2021

Forensic Genealogy: A Tool in DNA Analysis

Samantha Olsen

Forensic Genealogy: A Tool in DNA Analysis

An Honors College Thesis

By

Samantha Olsen

Fall, 2021

Department of Chemistry and Mathematics

Faculty advisor (Name) __Keri Wyllie, MPA_____

Faculty advisor Signature _____

Reader (Name) ___Pasquale Buffolino, Ph.D._____

Reader Signature_____

Date_____12/20/2021_____

TABLE OF CONTENTS

Abstract

DNA is the basic unit of which differentiates one individual to the next. Short Tandem Repeats and Single Nucleotide Polymorphisms are two locations of DNA that have been the most successful in DNA analysis. Unique DNA profiles can be created so that interpretations can be made about the individual who contributed to the DNA sample. Systems like STRmix are currently being employed to give unbiased assumptions about the DNA profiles. There are several different statistical approaches in assigning probabilities to DNA evidence, likelihood ratio and random match probability are two common approaches. After assumptions are made, comparisons can be made between DNA profiles to find related familial members due to similarities in DNA profiles. Maternal kinship can be considered based on the mitochondrial DNA while paternal kinship is based on the inheritance of the Y chromosome. The Golden State Killer was a famous case that opened the door of applying forensic genetics and consumer genealogy sites to investigations. The future of forensic genealogy is endless, but first policies must be put out in place to protect the privacy of individuals DNA.

CHAPTER 1: HISTORY OF DNA ANALYSIS

The human body is composed of around 100 trillion cells, or the basic units of life (Butler 2010). Within each cell is the nucleus, and within the nucleus are chromosomes which hold deoxyribonucleic acid, or DNA (Butler 2010). DNA contains all necessary information for cell replication and protein synthesis (Butler 2010). DNA itself is a molecule with three main parts. The phosphate group and deoxyribose sugar combine to form the sugar-phosphate backbone (Butler 2010). The third part of DNA is the nucleotide base: adenine; guanine; cytosine; or thymine (Butler 2010). The sequence of these bases on the DNA strand is what is responsible for the differences within human beings (Butler 2010). DNA is known to be an antiparallel double helix. This means that there are two strands of DNA of which run in opposite directions and form a twisted staircase structure (Butler 2010). These two stands of DNA are connected because the nucleotide bases have complements. Adenine bonds to thymine and cytosine bonds to guanine via hydrogen base pairing (Butler 2010).

While the discovery of DNA has been around for hundreds of years, the more recent development of DNA typing has only been around since 1985 (Butler 2010). The phrase "DNA typing" can be interchanged with "DNA fingerprinting" or "DNA profiling", but all pertain to looking into DNA sequences of individuals to perform identity testing based off these sequences (Butler 2010).

Because DNA is individual to each person, a DNA profile can be compared to a fingerprint. If there is a sample of DNA left at a crime scene, then DNA testing is necessary to figure out who contributed to that sample.

To properly understand DNA typing, we must first get a good understanding of DNA analysis. The importance of DNA in forensic-type cases is due to the fact that the human genome is unique to every induvial, except for identical twins (Butler 2010). Half of the human genome comes from the mother and the other half from the father (Butler 2010). One would think that this means that every single individual must have completely different genomes, but all humans actually share around 99.9% of their DNA. That 0.1% of DNA is what is responsible for the differences between each individual. Because there is only a small portion of our genome that is different, scientists only look at a small subset of genetic variation (Butler 2010). Only 0.0006% of nucleotides are examined and tested from the human genome for forensic-type analysis (Butler 2010).

DNA analysis can be performed with a number of different methods. The common goal between these methods is to differentiate two individuals beyond a reasonable doubt (Butler 2010). Through the past few decades, this process has improved exponentially (Butler 2010).

One of the earliest methods to differentiate was through blood group testing to be used in court. Every individual has one type of blood based upon their cell surface antigens: A, B, AB, or O (Butler 2010). The ability to identify a blood type of a sample can be best used to exclude someone of a different blood type. However, this should not be used alone to convict one individual (Butler 2010). Millions of people can have the same blood type. For example, if the DNA showed that the perpetrator was blood type A, and the person of interest was blood type B, then we know that individual did not contribute to that evidence. However, if the second individual was also blood type A, we wouldn't be able to exonerate or convict them on that evidence alone. About 42% of the population also has blood type A and there is no way to directly relate the suspect to the evidence through blood typing (Butler 2010).

Restriction Fragment Length Polymorphisms, or RFLPs are proven to be polymorphic meaning they are different between individuals. They can commonly be found in the noncoding region of the genome and used to track genetic inheritance and differentiate individuals (Butler 2010). The original process consisted of extracting and unwinding DNA, adding a restriction enzyme to cut the DNA strand at specific sequences, and then running the resultant fragments on agarose gel so that the fragments will be separated by molecular mass, or size (Butler 2010).

A southern blot test would then follow to ensure that the DNA fragments were in a denatured, or single stranded, form for Variable Number of Tandem Repeats (VNTRs) to bind. VNTRs are pieces of DNA tagged with chemiluminescent probes that would hybridize completely with specific sequences of the single stranded DNA (Butler 2010). The chemiluminescent probes had tags that allowed the analyst to track each sequence and marker to properly identify the suspect's DNA to the evidence (Butler 2010). Whether the RFLPs had multi-locus or single-locus probes, the result had a high power of discrimination. However, the process was relatively slow compared to other methods of DNA typing (Butler 2010).

In the late 1980's a new method was developed using short tandem repeat, or STR analysis (Butler 2010). STRs utilize shorter repeat units of genetic markers than RFLP. Since the late 1980's, this method has been repeatedly improved and is one of the more commonly used methods of DNA analysis today (Butler 2010). Typically, STR analysis is sensitive and rapid. However, it does have some of its own limitations such that they are less suitable for degraded DNA when compared to Single Nucleotide Polymorphisms (Butler 2010).

DNA is present in biological samples such as blood, semen, saliva, hair follicles, skin cells, and so much more (Butler 2010). The only way that this type of evidence from crime

scenes can be utilized in a forensic laboratory, is if the DNA is collected and stored properly from the source material. For example, each item of evidence should be packed separately and properly labelled. Any evidentiary items must be marked with the case number, sample information, and collection date (Butler 2010). The collector should sign their initials and date across the package seal. Stains should be air dried before sealing a package so the sample will be maintained to the highest degree of certainty. Left over water can speed the degradation of DNA molecules (Butler 2010). In order to further maintain the samples, they should be placed in paper envelopes or bags (Butler 2010). Storage of most biological evidence should be in a dry and cold environment until the time of testing (Butler 2010).

Once the samples are brought into the laboratory, the first analyses to perform are presumptive tests. This will determine whether the evidence is or isn't a specific biological material such as semen or blood (Butler 2010). If the presumptive tests are positive, then confirmatory testing can be performed to further prove the nature of the biological fluid (Butler 2010). If the confirmatory testing is also positive, then we can assume that biological fluid is human, and DNA is present in the sample.

The next step is to extract the DNA from the rest of the sample so it can be further examined (Butler 2010). There are many different methods in DNA extraction including organic, differential, and resins such as Chelex, depending on the sample's integrity (Butler 2010). The age of samples is one characteristic that can be used to determine which means of extraction is best suited.

