td>
<td>104.6 MyHeritage™</td>
<td>332521</td>
</tr>
<tr>
<td>3571.1</td>
<td>151.8</td>
<td>1.00</td>
<td>157.4</td>
<td>082.6 AncestryDNA™</td>
<td>96465</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of Autosomal SNPs Tested Across the Four Companies[27]

<table>
<thead>
<tr>
<th>23andMe™</th>
<th>FTDNA™</th>
<th>AncestryDNA™</th>
<th>MyHeritage™</th>
</tr>
</thead>
<tbody>
<tr>
<td>17,860</td>
<td>7,507</td>
<td>25,367</td>
<td>123,975</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison of Y Chromosome SNPs Tested Across the Four Companies[27]

<table>
<thead>
<tr>
<th>23andMe™</th>
<th>FTDNA™</th>
<th>AncestryDNA™</th>
<th>MyHeritage™</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,733</td>
<td>2,042</td>
<td>1,691</td>
<td>3,495</td>
</tr>
</tbody>
</table>

Table 3.5: Comparison of X Chromosome SNPs Tested Across the Four Companies[27]

<table>
<thead>
<tr>
<th>23andMe™</th>
<th>FTDNA™</th>
<th>AncestryDNA™</th>
<th>MyHeritage™</th>
</tr>
</thead>
<tbody>
<tr>
<td>259</td>
<td>12,802</td>
<td>13,061</td>
<td>13,164</td>
</tr>
</tbody>
</table>

Table 3.6: Comparison of Mitochondrial SNPs Tested Across the Four Companies[27]

<table>
<thead>
<tr>
<th>23andMe™</th>
<th>FTDNA™</th>
<th>AncestryDNA™</th>
<th>MyHeritage™</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,139</td>
<td>4,055</td>
<td>84</td>
<td>43,18</td>
</tr>
</tbody>
</table>

For this same reason, there was slight variation in the cM values of resulting matches when using the One to Many tool with, for example, the suspect profile from 23andMe™ and the suspect profile from AncestryDNA™. However, although the values were slightly different, the same individuals returned when using the tool with all four profiles – this was taken into account when creating the searchable candidate list.

The closest match, a profile originally submitted to 23andMe™ by an individual who will be referred to in this thesis as JW. This match was identified to have a 182.3 cM overlap with the suspect profile. This segment data can be seen in Table 3.7 and Figure [3.3]. There were eight significant matching segments of varying lengths and 176.3 cM of half-match segments. GEDmatch estimated
that this match was approximately 3.2 generations away from the MRCA. This is consistent with the possible familial relationships seen in Table 3.8 and Figure 3.7.

Table 3.7: Matching Segments between Suspect and JW. B37 refers to the human genome sequence reference used by GEDmatch

<table>
<thead>
<tr>
<th>Chr</th>
<th>B37 Start Pos’n</th>
<th>B37 End Pos’n</th>
<th>Centimorgans (cM)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>217,453,939</td>
<td>238,505,096</td>
<td>36</td>
<td>996</td>
</tr>
<tr>
<td>7</td>
<td>44,935</td>
<td>8,532,844</td>
<td>14.8</td>
<td>464</td>
</tr>
<tr>
<td>8</td>
<td>12,633,235</td>
<td>31,986,192</td>
<td>31.1</td>
<td>1,036</td>
</tr>
<tr>
<td>15</td>
<td>87,403,331</td>
<td>92,477,878</td>
<td>9.5</td>
<td>228</td>
</tr>
<tr>
<td>17</td>
<td>8,927,174</td>
<td>15,338,269</td>
<td>23.5</td>
<td>408</td>
</tr>
<tr>
<td>21</td>
<td>19,395,457</td>
<td>39,341,560</td>
<td>33.9</td>
<td>877</td>
</tr>
<tr>
<td>21</td>
<td>42,384,938</td>
<td>45,621,817</td>
<td>9.8</td>
<td>209</td>
</tr>
<tr>
<td>22</td>
<td>16,055,122</td>
<td>22,599,216</td>
<td>17.6</td>
<td>272</td>
</tr>
</tbody>
</table>

Figure 3.4: Matching Segments between Suspect and JW on Chromosomes 2, 7, 8, 15, 17, 21, and 22. Each chromosome is represented by a set of two color-coded horizontal bars, where appearances of colors are indicate relative base pair location on the chromosome. Color key defined in Figure 3.5.
Figure 3.5: Color Distinction Key for Matching Chromosome Segments.