Organic extraction has been used for many years and is best for high-molecular weight DNA (Butler 2010). However, this method uses a very toxic chemical, phenol: chloroform, and

is very time consuming (Butler 2010). With this method, sodium dodecylsulfate (SDS) and proteinase K break down the cell walls, then phenol:chloroform separates the proteins from DNA. Centrifugation in the last step of the procedure separates the DNA from the residual cell debris (Butler 2010).

An inorganic extraction, such as with Chelex, is cheaper, more rapid, and utilizes less transfer from tube to tube to decrease contamination (Butler 2010). Chelex is an ion-exchange resin with a high affinity for polyvalent metal ions (Butler 2010). The protocol involves the sample being added to a solution of Chelex (around 5%). The mixture is then boiled to lyse the cells. A saline wash is used to remove other cellular materials as the DNA is denatured and centrifuged to be separated (Butler 2010). However, this type of extraction results in single stranded DNA, which limits further analysis to polymerase chain reaction (PCR) (Butler 2010).

Lastly, differential extraction is used to extract the male DNA profile from sperm cells in mixed samples found in cases such as sexual assault. The mixture of male sperm cells and female epithelial cells is suspended in SDS/proteinase K solution (Butler 2010). SDS and proteinase K are used to break down proteins and lyse the membranes of epithelial cells (Butler 2010). The membranes found in the heads of spermatozoa will not be lysed by SDS/proteinase K alone due to the presence of disulfide bridges and the presence of cysteine (Butler 2010). Then, once the supernatant containing all the epithelial DNA is removed, dithiothreitol (DTT) can be used to further lyse sperm heads so the male portion can be used to testing (Butler 2010).

Whichever method of DNA extraction is used, it is important that DNA is kept preserved to best degree and to remove any potential inhibitors of PCR (Butler 2010). This is because the next step for the forensic laboratory is to quantify this DNA. The quantification of DNA is

important to first make sure that there is a sufficient amount of DNA in a sample for further amplification (Butler 2010). The amount of DNA to be tested is important because the polymerase chain reaction can only run on a small range of DNA concentrations (Butler 2010). The desired range of template DNA is 0.5 to 2.0ng (Butler 2010). The amount of DNA can also be used to determine the degree and method of amplification in the future. For example, a quantity of 1.0ng of DNA will be amplified to about 303 copies of each locus (Butler 2010).

Once the quantification of DNA is completed, a small amount of the DNA is amplified. Copies of DNA can be made from the sample through a process known as the Polymerase Chain Reaction, or PCR (Butler 2010). PCR takes a specific region of DNA, part of that 0.1% that is different between every individual and replicates it over and over so that there are millions of copies of a specific sequences (Butler 2010). The process in a forensic setting is around 32 cycles of pre-specified heating and cooling steps. Each cycle causes the number of target molecules to approximately double (Butler 2010). There are many components needed for a successful run of PCR to occur. This includes primers, buffers, nucleotide, template DNA, and DNA polymerase.

The steps of PCR are denaturation, annealing, and extension. First, denaturation opens the DNA strand from being double stranded to single stranded. Buffers, like sodium chloride, are used to decrease the melting temperature of the strand to allow for quicker reaction times. In the annealing step, primers bind to the single strands of DNA (Butler 2010). Primers are short DNA sequences that can be used to target a specific sequence of DNA through complementary pairing. In PCR, there are forward and reverse primers, depending on whether it is the forward or lagging strand of DNA. Lastly, in the extension step then the DNA polymerase will add

complementary nucleotides in the 5’ to 3’ direction on both single strands to form 2 double stranded molecules which are genetically identical. (Butler 2010).

Through PCR, an *in vitro* method, DNA is replicated so that it can be better tested. *In vivo*, DNA replication is non-symmetrical. This is because it is semi-conservative, so each strand of DNA is synthesized differently due to the need of being synthesized in a 5’ to 3’ direction (and DNA is double stranded, antiparallel). The 3’ end OH group is able to form a phosphodiester bond with the “new” phosphate group at the 5’ end (Butler 2010). The leading strand of DNA is synthesized continuously. DNA Primase can begin replication when it finds the correct base “code”. DNA Polymerase can then start to read the template strand and find the complement base pair (Adenine to Thymine, and Guanine to Cytosine) and add the “new DNA” to the 3’ end. The lagging strand is then synthesized discontinuously. Due to Helicase working alongside DNA Polymerase from the leading strand, Helicase can denature DNA by breaking hydrogen bonds between adenine to thymine and cytosine to guanine as the leading strand is synthesizing. This allows for DNA Primase to synthesize RNA Primers in a 5’ to 3’ direction of short segments of the lagging strand (Butler 2010). These RNA primers can then “sit” on the complementary strand and allow for the replication of Okazaki fragments. DNA Polymerase then comes into play to finish adding fragments of DNA by replacing the RNA that was added. Now, there are short sections of DNA fragments that can be “glued” together by ligase.

The last step in the sequencing of DNA involves a method known as capillary electrophoresis. This was introduced as an alternative to slab gel (agarose) electrophoresis, a process that separated DNA fragments based off size (Karger 2011). Capillary gel electrophoresis offers a higher resolution of the separation of DNA. The analysis utilizes a polyacrylamide column and computerized system with ultraviolet and visible light detection

(Karger 2011). The results are generated with increased speed and efficiency. Crosslinked polyacrylamide capillary columns achieved high efficiencies in DNA analysis, meaning it can be used to sequence single stranded oligonucleotides (Karger 2011). Already, compared to the polyacrylamide slab gel electrophoresis, the capillary gel electrophoresis was fivefold faster in separation speed (Karger 2011). Some issues did arise, like the fact that bubbles formed near the injection end of the capillary and ruined many procedures. They were likely formed from osmotic shock from the high salt concentration that migrated into the capillary (Karger 2011). One fix to this solution was to use non-cross linked polymer matrices. Another issue was that the polymer matrix needed to be automatically replenished, which isn't ideal for any geneticist to be working with (Karger 2011). Replaceable linear polymer matrices could be used in capillary electrophoresis for DNA sequencing, single stranded DNA, double stranded DNA, and RNA (Karger 2011).

The sequencing of DNA with capillary electrophoresis was effective, but not feasible for sequencing the entire human genome of 3 billion bases (Karger 2011). A 96-column format was made for this larger scale work. One way to improve separation is to increase the capillary column temperature. This is useful when there is a broad band that is forms from the decrease in size-based electrophoretic mobility (Karger 2011). Another way is increasing the chain length of the sieving polymer if the lengths that pass through are all too long. Because processing a single column at a time is expensive and time exhaustive, multicapillary sequencing was very necessary (Karger 2011). This can handle hundreds of DNA samples and sequence billions of bases at once (Karger 2011).

DNA itself is an important building block of forensics. The way that DNA naturally replicates and get passed down is an important concept so that it can be examined in humans.

Understanding what it is and how to handle it in the lab is important for further investigation and experimentation of the DNA. Processes like DNA extraction, quantitation, and amplification are essential beginning steps so that a clean, and comprehendible DNA profile can be created. Further investigation of this DNA will allow for scientists to study the differences between samples of DNA to work and identify individuals from DNA samples. There are many different methods of DNA analysis that can be done to create DNA profiles.

CHAPTER 2: INTERPRETING DNA PROFILES

A DNA profile is a pattern of specific DNA segments that are unique to one individual. The majority of people have identical DNA sequences except for a small percentage which can be used to differentiate one person from the next. The portions that make up this percentage which varies between individuals is what forensic scientists study and sequence during DNA profiling (Panneerchelvam 2003).

DNA Analysis was first known to be true in 1985 by Dr. Alec Jeffreys (Jeffreys 1985). He discovered that there are regions of DNA sequences repeated right next to each other. Dependent on the individual is how many repeated sequences there are in one sample (Butler 2010). As discussed previously, there are different methods that are used in the forensic lab to achieve this goal. Restriction Fragment Length Polymorphisms (RFLP), Variable Number of Tandem Repeat Sequences (VNTR), Short Tandem Repeats (STR), and Single Nucleotide Polymorphisms (SNP) are a few of the many ways in which forensic scientists can type DNA.

The first set of DNA regions studied were VNTRs. (Butler 2010). Variable Number of Tandem Repeat Sequences are segments of DNA which repeat the same sequence of base pairs found within the human genome (Nwawub 2020). VNTR's range from 8 to 100 bases in length. Due to this sizing, they are also known as minisatellites (Butler 2010). VNTR's have a high rate of mutation, higher than the average on the genome (Nwawub 2020). Especially when compared to STR's, there are more base pairs being studied which mean that there are more locations at which there can be a mutation in the base pairs.

Probes made from VNTRs can be used in different sequencing tools. These probes contain a VNTR sequence and are dyed with chemiluminescence (Butler 2010). After DNA

separation based on size, these probes can hybridize with the complementary sequence which is being amplified in order to create a profile. Because each probe is labeled by the dye of the chemiluminescence tag, after hybridization the probe can be visualized and the specific sequence is known (Butler 2010). The first method consisted of multi-locus VNTR probes (Butler 2010). This means that a single labeled probe can bind to various regions of the human genome (Butler 2010). Eventually, this improved to single-locus probes, where there one or two known alleles that could be detected if a probe bound to a certain sequence.

Restriction Fragment Length Polymorphisms utilize restriction enzymes to cut DNA at specific sequences (Panneerchelvam 2003). These fragments will all have different lengths per person and can be used to identify individual characteristics. The DNA fragments are then placed on a gel to be separated based on their size (Panneerchelvam 2003). This method is most useful to determine if two DNA samples are different, rather than confirm that the same have come from the same individual (Panneerchelvam 2003).

While there are current advancements in SNP typing, STR typing does continue to be the main tool in forensic DNA analysis. DNA in our genomes consist of repeated sequences by a core unit repeat (Udogadi 2020). Short Tandem Repeats are microsatellites, which mean that they are two to six base pairs in repeat. Because these units are so small and do not degrade as fast as larger fragments, they can be best suited when DNA may be compromised (Udogadi 2020). STRs are highly polymorphic so there is a high degree of variability from one individual to the next (Udogadi 2020).

The regions of DNA being evaluated as STRs are known as loci. The loci are composed of core units of nucleotides repeated up to a length between 80 base pairs and 400 base pairs

(Panneerchelvam 2003). There are already a large range of loci that have been identified, characterized, and demonstrated within the human genome. Originally, CODIS looked at 13 core loci for identification of individuals. It was then determined that the more STR loci being used, the greater the discrimination value, which has led to the number of core loci being increased to 20 (Udogadi 2020). Most STRs are found in the noncoding region of our genome, with only 8% located in the coding regions (Udogadi 2020).

Because of their high rate of polymorphism, short tandem repeats are used frequently in DNA profiling (Panneerchelvam 2003). Polymorphism is the ability of a specific region of DNA to have more than one variant. Since there are a number of different regions throughout the noncoding region of the genome which exhibit this polymorphic ability, these specific areas differ among individuals and can be used to match DNA sample to individual (Butler 2010). Therefore, STR's are most commonly used for human discrimination.

STR Typing looks at the alleles at these specific loci. For most mendelian traits, an individual receives one allele from their father and one from their mother (Panneerchelvam 2003). This creates a combination of genetic heritage that neither exactly matches the maternal or paternal DNA profile. STRs at these loci are also inherited from either the maternal or paternal DNA.

Single Nucleotide Polymorphisms, or SNPs, are a single-based sequence variation between individuals on the genome (Butler 2010). They are abundant in the human genome and can be used to track genetic diseases. SNPs are usually less that 100 base pairs in size, so they can be better suited for DNA that is partially degraded (Butler 2010). SNPs are mostly biallelic, so there are three possible genotypes: heterozygous, homozygous for one allele, or homozygous

for the other allele (Butler 2010). Compared to STR loci, these are far less alleles, meaning they are less polymorphic (Butler 2010). Therefore, we do need more SNP markers for proper discrimination. For example, we would need around 25-45 SNP loci compared to the original 13 STR loci for proper discrimination (Butler 2010).

Additionally, we need to know how to analyze the SNP data so that we can correctly type the markers. We can examine multiple markers simultaneously because SNP are not as variable as STRs (Butler 2010). One method of SNP analysis is the TaqMan 5' nuclease assay (Butler 2010). A fluorescent probe with a reporter and quencher is included in the PCR reaction that is complementary to a specific region (Butler 2010). When the complementary region is amplified, the probe is cleaved off and results in fluorescence (Butler 2010). This allows for the visualization and knowledge of where the complementary region is located.

Another method for the analysis of SNPs is Luminex 100, or allele-specific hybridization. Different SNP types are represented by oligonucleotide probes, which are hybridized to dye-labeled PCR products. These are attached to colored beads which indicates whether a PCR product is attached while the beads pass through two laser of a flow cytometer (Butler 2010). Both this method and another called the SNaPshot minisequencing assay, allow for the multiplexed analysis of multiple SNP markers simultaneously (Butler 2010).

SNPs can be used in the forensic world for ancestry and lineage information (Butler 2010). It can be especially useful to predict an individual's ancestral background. SNPs have a relatively low mutation rate, changing once every 10^8 generations (Butler 2010). This leads to SNPs becoming fixed in certain populations and therefore are population specific (Butler 2010). SNPs are also being studied as potential tools to identify genetic variants that code for

phenotypic characteristics (Butler 2010). For example, a mutation in the human melanocortin 1 receptor gene is associated with red hair (Butler 2010).

Although this new idea is still being studied and experimented with, but with time and efforts, SNPs can provide information on phenotypic traits such as facial features. Many of the direct-to-consumer genealogical sites do use SNP markers for DNA analysis (Kling 2019). They analyze more than 600,000 autosomal SNP markers on high density microarrays (Kling 2019). Sites like Ancestry, 23andME, and GEDmatch do use SNP markers for their DNA analysis (Mateen 2020).

The Combined DNA Index System (CODIS) is a computerized tool that can be used to compare DNA profiles. It is a software run by the FBI to keep track of the DNA of criminals first created in 1990. In 1998, CODIS originally considered 13 core loci for DNA profile comparisons (Panneerchelvam 2003). As of 2017, CODIS considers 20 core loci when comparing DNA profiles (Panneerchelvam 2003). CODIS contains the DNA profiles of convicted individuals. Once a DNA profile is entered, the system can try to find matches based on similarities in profiles already in the CODIS system (Panneerchelvam 2003). Because of its accessibility by different levels of law enforcement, from local, to state, to national, and federal, profiles can be identified from other databases across the United States (Panneerchelvam 2003).

While creating these DNA profiles are extremely useful, the way in which they are interpreted is even more important. Most interpretation is done mathematically in order to quantify the probability of one DNA profile matching an individual's DNA profile. This is what can determine if a suspect is found guilty or not.