<table>
<thead>
<tr>
<th>Top Bar of Each Chromosome - Base Pairs</th>
<th>Full Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>Half Match</td>
</tr>
<tr>
<td>Red</td>
<td>No Match</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bottom Bar of Each Chromosome - Match Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
</tr>
<tr>
<td>No Match</td>
</tr>
<tr>
<td>Tan</td>
</tr>
<tr>
<td>Large Gap Between Adjacent SNPs</td>
</tr>
<tr>
<td>Purple</td>
</tr>
<tr>
<td>SNP Density Ratio (less than 0.12)</td>
</tr>
</tbody>
</table>

Figure 3.6: Relationship Probabilities for Suspect and JW. These probabilities were provided by DNAPainter’s Shared cM Tool.[30]

<table>
<thead>
<tr>
<th>50%</th>
<th>2GC</th>
<th>2C1R</th>
<th>Half 1C2R</th>
<th>1G3R</th>
</tr>
</thead>
<tbody>
<tr>
<td>38%</td>
<td>Half GG-Aunt / Uncle 2C</td>
<td>Half 1C1R</td>
<td>1C2R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Half GG-Niece / Nephew</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9%</td>
<td>Half 1C3R</td>
<td>1G3R</td>
<td>1C</td>
<td>2C</td>
</tr>
<tr>
<td>2%</td>
<td>Great-Great-Aunt / Uncle 1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Great-Great-Niece / Nephew 1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Half Great-Aunt / Uncle 1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Half Great-Niece / Nephew 1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Half 1C</td>
<td>1C1R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.79%</td>
<td>2C3R</td>
<td>1C2R</td>
<td>Half 3C</td>
<td>3C1R</td>
</tr>
</tbody>
</table>

† This relationship has a positive probability for 182.3 cM in the table’s range of probabilities, but falls outside the bounds of the recorded cM range (99th percentile).

Figure 3.7: Possible Relationships of Suspect to JW. Any faded relationships are not considered a possibility. The values below the relationship indicates the average cM value and the expected cM range for that relationship. Provided by DNAPainter’s Shared cM Tool.[30]
After identifying the closest match, the next step in the process was to perform genealogical research. This was completed using the AncestryDNA™ Library Edition Database available through West Virginia University Libraries. This resource provides access to census records, birth certificates, marriage licenses, obituaries, and other documents that can be used to identify generations worth of related individuals. Some of these records have been digitized, while others are simply scans of the original documentation. Additional resources include information sourced from school yearbooks and community-based websites such as Find a Grave. Relevant information was used to construct a family tree for JW. This tree was comprised of direct relatives of JW, spanning up to ten generations in certain family lines with ancestors dating back to the seventeenth century.

This process of identifying a DNA match and building the relevant family trees was repeated multiple times throughout the descendency research step of IGG until a “union couple,” or MRCA, was found. While all matches are related in some form to the suspect, there is not always a relation between two matches that allows the researcher to link two family trees. In addition, the genealogy records necessary to link two individuals may not be publicized, searchable, or otherwise accessible. For this thesis, the union couple was identified after beginning to build out the family tree of one AncestryDNA™ DNA match which will be referred to as GM. This match was identified to have a 37.7 cM overlap with the suspect profile. This segment data can be seen in Table 3.9 and Figure 3.8. There were two significant matching segments of varying lengths and 37.6 cM of half-match segments. GEDmatch estimated that this match was approximately 4.5 generations away from the MRCA. This is consistent with the possible familial relationships seen in Table 3.10 and Figure 3.10. GM’s family tree was comprised of their direct relatives, spanning a maximum of eleven generations in certain family lines with ancestors dating back to the sixteenth century.

<table>
<thead>
<tr>
<th>Chr</th>
<th>B37 Start Pos’n</th>
<th>B37 End Pos’n</th>
<th>Centimorgans (cM)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>127,927,990</td>
<td>133,728,225</td>
<td>7.5</td>
<td>233</td>
</tr>
<tr>
<td>6</td>
<td>3,836,353</td>
<td>18,439,027</td>
<td>30.2</td>
<td>1,050</td>
</tr>
</tbody>
</table>

**Figure 3.8**: Matching Segments between Suspect and GM on Chromosomes 2 and 6. Each chromosome is represented by a set of two color-coded horizontal bars, where appearances of colors are indicate relative base pair location on the chromosome. Color key defined in Figure 3.5.
Figure 3.9: Relationship Probabilities for Suspect and GM. These probabilities were provided by DNAPainter’s Shared cM Tool. [30]

<table>
<thead>
<tr>
<th>Family Relationship</th>
<th>Probability</th>
<th>cM Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half-Grand-Aunt/ Uncle</td>
<td>50%</td>
<td>5C3R</td>
</tr>
<tr>
<td>Grandparent</td>
<td>19%</td>
<td>Half-3C</td>
</tr>
<tr>
<td>Parent</td>
<td>18%</td>
<td>4C</td>
</tr>
<tr>
<td>Aunt/Unclce</td>
<td>10%</td>
<td>3C</td>
</tr>
<tr>
<td>Self</td>
<td>3%</td>
<td>Half-2C</td>
</tr>
<tr>
<td>0%</td>
<td><strong>1C2R</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** This set of relationships is just within the threshold for 37.7cM, but has a zero probability in the phage’s table of probabilities.

† This relationship has a positive probability for 37.7cM in the phage’s table of probabilities, but falls outside the bounds of the recorded cM range (99th percentile).

GM was able to be linked to JW through the identification of the “union couple,” or two individuals whose marriage/union combines two otherwise unrelated families. The union couple, comprised of JL and MH, will be referred to collectively as JL/MH. GM and JL are second cousins once removed, and JW and MH are first cousins once removed. GM and JW themselves are not related to each other (Figure 3.15).

GM and JL are four generations away from each other in the combined family tree. This aligns with the predicted 4.5-gen relationship provided by GEDmatch. JW and MH are three generations away from each other in the combined family tree. This aligns with the predicted 3.2-gen relationship.
Once the union couple has been identified, the next step in the investigative process is to build down their family tree until it is up to date with the year in which the crime had been committed. Building this family tree involves returning to genealogy documents to identify any individuals directly descending from the union couple. Typically, any closely related descendants within the same generation are considered possible suspects, depending on certain characteristics (e.g., age) of the individuals at the time of the crime. In this instance, there is only one direct descendant of JL/MH, so a filtered list of multiple suspects could not be created (Figure 3.11). This suspect, SH, is five generations away from GM and four generations away from JW. SH and GM are third cousins and SH and JW are second cousins once removed. Triangulation depicting how JL/MH, SH, GM, and JW are related is visualized in Figure 3.13 - Figure 3.15.

![Diagram](image)

**Figure 3.11:** Relationship between Union Couple (JL/MH) and SH (denoted in figure as Doe, John-Jane).

![Table](image)

**Figure 3.12:** Key for Figures 3.8 - 3.11
Figure 3.13: Relationship between Union Couple (JL/MH), JW and SH (denoted in figure as Doe, John-Jane).
Figure 3.14: Relationship between Union Couple (JL/MH), GM and SH (denoted in figure as Doe, John-Jane).
Figure 3.15: Relationship between Union Couple (JL/MH), JW, GM, and SH (denoted in figure as Doe, John-Jane).
Initial searches for JW did not return any leads. However, further investigation using 1994-2019 U.S. Index to Public Records and 1993-2002 U.S. Phone and Address Directories, as well as present-day state address directories, determined that a name change took place through marriage. From here, now that the original name was established, this original name could be linked to a document from the West Virginia, U.S., Marriages Index from 1785-1971. This 1959 document not only contains the names of JW’s parents, but it also lists the names of both JW’s maternal and paternal grandparents. There is not much clear information on JW’s paternal genealogical trail beyond this point in time. This could be indicative of a name change or other event that led to a break in documentation. Alternatively, on JW’s maternal side, many documents could be found for JW’s parent GG. In the 1910 U.S. Federal Census, GG, living with their parents, was listed as one of many children of DG and AG. DG and AG were later established to be the union couple between SH and JW, specifically. Since this could not be determined at the time of the investigative process, family trees were continued to be built with the same document-finding process until records could no longer be established, leading to the late 1600s in six generations. This included documents such as those from the U.S., Find a Grave Index (1600s-current), the 1850 U.S. Federal Census, the 1567-1945 Germany, Lutheran Baptisms, Marriages, and Burials, and the 1560-1900 U.S. and International Marriage Records.