DNA profiles can be made from almost any type of biological sample. As technologies and advancements have improved with time, so has the ability to form DNA profiles from poor samples (Puch-Solis 2013). Whether these samples were subject to degradation or were just so miniscule at a scene, we can still work to make suitable DNA profiles (Puch-Solis 2013). DNA is collected, extracted, amplified with PCR, and then separated via capillary electrophoresis (Bright 2014). We can study alleles and loci of DNA profiles by looking at electropherograms, or epgs. Certain peaks on an electropherograms indicate a DNA allele. There are a set of peaks made on a fluorescence versus time plot (Bright 2014). The height of the peak exemplifies the amount of amplified DNA and the length can be determined by the x-axis (Puch-Solis 2013). But when DNA samples are not the best, issues often arise in the electropherogram data.

A phenomenon known as allelic dropout exists when an allele from a gene doesn't generate an electropherograms peak (Puch-Solis 2013). This phenomenon can also form small peaks, but if they do not meet the predetermined threshold height, then it can be considered dropout (Puch-Solis 2013). Another artifact to be aware of when studying electropherogram peaks is stutter. Stutter can be compared to "background noise" (Puch-Solis 2013). It is formed because of the PCR process when an allele is one repeat unit shorter than the target allele (Bilie 2014). Lastly, when a DNA sample comes from multiple contributors, there are varying levels of DNA template and therefore alleles (Puch-Solis 2013). This is where known peak heights and estimations are used to make the best interpretation as to which peaks belongs to which individual. It is key to identify both the major and minor contributor to the sample. Sometimes, mix ratios can be calculated to estimate the amount of one contributor compared to the other (Billie 2014). This can be used when looking at allelic frequencies to best attribute each allele to a contributor.

Population statistics is a tool used in probabilistic genotyping. It takes a subset of the population into account when looking at a particular sample. When looking at a sample it is important to note that if two samples don't match, then we can presume the sample did not come from the individual in question (Buffolino, 2021, *Population Statistics*) However, if the DNA profiles do match, there is still the question of whether the sample could have come from only that individual or if the sample matches that of a different individual with the same DNA profile (Buffolino, 2021, *Population Statistics*). We can look at random match probability and the likelihood ratio to measure this to the highest degree.

Both random match probability and likelihood ratio take population allele frequency into account (Buffolino, 2021, *Breaking Traditions*). Allele frequency is calculated by counting the number of times each allele is observed in a population (Buffolino, 2021, *Population Statistics*). The subset of a population is studied and used to create a genotypic array at each location. The allele frequency is calculated by dividing the number of times an allele is present over the total number of alleles studies. These allele frequencies can be further used to calculate random match probability by finding the inverse of the product of each allele frequency (Buffolino, 2021, *Population Statistics*).

Random match probability is the likelihood of choosing an unrelated individual at random in the population with the same genetic profile as the one in question (Buffolino, 2021, *Population Statistics*). Another way to approach this is the frequency at which a DNA profile can be found in a given population.

Random match probability is a "rarity" statistic. When forensic scientists testify, they would use statements which describe how often the DNA match would be present as a

probability (Buffolino, 2021, *Breaking Traditions*). For example, when explaining the random match probability of a specific profile, the forensic scientist would say, "This DNA profile would be expected to be found in one in one billion individuals" (Buffolino, 2021, *Breaking Traditions*).

Likelihood ratios are best used to determine how evidence can be weighted in trial (Puch-Solis 2013). This can help to give clear cut statistics to the probabilities of scenarios occurring to either convict or acquit an individual. This value considers the probability of obtaining some evidence given two competing propositions (Bright 2012). Likelihood ratios compare probabilities under a null and alternative hypothesis or the hypothesis of the prosecution and defense (Buffolino, 2021, *Population Statistics*). The difference between this method and random match probability is that likelihood ratio takes the evidence and proposition of situations into effect when calculating the ratio and explains how much more likely an individual contributed DNA than not (Buffolino, 2021, *Continuing with Continuous Models*). Each hypothesis is mutually exclusive of the other but favors the ideal that side is representing. An example of a hypothesis of the prosecutor could be "that the DNA originated from the person of interest and one unknown contributor" (Billie 2014). Meanwhile, the defense hypothesis may consist of "that the DNA originated from two unknown contributors" (Billie 2014). These hypotheses can then be transformed to denote symbols and numbers based off the evidence. Then, once both hypotheses are created, we can solve for the value of the likelihood ratio. The likelihood ratio is a fraction with the numerator being the hypothesis of the prosecutor and the denominator being the hypothesis of the defense or H_0/H_1 (Buffolino, 2021, *Breaking Traditions*). Usually, the prosecutor's hypothesis will be that the DNA in question originated from the suspect while the defense would be saying that the DNA randomly matches the

defendant and is actually from an unknown person from the population. A likelihood ratio could be explained by saying "The DNA profile is around 1,000 times more probable if the sample originated from Individual A than if it came from an unknown individual" (Buffolino, 2021, *Breaking Traditions*).

Both likelihood ratio and random match probability are being used in forensic science. There is currently not a definitive more preferred tool in population statistics at the moment but with more advancements and statistical analysis being done, there can be in the future. Many scientists prefer to use likelihood ratio calculations because it takes more factors into account when making interpretations.

Sometimes assumptions need to be made in DNA analysis for interpretation. This, will in turn, affect the likelihood ratio. Some circumstances that can affect the likelihood ratio can be increasing the profile complexity, assigning the number of contributors, or to replicate amplifications (Taylor 2014). In this experiment, likelihood ratios were calculated under varying propositions. The first experiment tested the outcomes based off the correct number of contributors inputted (Bright 2012). This experiment showed that the likelihood ratios ability to differentiate between true and false propositions declines as less true information is inputted (Bright 2012). This is because less information could lead to smaller or fewer peaks which then leads to more uncertainty of the DNA information (Bright 2012). Even if the correct number of contributors are inputted, the likelihood ratio does get more skewed as the number of contributors increase. This is because there are more genotype sets in the profiles and can be more difficult to differentiate between which allele belongs to which contributor (Bright 2012). The second experiment tested the use of replicate PCR's in each analysis (Bright 2012). The third experiment tested inputting three out of the four contributors were known so that 4 analyses

of each sample was tested (Bright 2012). This is another example of adding more relevant and correct information.

When more contributors are knows, we can remove many of the possible genotype sets in the combined DNA profile, this leaves a restricted set of genotypes that can be available for what is unknown (Bright 2012). This allows for the weights to be more concentrated on the true contributors' genotype (Bright 2012). By providing more and more of the true contributors' profiles, we can increase the resolution of the remaining contributors genotype (Bright 2012). The fourth experiment inputted incorrect information, assuming that a non-contributor was a part of the deconvolution (Bright 2012). Each time the outcome of the likelihood ratio can vary drastically, even to completely excluding a known contributor (Bright 2012).

In the fifth experiment there was unnecessary information added. There was an additional contributor to the "true" known contributors, but the non-contributor was chosen at random from a database (Bright 2012). Their contribution to the profile was essentially zero. Because of this, the change in likelihood ratio can be considered negligible (Bright 2012). These experiments can all show the effect that the analyst has when inputting certain information when calculating likelihood ratios.