Moving forward in time from DG and AG, most of their children only have one or few associated birth records – none of which result in individuals that would be alive in the present day. One of the exceptions, and the cause of a major hiccup in the investigative process, is LG, GG’s sibling. LG goes on to marry AD, documented in both the 1930 and 1940 U.S. Federal Census. However, beyond having children of their own, both of these documents list that two siblings, listed as the child of either LG or AD’s sibling, are living in their household. Additionally, these children, RH and JGH, are not listed in LG and AD’s household in the 1920 Federal Census. This took an even more confusing turn when a card from the U.S., World War II Draft Cards Young Men, 1940-1947 Index, documenting JGH’s enlistment, depicts LG as JGH’s parent, not an aunt or uncle. At this point in time, a plausible working theory was created, backed up by suitable documentation. VG, one of LG’s siblings later marries LH about a decade later, which was determined using the 1920 U.S. Federal Census and a document from the West Virginia, U.S. Marriages Index from 1785-1971. In both the 1920 Census and a document from the 1936-2007 U.S. Social Security Applications and Claims Index, VG and LH are recorded as having a child: RH. Therefore, it can be assumed that JGH, the sibling of RH, is also the child of VG and LH. It is unclear how RH and JGH came into the custody of LG and AD. No reliable records could be located for either VG or LH beyond 1920. JGH is later included on a document from the 1908-1998 Ohio, U.S., Birth Index. This document depicts the birth of MH. MH and JL later have a child, SH, also documented in the 1908-1998 Ohio, U.S., Birth Index. Address records found in the 1994-2019 U.S., Index to Public Records list an address for SH as current as 2020.

The family tree building process described above was repeated for Matches 2-13, as listed in Table 3.1. The time and effort it took to build these clusters varied dramatically between the twelve investigative repetitions, as some individuals’ ancestors were frequently found in genealogy records, while others were rarely mentioned. If previous family members had made significant, cross-country moves, linking documentation was harder than if the family had state mostly in one state. Certain matches had very large families, creating more family tree branches, while others included a much smaller number of individuals. None of the family trees built for Matches 2-12 could be linked to the tree built for JW. Match 13, however, not only could be associated with JW’s family tree, but was linked in such a way that it drastically narrowed the list of potential suspects.
GM is the child of DM and CJ, a married couple established by a document from the Ohio, U.S., Marriage Abstracts of 1970 and 1972-2007. In the Ohio, U.S. Birth Index from 1908-1998, DM is listed as the child of GM and KM. KM, formerly KK, can be found on the 1950 U.S. Federal Census as one of three children living with their parents, IK and EK. EK, formerly EE, is listed on the 1920 U.S. Federal Census as the child of WE and LE. WE and LE were later established to be the union couple between SH and GM, specifically. Just like for JW, family trees were continued to be built with the same document-finding process until records could no longer be established, leading to the 1500s in twelve generations. This included documents such as those from the 1558-1929 Germany, Select Marriages and 1558-1898 Select Births and Baptisms Indexes, the 1870 U.S. Federal Census, and 1800-1970 New Jersey, U.S., United Methodist Church Records.

In the 1940 U.S. Federal Census, EE’s sibling WE Jr is listed. WE Jr goes on to marry AS, which is documented by the 1774-1973 Ohio, U.S., County Marriage Records. They later have two children: HE and GE, who are recorded on the 1950 U.S. Federal Census. There are no other records of GE that were found, but HE’s obituary appears in the 1800s-current U.S., Newspapers.com Obituary Index. This obituary lists JoL as a spouse and HL and JL as their two children. At the time of the obituary publication, JL has previously married MH, depicted by the 1970, 1972-2007 Ohio, U.S. Marriage Abstracts. The appearance of JL and MH in GM’s family tree provides the necessary link to JW’s family tree, making them the union couple for the combined tree. As mentioned previously, JL and MH only have one child, making SH the primary candidate for further investigation.

Once a cluster of possible suspects has been identified, further steps can be taken to increase investigator confidence in the resulting names and familial relationships. One option is to create an all-encompassing family tree based around the suspect(s). This extends beyond the scope of the trees created during the pre-identification steps and focuses more heavily on ancestral relatives rather than descendant ones. Creating this tree provides the investigator with a larger understanding of a family background, can confirm or deny the presence of certain ancestral traits in potential suspects, and can, more broadly, be used to ensure the correct inferences have been made. For example, if further genealogical research displays discrepancies with previously determined information, such as a family line ending in one century, but the suspect(s) have an ancestor from the line years later, this could be a sign that an error has been made. If all the pieces fit together to form a coherent family tree, this increases the strength of the investigative lead. An example of this combined tree can be seen in Figure 3.16. This tree includes the smaller trees built around JW and GM to create a larger, more detailed, tree for SH. This tree includes 191 people, 97 unique families and goes back a maximum of thirteen generations in certain family lines. Names of individuals have been obscured to protect confidentiality.
Another additional step is chromosome mapping for each potential suspect. This process involves determining which segments of an individual’s DNA come from which ancestral relative(s). For this research, chromosome mapping was completed using DNA Painter. Segment data was pulled from the 19,000+ DNA matches returned from 23andMe™, MyHeritage™, Family Tree DNA™, and GEDmatch. Matches from AncestryDNA™ were not used as a part of this process, as segment data cannot be accessed from this website. Initial uploads of DNA segments were performed manually. However, to maximize productivity and minimize investigative turnaround time, DNA Painter’s mass import tool (available with subscription) was used to upload the remaining segment data. Certain parameters were set to avoid attributing IBC segments as actual matches, as this could lead to errors in chromosome mapping, phasing, and suspect identification. For example, to limit this possibility, individual segments less than 10 cM were excluded, as well as matches with a total of less than 15 cM across multiple segments. During the manual upload process, each match was categorized as either an overlap, an unknown, or attributed to a specific ancestral couple. An overlap match meant that the individual contributed a DNA segment at the same chromosome location as another DNA match, while an unknown meant it was a unique location. Matches were only attributed to a specific ancestral couple if there were multiple instances of definite IBD matches - this did not occur frequently. An example of this can be seen in Figures 3.17–3.20.
Figure 3.17: To begin chromosome mapping, DNA segment data is uploaded into DNA Painter.

Figure 3.18: The DNA segments are mapped to the corresponding locations on the appropriate chromosomes.

Figure 3.19: On Chromosome 2, Match A has shares a 996 SNP DNA segment that starts at base pair 217,453,939 and ends at base pair 238,505,096.
Figure 3.20: This process is repeated for all matches for which segment data is available. In this example, Match B also shares a segment with the individual on Chromosome 2, starting at base pair 152,022,950 and ending at base pair 161,220,126.

Once mass data import began, automated uploads were manually verified as being likely IBD matches to remain in the chromosome map. In instances with duplicate matches, such as individuals who had submitted DNA samples to multiple companies, excess matching segments were removed. After all matches had been uploaded and manually verified, a total of 5,423 segments remained. The process of building this map, as well as the version with all matching segments, can be seen in Figure 3.21 – Figure 3.24. Names of individuals have been obscured. When viewing this map in DNA Painter, each of these chromosomes can be expanded to view the contributor names, locations, sizes, and cM values relative to each chromosome.

Figure 3.21: Chromosome Map Comprised of Segment Data from GEDmatch.
Figure 3.22: Chromosome Map Comprised of Segment Data from GEDmatch and 23andMe™.

Figure 3.23: Chromosome Map Comprised of Segment Data from GEDmatch, 23andMe™, and FTDNA™.
The chromosome map can be specified further by phasing, which refers to assigning the matching segments as coming from either a paternal or maternal ancestor. Based on the family tree constructed for SH, JW was identified as being a paternal relative. Therefore, comparing the DNA segments from JW to those who have segments at the same chromosomal positions allowed for the assignment of these relatives as well. The segment comparison tools on GEDmatch and 23andMe were used to complete this process. If segments from different individuals were consistent at the same position, then that match was classified as a paternal relative. Conversely, if the two segments were not a DNA match but were in the same chromosome position, then the other individual must be a maternal relative. An example of this can be seen in Figures 3.25 and 3.26. As additional relatives were assigned as maternal or paternal, these matches could be used to compare to other individuals that had matching DNA segments at different locations than JW. This method was repeated until the chromosome map was phased as completely as possible at 55%. Those matches that were included from Family Tree DNA and MyHeritage could not be used for this portion of the research process, but, if possible, would contribute to a much higher completion percentage. The data from these companies were unable to be used since a point of comparison did not exist. For example, 23andMe match data was able to be used since a paternal match, JW, had been established that segment data to which an unphased match could be compared. Since a point of comparison could not be established, phasing could not be completed for these matches. The phased chromosome map can be seen in Figure 3.27. In this final map, there are areas in which segment information is missing, such as on chromosomes 13 and 22. Some of these locations did