There are three main methods utilized when looking at STR interpretations: the binary model, semi-continuous model, and continuous model. The binary, or discrete model assigns a DNA profile the value of 0 or 1. The interpretation is done manually, so one person could come to a different conclusion than another individual (Buffolino, 2021, *Breaking Traditions*). Usually, random match probability is associated with the binary model.

As discussed, prior, the more traditional method of DNA profile interpretation is known as binary. It can take real life complications into account but is stuck due to human subjectivity of the model (Billie 2014). As more advanced models come into play, they are slowly replacing the binary model. Due to the need for analyst decisions in determinations, this method can be inconsistent (Billie 2014). Identifying the molecular weight of loci, or individual contributor degradation, or the number of contributors all do affect the outcome of the DNA interpretation (Billie 2014).

Semicontinuous and continuous models can better take stochastic events into account (Billie 2014). Stochastic events are events that are unpredictable and sometimes uncommon. The semicontinuous model is an improvement because it does take the probability of dropout into account (Billie 2014). The fully continuous model takes peak height and peak height ratios into account (Billie 2014). Models like this, take the quantitative information from the electropherogram and conclude probabilities of the peak heights given all possible genotype combinations (Bright 2012).

The continuous model of interpretation assigns a DNA profile a value ranging from 0 to 1. This method utilizes more factors from the DNA sequencing results into account like dropout, peak height ratios, and even the number of contributors (Buffolino, 2021, *Breaking Traditions*). Depending on which factors are used determines how continuous the model is. The semi-continuous model doesn't necessarily utilize peak heights but does take dropout into account (Buffolino, 2021, *Continuing with Continuous Models*). The fully continuous model is even more objective and takes more factors into account. These factors include the peak heights, stutters, and drop-out (Buffolino, 2021, *Continuing with Continuous Models*). The likelihood ratio is usually associated with the continuous models of interpretation.

STRmix is an extremely useful and current tool in DNA interpretation. It is a software that takes electropherogram calculations into account to predict the probability of genotype profiles (Bright 2016). This program assigns a statistical weight to the probability of a possible genotype at a specific locus (Bright 2016). It is especially useful in the continuous model of interpretation (Buffolino, 2021, *Continuing with Continuous Models*). Because STRmix is a computerized program, it can assign weights and ratios to certain scenarios. For instance, STRmix applies a per allele stutter ratio so that individual stutter contributions to peaks can be considered (Billie 2014). STRmix can make the best use of data across profiles. The binary model produces less discrimination when compared to the semicontinuous and continuous models (Billie 2014).

STRmix takes all factors into account and uses quantitative information from the sequencing results. One factor, stutter, is the appearance of smaller, artifactual peaks that are made during the process of PCR (Bright 2016). Drop-in is when there are low amounts of DNA present in the profile that do not actually appear to be from the DNA sample (Bright 2016). Allelic drop-in results in partial DNA profiles due to low or degraded DNA (Bright 2016).

STRmix does not take reference profiles into account when deconvoluting a sample, unless it is a reference of a known contributor (Bright 2016). For example, if there is a sample from saliva of a victim, the reference DNA sample of the victim will be inputted along with the DNA sample in question because there is reasonable probability to believe that the victim's DNA will be present in their own saliva. Therefore, when the DNA profile is created from the potential combined sources, the victim's DNA profile can be extracted, creating a DNA profile for the perpetrator only. Likelihood ratios can then be created when we have the DNA in question from the scene and DNA sample from a person of interest.

One advantage of STRmix is that it can still provide a likelihood ratio while taking relatives of the person of interest into account. STRmix can consider all members of the population without having to specify a related or unrelated individual. (Bright 2016). Usually, with two closely related family members, it may be difficult to differentiate who may have contributed to a DNA sample. If this is the case, then the two DNA profiles can be further provided to STRmix so that it can further deconvolute the sample and create new likelihood ratios with the additional DNA information (Bright 2016). However, one disadvantage of STRmix is that if the process is repeated, the conclusions will not always be the same (Bright 2014). Each time STRmix is ran, there is a different outcome which questions which outcome is the "most correct" (Bright 2014).

Based off the values outputted, there are three possible outcomes in DNA interpretation. The possibilities are that the evidence cannot exclude the DNA profile, the evidence can exclude the DNA profile, and the results are inconclusive (Bright 2016). It is important to note that STRmix won't conclude that the DNA profile is only the DNA of the person of interest. There is always a percentage of uncertainty. Therefore, likelihood rations and random match probability values are used.

Sensitivity and specificity are two important factors when evaluating any scientific method, including DNA interpretation. Sensitivity is the ability to identify the DNA profile of a known contributor in a mixed DNA sample when the details of the template are unknown (Bright 2016). Specificity is the ability to exclude non-contributors that have been identified from a mixed DNA profile (Bright 2016).

Both factors are important because case-type DNA are often not found in ideal conditions. DNA can be degraded, there can be a very small amount obtained, or DNA can be from multiple people, and it is hard to differentiate. All of these factors can lead to incomplete or inconclusive DNA profiles.

The number of contributors in each sample is an important aspect in relaying outcomes from STRmix. The number of known and non-known contributors do influence the likelihood ratio outputted by STRmix (Bright 2016). In reality, the true number of contributors is always unknown. If the number of contributors was higher than that of actuality, then the likelihood ratio produced would be lower (Bright 2016). If an individual were to underestimate the number of contributors, then the likelihood ratio value would be exclusionary and therefore favor an incorrect genotype (Bright 2016).

As one would expect, the results of the likelihood ratio that STRmix outputs depends on both the sample used and the parameters set (Bright 2016). The sample is dependent on the number of contributors, the quality of the DNA, the probability that the data regard the person of interest as a contributor, and the amount of STR information (Bright 2016). The run parameters are dependent on the numbers set for that of the programming, like the weights set by Markov chain Monte Carlo (MCMC) and the Random Walk Standard Deviation (RWSD) (Bright 2016).

All in all, STRmix is an extremely valuable tool in the interpretation of various types of DNA profiles. The Scientific Group on DNA Analysis Methods published guidelines for the Validation of Probabilistic Genotyping Systems in 2015 (Bright 2016). This was done to verify (or nullify) the functionality of the STRmix system and its results (Bright 2016). The Scientific Working Group on DNA Analysis Methods (SWGDAM), a group of CODIS Administrators

from across the nations, find this method of interpretation to be precise, accurate, reproducible and repeatable according to their guidelines (Bright 2016).

STRmix is one useful tool that can be used to interpret DNA profiles, like how the binary and continuous models can be used. To create accurate and unbiased assumptions about these DNA profiles both likelihood ratios and random match probabilities can be calculated. Short Tandem Repeats and Sigle Nucleotide Polymorphisms are two major aspects of DNA that can be used to differentiate individuals. Even with the tools and advancements in differentiation and identification of individuals from DNA, there are sometimes issues and roadblocks in finding these individuals.

CHAPTER 3: FAMILIAL DNA ANALYSIS AND MEANS OF COMPARISON

When DNA is found at the scene of a crime, it is sometimes thought to be the key to solving the case. However, despite the quantity of DNA and quality of the DNA profile recovered, it would prove to be useless if there is nothing to compare the profile against. If there is no person of interest and CODIS has failed to produce a viable profile match, then further investigation must be undertaken.

One potential pathway of investigation would be familial DNA searching. While familial DNA searching has been used in New York State for the past two years, it still has a lot of potential yet to be explored (Debus-Sherrill 2019). This technique utilizes DNA databases to identify family members of those who contributed to a DNA sample (Debus-Sherrill 2019). Instead of searching for an exact match at specific loci, software can be used to find similar matches at those loci (Debus-Sherrill 2019).