Figure 3.24: Chromosome Map Comprised of Segment Data from GEDmatch, 23andMe, FTDNA, and MyHeritage.
have segment data present, but because the matches were less than 7cM, they were not included. This resulted in what appears to be a less-than-complete phasing percentage, but the data was phased to the highest possible extent in regards to the confines of this research.

Figure 3.25: In this example, Matches A and C both share DNA segments with the individual on Chromosome 21. There is a significant overlap between this segments that consists of 11,546,542 base pairs.
Figure 3.26: Since both matches were retrieved from 23andMe™, the company’s Advanced DNA Comparison tool can be used to compare the two. Even though both segments are in approximately the same location, there is no DNA overlap between them. Therefore, they can be attributed to different lines of the individual’s ancestry (i.e. maternal or paternal).
Figure 3.27: Phased Chromosome Map. Green indicates maternal segments and maroon indicates paternal segments.
4. Discussion

A summary of the methodology determined through this research can be seen in Figure 4.1. It also includes the surrounding investigative procedures that are not IGG specific. It is important to note just how non-linear the process is. Even in a completely ideal, straight-forward situation, the investigator would still be cycling back and forth between finding matches and building out their family trees, doing this as many times as necessary to find a union. As the process goes on, and union descendants are identified, the investigator might realize there is not enough information available for this set of family members or that it leads to a dead end. In this instance, another union couple from a different combination would be necessary. If a list of five potential candidates is formed, that could lead to possibly five different DNA collection scenarios. If the DNA sample is a match to the crime scene sample, law enforcement has gotten their answer. If not? If the DNA sample is similar to the crime scene sample, it is likely a relative, which lets law enforcement know they are on the right track. If it is completely wrong, however, this would lead the investigator to start the entire process over again, backtracking all the way to finding additional public use matches.

![Figure 4.1: IGG Method Summary](image-url)

Because of the design of this research, the identity of the suspect was known from the beginning, but that did not guarantee that the investigative process would necessarily yield that answer. The process resulted in the identification of a single potential suspect. Typically, a list of multiple potential suspects would be generated, but in this instance, the union couple only had one descendant.
The greater the familial distance between the union couple and the generation(s) in the right age bracket to have committed the crime, the larger the list of possible suspects. An example of this is provided in Figure 4.2. Let’s say the estimated age of the person of interest falls within Generation 3. If the union couple is identified in Generation 1, like in the tree on the left, there are four potential candidates for the person of interest in Generation 3. If the union couple is identified in Generation 2, like in the tree on the right, there are only two potential candidates.

![Figure 4.2: The Farther Back the Union Couple, the Greater the Candidate List.](image)

Now that a potential suspect has been identified, what are the next steps? It is important to remember that this entire process has simply found an investigative lead and has not positively identified the suspect. At this point, law enforcement must investigate the person(s) named by IGG and legally obtain a DNA sample. This is typically done by collecting a discarded DNA sample from the person of interest. STR analysis is performed and the sample is compared to the one found at the crime scene. If there is not a match, law enforcement continues down the list generated from IGG. If there is a match, this is grounds for an arrest warrant, leading to suspect interrogation and potentially culminating in criminal court.

The workflow of this particular research shows that an investigator cannot necessarily complete the process with only the first two closest matches. As discussed previously, a union couple could not be identified until the first match was compared to the thirteenth match. Instances such as this are the results of breaks in family lines, such as adoptions or misattributed parentage, inability to access the needed genealogical records, or immigration/emigration.

There are certain assumptions that are made during the IGG process. To narrow down the list of possible suspects, law enforcement must use the nature of the crime to infer characteristic information about the perpetrator. For example, a second-floor break-in was likely committed by an able-bodied individual, so an elderly relative could likely be removed from the potential suspect list. An other example of an ineligible candidate would be an individual who was incarcerated at the time of the crime.