These DNA similarities can identify a close familial relative. The degree to which the loci are alike is directly correlated to how close in lineage two individuals can be. The types of genetic similarities are what is measured to determine the type of relationship (Debus-Sherrill 2019). This software is able to search the DNA database and rank a list of potential, close biological relatives to an unknown individual (Debus-Sherrill 2019). More testing must then be done to either support or refute the claim of relatedness. Lineage testing consists of the additional DNA testing like Y-STR and mitochondrial DNA analysis to further determine if two individuals are related (Debus-Sherrill 2019).

It is important to note a similar approach called partial matching. Partial matching uses CODIS alone to identify DNA profiles that are alike to the one in question (Debus-Sherrill

2019). These similarities in the DNA profile do not necessarily mean the individuals are related but instead infers that the population may have a random consistency at a certain locus (Debus-Sherrill 2019). CODIS can be set to search for different stringency levels of comparison. Moderate stringency matches can allow for one base mismatch and partial matches. High stringency would mean that all loci would match while low stringency would mean that less loci match and are therefore partial matches (Debus-Sherrill 2019).

Kinship testing can be used to identify the relationship between two individuals who may be biologically related. This utilizes certain genetic markers like STRs, SNPs, and any type of polymorphism (Zhang 2020). STRs are the most widely used due to their high degree of polymorphism. Most kits analyze from 15 to 23 STR loci in a single PCR analysis tube. SNPs, on the other hand, have a lower mutation rate, smaller amplicons, and can be used even on degraded samples (Zhang 2020). SNPs could identify from fifth-degree relatives to ninth-degree relatives, or an unrelated pair of individuals. While both have their perks and disadvantages, in the past they have never really been tested simultaneously.

When testing STR markers in databases with convicted individuals, like CODIS, only first-degree relatives can be easily identified (Kling 2019). However, to use familial DNA searching, it is necessary to identify more distant relatives, like 2nd cousins or even 9th cousins. Studies have shown that this can be done by using dense sets of SNP markers (Kling 2019). One way that we can do this is through the Likelihood Approach, which measures the Identical By Descent degree (Kling 2019). The likelihood ratio is calculated to measure the weight of evidence (Kling 2019). By observing genetic markers for a set of individuals and hypotheses about the relationship, the conditional probability of relatedness can then be calculated (Kling 2019). Because statistics like population allele frequencies and population substructures are

included, this approach is sensitive to the estimation of certain population parameters (Kling 2019). This method proved to have the highest classification rates for all related relationships (Kling 2019).

The next two methods are through identity by state (IBS) relations. The first of the two is known as the KING method (Kling 2019). This method counts the number of shared alleles IBS for each marker and then averages that over many markers to infer the IBD relationship (Kling 2019). The second identity by state method is called the Segment approach. This method measures segments of chromosomes where there is a shared allele (Kling 2019). The length of each segment is measured in centimorgan (cM) and then the total length can determine a measure of relationship (Kling 2019). This method is used because stretches of DNA are inherited and unchanged through generations, without recombination (Kling 2019).

Due to the randomness that DNA is transmitted through generations, the probability that distant relatives share segments in their genome identical by descent will decrease inversely proportional to the number of generations (Kling 2019). We can then identify the relationships between two individuals like siblings, cousins, or second cousins. For both identical by state methods, they proved to work better for individuals with a higher degree of relatedness like siblings rather than second cousins (Kling 2019). However, both the segment and the likelihood approach did have true classification rate above 90% for individuals related up to being second cousins (Kling 2019). The segment approach was the most successful with true classifications of unrelated individuals (Kling 2019).

Some important numbers we can infer from this article is the amount of DNA shared within each degree of familial relationship. Identical twins share 100% of their DNA while

parent-child and sibling-sibling both share 50% of DNA (Kling 2019). Grandparent-grandchild and aunt/uncle-niece/nephew share 25% of their DNA (Kling 2019). First cousins and great grandparent-great grandchild share 12.5% of their DNA while second cousins share around 3% of DNA (Kling 2019).

One way to try to improve any classification method is to combine all three approaches. This would mean that all classifiers would need to assign the same degree of relatedness to two individuals, and if this was not matched across the board, the individuals would be classified as "undefined" (Kling 2019). The high lack of agreement between methods shows why there will be a huge increase in individuals labeled "undefined". While the false classification rates do decrease, it is difficult to favor the combined approach due to the high "undefined" individuals (Kling 2019). With this whole study, one issue was that it was difficult to tell apart relationships that had a similar amount of average shared genome (Kling 2019). The classification rates do not currently hold enough weight to hold up in court but can be very useful in investigative purposes.

Another experiment was done to study how to connect linkage analysis in loci and kinship testing. This experiment took different combinations of STR and SNP marker sets and calculated likelihood ratios under the hypothesis that the two individuals were related and that they were not related (Zhang 2020). Some of the marker combinations were 40 STRs alone, 27 STRs with 91 SNPs, and 40 STRs with 91 SNPs (Zhang 2020). The subjects in this test consisted of different pairs of individuals, ranging from relatives such as grandparent to grandchild to full siblings (Zhang 2020). The goal was to study how accurate the probability of classifying two related or unrelated individuals could be (Zhang 2020). All three marker combinations were successful, with the 40 STR and 91 SNP having the strongest discrimination

power (Zhang 2020). This set also has proven that there is a possibility to fully separate full siblings from unrelated pairs (Zhang 2020).

Additionally, the impact of linkage on relationship testing was also evaluated. While the linkage effect wasn't directly correlated to likelihood ratios, it was affected by different relationships. This proved that linkage does need to be considered when analysts look at larger numbers of loci in kinship analysis (Zhang 2020).

It is also important to note the importance of mutations on relationship testing. It was concluded that in unrelated individuals, if mutations were ignored then the likelihood ratio would be underestimated (Zhang 2020). When multiple markers are used, multiple loci that are positioned closely on the same chromosome can be evaluated to calculate likelihood ratios. By combining both STR and SNP loci, scientists can find the most efficient method to solve complex kinship testing (Zhang 2020). Looking forward this theory can be applied not only to autosomal loci, but potentially loci on that of mitochondrial DNA or Y-chromosomes.

While genetic fingerprinting has most popularly been done on autosomal STR DNA, there are other ways to form these DNA profiles too. Mitochondrial DNA, or mtDNA, is the DNA that is located in the mitochondria of cells. mtDNA was first sequenced in 1981 and was known to contain an extrachromosomal genome as opposed to a nuclear genome (Amorim 2019). Mitochondrial DNA has its own benefits and deficits, but its differences are what make is useful for forensic investigations. In mtDNA, there is a lack of repair mechanisms and a low accuracy to replicate a template, which leads to a higher mutation rate (Amorim 2019). This allows for more variation in mtDNA sequences, which creates more discrimination between mtDNA sequences.

The mtDNA every person has is passed down from the maternal parent only, while autosomal DNA is a combination of both parents. Without potential mutations, mtDNA sequences of maternal relatives and siblings should all be identical (Amorim 2019). This means that a person has the same mtDNA sequence as their mother, grandmother, great-grandmother, and any maternal siblings. This can be helpful in cases where maternal relatives are known. While the mtDNA sequence between two brothers with the same maternal lineage cannot be distinguished, it is useful to use as a reference or when ruling suspects out. Although sperm does contain a few mitochondria, the whole male genome is destroyed and disappear during or after fertilization (Amorim 2019).

When looking at mtDNA it is just as important to compare a reference sample to an unknown. It is feasible to exclude a sample originating from a known when the sequences within the two samples are very different. If two mtDNA sequences are identical, we can confer that they may have the same maternal lineage (Amorim 2019). One example of an individual that can be considered heteroplasmic is if they carry more than one mtDNA type within the same cell (Amorim 2019). This can be due to a length polymorphism or point mutation substitutions (Amorim 2019). The point substitution is more useful in forensic investigations and has more research performed to lend to its credibility. There is not much research behind heteroplasmy, which raises issues when interpreting mtDNA evidence. Furthermore, there are different ways that heteroplasmy can be exhibited in a person: heteroplasmic in one tissue and homoplasmic in another; or show more than one mtDNA profile in one tissue and a different profile in another (Amorim 2019).

The DNA Commission of the International Society of Forensic Genetics updated the guidelines and recommendations regarding mitochondrial DNA typing in 2014, including

mtDNA database searching (Amorim 2019). One of the more popular mtDNA databases is the European DNA Profiling (EDNAP) Mitochondrial DNA Population Database (EMPOP) (Amorim 2019). EMPOP is a database that stores mtDNA haplotypes that can be used for reference comparison of individual DNA or as a population database (Amorim 2019). *Mitomap* is another mtDNA online database that contains the human mtDNA variation and specific disease variants of the mtDNA (Amorim 2019). *Mitomap* can be used to study human mitochondrial DNA, haplogroups, and allele frequencies (Amorim 2019). *Mitomaster* is a program used that can identify polymorphic regions of mtDNA and calculate the variant statistics (Amorim 2019).

Another means of analyzing DNA is through Y-chromosomes, which are different yet equally important to mtDNA with male DNA profiles. Y-STR haplotyping is a much more suitable way to identify male contributors in STR profiling (Kayser 2017). We can search the source Y-STR haplotypes on the databases against the reference samples to calculate the match probability (Kayser 2017). By creating the Y-STR haplotypes, we can exclude male suspects from a crime, identify the paternal lineage of a suspect, find leads for suspects, or count the number of contributors to a sample (Kayser 2017). Not only is Y-STR analysis useful in crime scenes but can also be used in paternity disputes or for kinship analysis.

On the Y chromosome of males, STRs can be studied in forensic investigations. Y-STRs can be used to characterize paternal lineage of males because only males have the Y chromosome, and this gets passed down from male to male (Kayser 2017). Even if the father is deceased, paternity can still be determined from a paternal relative, such as an uncle for example. The father and the father's brother both received the Y chromosome from the same grandfather, so it is expected that the son would have a similar Y chromosome to the uncle, if paternity was

confirmed. These tests can only be successful if the Y-STR locations being analyzed have a low to medium mutation rate (Kayser 2017). Y-STR analysis can also be utilized in identifying males from human remains whether it be a missing person or from a natural disaster. If there is no specific individual in a database to compare to, the Y-STR profile of a close male relative can be used (Kayser 2017).

In 1992, the first polymorphic STR was discovered on the nonrecombining region of the Y-chromosome (Kayser 2017). The product rule is normally applied after a genotypic array is created and each allelic frequency is multiplied together (Buffolino, 2021, *Population Statistics*). Since the Y chromosomes is not homologous to the X chromosome, there is a large portion of it that is nonrecombining (Kayser 2017). Therefore, the product rule for allele frequencies that cannot be applied here, as with other autosomal loci. A complete haplotype frequency can then be created to calculate Y-STR match probabilities (Kayser 2017). A haplotype is a set of genetic information that has been inherited together from one individual. Y-STR haplotypes are directly transferred from father to son (Butler 2010). As expected, to help identify male contributors, the DNA at the scene must be compared to that Y-STR haplotype to a reference. The more Y-STR markers that are identified, the more accurately the paternal lineage can be estimated (Kayser 2017).

Because Y-STRs do normally consist of high mutation rates, it may be more difficult to characterize individuals as related. The more loci that can be accounted for, the more likely the risk that there is a mutation between the individuals (Quian 2017). Haplogroups can be marked as a set of co-ancestors with the same Y-SNPs and similar haplotypes. Because they have similar haplotypes, it can be assured the individuals are paternal relatives (Quian 2017). The Y-STRs especially, are a great marker for the pedigree searches due to their mutation rate (Quian

2017). These pedigrees can be formed with the aim to find relatives with an identical ancestral origin to date as far back as the most recent common ancestor (Quian 2017).

Two huge steps in analyzing Y-STRs to assist in creating these haplotypes occurred in the early 2000s. The first multiplex PCR kit that was created could develop up to 12 loci. Then the number increased to 23 loci and continues to increase (Roewer 2019). As of 2019, commercial Y-STR mutiplex kits are able to amplify 40 Y-STR sequences, which is a great number to use in forensic casework (Roewer 2019).

Additionally, the use of an online reference database with Y-STR profiles was established. Known as Y-Haplotype Reference Database, or YHRD, it includes more than 100 national databases in one place (Roewer 2019). This database does not assist investigators in finding the identify of unknown individuals but instead holds the information of anonymous Y-STR profiles (Roewer 2019). It is used to study larger populations to calculate allele frequencies. One method is called Analysis of Molecular Variance (AMOVA). This method can be used to describe the pattern of a Y-STR profile over a certain territory. YHRD can utilize this metadata to assign haplotype frequencies to Y-STRs (Roewer 2019).

The Discrete Laplace method can be used to calculate the probability of occurrence of a haplotype in any population (Roewer 2019). The Discrete Laplace method utilizes the YHRD to estimate the haplotype frequency with a better approximation than AMOVA. This can be extremely useful in studying DNA at a crime scene to figure out how common the Y-STR is and if it can help to narrow down a suspect pool. Another way to narrow down this suspect pool is to use the Y-based ancestry prediction from YHRD. This database will contain the probable ancestry of specific STR markers based on past submissions (Roewer 2019).

While Y-STR can be a great tool, there are some drawbacks. While it can help to identify a male compared to a close male relative, it may not be able to differentiate between two closely related male Y-STR haplotypes because they have the same paternal lineage (Kayser 2017). Due to this, it cannot allow for individual identification based on the Y-STR.

In the past, mutations in the Y chromosome may have been an issue, but today these mutations can be used to better separate and identify related males (Kayser 2017). There are times in which the Y-STR haplotype may be identical between two males. There can be two reasons that haplotypes are identical. Identical by state, or IBS, is due to similarities from random mutations. This means that two individuals who aren't related have similar haplotypes because one individual experienced a mutation (Roewer 2019). The second means is identical by descent, or IBD. These similarities in haplotypes are because two individuals are closely related so they have the same paternal lineage of the Y-chromosome (Roewer 2019).

With enough Y-STR markers, closely related male relatives can be separated based off mutations (Kayser 2017). These differences can be understood by studying the number of loci, the mutation rates of these loci, and the size of the database being studied. While the higher the number of markers and higher the mutation rate can lead to the belief that these two individuals are identical by descent, the issue of mutations do arise (Roewer 2019). The rapidly mutating Y-STR markers are useful is solving the question if similarities are due to IBD or IBD. This high mutation rate can help to further discriminate even males who are more closely related (Roewer 2019). This will then help to reduce the proportion of a coincidental IBS to the IBD haplotypes (Roewer 2019). One end goal of investigating with this method is to potentially reach the point where investigators can say "The suspect or someone from his family is the source of the trace. Persons unrelated to the family can be excluded as the source of the trace" (Roewer 2019).