This process raises two ethical and legal issues – the use of discarded DNA and investigatory privilege. While neither are specific to IGG, both are factors that go hand-in-hand with the technique. Once discarded, DNA is considered public property and a warrant is not required for collection. In the opinion of privacy activists and entities such as the American Civil Liberties Union, this raises concerns of privacy violations and pushing the limits of law enforcement boundaries, potentially violating the Fourth Amendment. There have been recent pushes in court to contest the legality of discarded DNA collection, but there is yet to be an official ruling against it.\[31\] The second issue is that of investigatory privilege. Law enforcement does not have to reveal the process taken to
produce an investigative lead, meaning that IGG does not have to be discussed in a criminal court case. Therefore, IGG as an investigative method does not have the opportunity to be verified, challenged, or established in a court setting. This brings the legal standing of IGG into question, as it also allows incorrect or dishonest IGG practices a place in the courtroom. However, this would likely be mitigated since a CODIS hit or STR analysis match must be established before a case is taken to court.

As mentioned previously, most genealogy companies do not allow for law enforcement access to DNA matches. GEDmatch does have this option, but it is up to the user whether their profile will be publicly accessible. These decisions are made so that users can be assured of their privacy and can take an active role in volunteering information to law enforcement. This is a deliberate step taken by these companies to maintain ethical business practices and is typically praised by consumers. However, this severely limits the information that law enforcement can use to generate an investigative lead and could potentially be the difference between identifying a criminal or letting him run free. This is not to say that these companies should automatically make all DNA profiles public, as this would likely result in a myriad of legal issues, but it is worthy to note when considering the potential success rate of IGG. In this research, the base identity of the suspect was able to be determined using only public GEDmatch segment data and genealogy records – resources accessible to law enforcement. However, accompanying information, such as the chromosome map, would likely not be law enforcement accessible and could potentially hinder investigation. Consider the GEDmatch only chromosome map seen in Figure 3.21. There is a massive lack of information when compared to the chromosome map seen in Figures 3.24 and 3.27. Per SWGDAM and DOJ policies, law enforcement agencies must identify themselves as law enforcement to genealogical services prior to use for IGG, so this angle could not be used as a “work-around” for such a challenge. [13, 15] However, GEDmatch is open for law enforcement use, allowing agencies to upload and process DNA profiles that may have been developed elsewhere, such as a private forensic laboratory. IGG is not automatically impossible without the contribution of these additional sources, but it may prove more challenging.

Investigators interested in pursuing IGG for criminal cases should take the budget and timeframe of the method into consideration. A DNA sample must be sent to at least one DTC genealogy company, which involves buying a kit and possibly a subscription. There is a multiple month turnaround time for these companies to return a profile. GEDmatch is free at the base level, but certain helpful features, such as segment searching and triangulation, are only available behind a paywall. Certain genealogical records, while still considered publicly available, are only available via a reoccurring purchase. Even with spending the funds on premium subscriptions to access more information, the descendancy research process can take a significant period of time. Identifying the suspect in this “case” took months to build out multiple family trees, on top of the extended turnaround time experienced by the DTC companies. For reasons such as this, IGG is currently a highly privatized industry, as it could bear a significant load on the average investigations department. Tables 4.1 and 4.2 summarize the timeline and costs incurred for this research. It is important to note that more experience in the technique does not necessarily mean a shorter timeline. The amount of time it takes to establish a person/s of interest strongly depends on the availability of the needed genealogy documents. While the process would, ideally, get faster with experience, because it is not a straightforward process, even experienced investigators could hit snags.
Table 4.1: Approximate Timeline comparing Calendar and Hourly Effort

<table>
<thead>
<tr>
<th>Process</th>
<th>Turnaround Time</th>
<th>Working Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submission to DTC Companies and Receive Results</td>
<td>4 to 6 months</td>
<td>0.5 hours</td>
</tr>
<tr>
<td>Upload to GEDmatch and Establish Candidates</td>
<td>2 weeks</td>
<td>20 hours</td>
</tr>
<tr>
<td>Family Tree Building</td>
<td>6 months</td>
<td>480 hours</td>
</tr>
<tr>
<td>Chromosome Mapping and Phasing</td>
<td>1 month</td>
<td>80 hours</td>
</tr>
</tbody>
</table>

Table 4.2: Cost of Tools Used

<table>
<thead>
<tr>
<th>Tool</th>
<th>Cost</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AncestryDNA™ Subscription</td>
<td>$79.95</td>
<td>6 months</td>
</tr>
<tr>
<td>GEDmatch Tier 1</td>
<td>$10</td>
<td>1 month</td>
</tr>
<tr>
<td>MyHeritage™ Kit</td>
<td>$79</td>
<td>One-time fee</td>
</tr>
<tr>
<td>Family Tree DNA™ Kit</td>
<td>$79</td>
<td>One-time fee</td>
</tr>
<tr>
<td>23andMe™ Kit</td>
<td>$99</td>
<td>One-time fee</td>
</tr>
<tr>
<td>AncestryDNA™ Kit</td>
<td>$99</td>
<td>One-time fee</td>
</tr>
<tr>
<td>DNA Painter</td>
<td>$55</td>
<td>12 months</td>
</tr>
</tbody>
</table>

Steps were taken to avoid confirmation bias during the research process. No family members were contacted throughout descendancy research, even though individuals with shared DNA matches did attempt to connect through the various company websites. No family tree inferences were made based on personal knowledge of SH’s family lineage. Genetic information of close family members (i.e., within one to two generations) was not used as a starting point to provide a more realistic scenario. All family tree information generated from unofficial sources was manually verified with documented records prior to being included in family trees. For the purpose of this research, the person of interest was not assigned a gender during the investigative process. However, since general practice is to complete an STR profile and submit to CODIS prior to using IGG, the gender of the individual could already be known before this process begins. This could allow for a smaller candidate list and less investigative time as individuals of a certain gender could be eliminated prior to DNA collection.
5. Conclusions and Future Research

Going into this research, there was no way of knowing whether the technique could be utilized successfully or not for this particular case. However, the goal was met and a more defined idea of how this process works was developed. Using the four DTC companies, an extensive list of matches was returned and narrowed to useful individuals. 13 different family tree clusters were compiled and the connection between two of them allowed for the identification of a single suspect, rather than a list of multiple. This success was used to construct a streamlined method as a deliverable, which will serve as the base for gaining a better understanding of IGG as a whole.

It was relatively easy to retrieve the list of matches and narrow it to a workable candidate list, although it was time-consuming. The family tree building led to the greatest hurdle. The more clusters that were built, the easier the process became. Familiarity with document availability and the library database increased with time, cutting down on time spent searching for usable documents and establishing individual connections between records. This being said, familiarity will not necessarily result in a decreased turnaround time for the overall connection to the suspect, but rather on building individual clusters from the candidate list.

There are unavoidable challenges of IGG due to the nature of the technique, primarily rooted in the unpredictability of results. IGG does not have a method validation in place or a standard operating procedure for which to follow. There is no way to guarantee any useful matches will be returned, if any, or if the technique will identify an individual. This particular research submitted samples to four different DTC genealogy companies to increase the chance of a match. This will not necessarily be true for suspect samples or individuals submitting their own DNA. As touched on previously, there is also no set number of matches to find a connection between clusters. This leaves a large variability in the expected turnaround time to determine an individual’s identity. Other unavoidable challenges include the accessibility of genealogy records and DNA profiles. Because certain companies are utilized at different levels of popularity in different countries, one company might have more records of one country’s population than another. Genealogy records become increasingly difficult to find and use the older they are, as well as those closer to the present day. Older records may have unclear handwriting, missing pages, translation issues, or are just not available online. Recent records may not be available for public access. For example, census records cannot be accessed by the general public until 72 years after the census date.

This research would benefit from multiple repetitions. In order for IGG to be implemented into a reliable investigative technique, a better understanding of what level of connection is required for success is necessary. This research took 13 matches to find a union couple. Other matches were more closely related to the suspect, but they did not end up being helpful. Since an answer cannot always be found from closer matches, how far down the candidate list an investigator would have to go will vary - at what point should they stop and pursue another technique? A concerted effort was put into building out family tree clusters that were not useful, raising the question of
weighing effort against results. More repetitions with different cases and individuals would help answer cost/benefit questions like this.

Additional future direction of this research would be to publish a guide on the IGG methodology. This would involve useful genealogical databases, reference websites, pricing guides, current practices (from both individuals and companies), and other helpful information and tips. While IGG is not specifically an analytical method, perhaps a modified method validation protocol or similar guidelines in a forensic context could be developed.
6. Bibliography