The Y-chromosome can also be especially useful to provide information about where, geographically, an individual's ancestors were from (Kayser 2017). Especially when the male perpetrator is completely unknown, the investigator can use the Y-chromosomes to narrow down potential ethnicity. Because the Y-chromosomes lack recombination, it can be used for bio-geographic ancestry (Kayser 2017). Once a mutation has occurred in the Y-chromosome it will not be eliminated unless the male has no male offspring, because the female offspring will not obtain that Y chromosome (Kayser 2017).

Y-STR analysis is especially useful in sexual assault cases involving one female and one male. This is especially true when the DNA at a scene is a mixed sample between the male perpetrator and the female victim. This is because only the male perpetrator will have a Y chromosome, so when analyzing the Y-STR profile, it can be attributed to the perpetrator. Due to differential separation and preferential PCR amplification of the DNA components, the female epithelial cells, like from a vaginal swab, can be separated from the male sperm cells (Kayser 2017). This can even be done when the mixed stains have a high quantity of female DNA when compared to a low quantity of male DNA (Kayser 2017). This method proves to be extremely helpful when the analyst can say that one out of ten of these cases would have been deemed inconclusive without the use of Y-STRs (Kayser 2017). Especially when compared to the identification of male contributors when using autosomal STR profiling, the Y-STR haplotyping can be three times more suitable (Kayser 2017).

In the future, a predictable outcome is that the Y-STR kits will have more markers to account for even more accuracy. Both high and low mutation rates can be used for different aspects of Y-STR analysis. Low mutation rates of Y-chromosomes can be used for familial searching and kinship testing while locations of high mutation rates can be used for paternal

lineage identification (Kayser 2017). Another consideration for the use of Y-STR haplotyping in the future is predicting a male's surname from his Y-chromosome (Kayser 2017). Because many societies are patrilineal this means that the male surnames are transferred from the father to their offspring. It can be expected that, from father to son, there is co-ancestry of surnames and Y-chromosomes. One limitation to this is if a common surname is given to an unrelated man because they will share the same surname, but not Y-chromosome (Kayser 2017).

Amelogenin can further be used to determine if a DNA profile originated from a male or a female (Mateen 2020). Amelogenin is actually a gene that codes for proteins in tooth enamel (Butler 2010). It has a length polymorphism between its X and Y chromosome as it get copied and detected in analysis (Kayser 2017). The primers flank a six base pair deletion within one intron of the amelogenin gene of only the X chromosome. This means that during PCR the amplification of this area of the genome will result in a 106 base pair amplicon from the X chromosome and a 112 base pair amplicon from the Y chromosome (Butler 2010). The amelogenin gene is tested differently than any regular STR in any normal STR profiling kit. It can be used as a sole sex marker to infer the biological sex of the trace donor.

Other methodologies and techniques are already being explored for their use in a forensic role. Next Generation Sequencing (NGS) is a huge field that is still advancing in forensic science. It can be used to determine the sequence the human genome in a time and cost-efficient manner (Butler 2009). NGS+ is a combination of Next Generation Sequencing, capillary electrophoresis, and pyrosequencing. Y-STR haplotypes analysis was based on capillary electrophoresis, and Y-SNP haplogroup analysis was based on NGS and pyrosequencing (Quian 2017). The NGS+ system can then be utilized to type Y-STRs an Y-SNPs. This can help in forensic pedigree searches with Y-chromosome STR profiling. On the nonrecombining region

of the Y-chromosome, the Y-STRs are what is used to characterize paternal lineages (Quian 2017). This is because male relative usually share identical Y-STR haplotypes. As more Y-STR loci are studied, and relationships between males are more distant, the probability of encountering a mutation will increase (Quian 2017).

NGS+ can be used to divide current haplogroups through the typing of STR and SNPs. It is a highly suitable approach to construct a forensic pedigree search in a reasonable and efficient manner (Quian 2017). This study developed a formula, FSindex, to estimate the likelihood for each retrieved pedigree (Quian 2017). This value could then help to determine if a pedigree needed further investigation. FSindex takes into account both the Y-STR haplotype frequency and Y-STR mutation rate, and the resolution level of the Y-SNP haplogroups of the target population (Quian 2017). This allows for positive identification of pedigrees from mismatches Y-STR haplotypes whether it be from mutation or unrelatedness. NGS+ is a tool that can continue to help construct pedigrees based off the Y-chromosome.

Familial DNA analysis can be especially useful when conventional identification techniques aren't efficient. After evaluating the measure of similarities within two genotypes, the degree of relatedness can be inferred about the two individuals. The analysis of mitochondrial DNA can be done to trace maternal lineage. Because males can only pass down the Y-chromosome, Y-STR analysis can be used to trace paternal lineage. One goal in mitochondrial DNA, Y-chromosome, or autosomal DNA testing is to find related individuals to the contributor of a sample. Databases such as CODIS hold a lot of this genetic information, but consumer genealogy companies also hold similar data. Consumer sites will later be used to find an infamous killer by forensic genealogy testing.

Chapter 4: The Golden State Killer and His Effect on Forensic Genetics

Our current knowledge of DNA profiling (autosomal, mtDNA, and Y-STR) also allows us to solve cold cases that did not have these techniques at the time of their occurrence. In recent media, the most well-known cold case that has been solved with revolutions in DNA technology is that of the Golden State Killer. This utilized both familial DNA searching and the use of ancestry DNA databases to locate the eventual convicted murderer.

In California in the 1970's and 1980's, there was a series of rapes, robberies, and murders that took place over the span of 13 years (Garbus 2020). Investigators could not find who was responsible for these crimes at the time and the case went cold. Because of this, the individual brought so much terror onto the state of California. Victims were taunted, threatened, and lives were ruined. This made it even more hard-hitting that the case remained unsolved for over thirty years (Garbus 2020). Throughout his "reign," the perpetrator coined many different names such as the East Area Rapist, the Original Night Stalker, and finally his most well-known, the Golden State Killer.

Between 1973 and 1976, in Visalia, California, there were a series of burglaries that occurred. This individual coined the name of the "Visalia Ransacker" (Morford 2018). In Sacramento between 1976 and 1979, there were dozens of rapes that had occurred. These crimes were all attributed to the criminal by the name of the "East Area Rapist" (Morford 2018). While this was not the only serial criminal at this time, this was the most violent and brought on the most terror. Then, in southern California from 1979 to 1986, there was a series of murders (Morford 2018). This individual coined the name of the Original Night Stalker. This is not to

be confused with the Night Stalker, Richard Ramirez who was convicted in 1989 (Morford 2018).

In 2001, investigators were able to connect the crimes believed to be done by East Area Rapist and crimes done by the Original Night Stalker. They were able to match the DNA at some of these crime scenes to each other, meaning that the same person committed all these crimes (Garbus 2020). Afterward, the name of "EAR/ONS" or EARONS was used by the media to identify this perpetrator.

Traditional police work in the 1970s did not have the proper tools to evaluate all possible evidence. Investigators were able to connect many of these crimes to each other through classic police work. But this individual did not have their DNA in any system so they could not be identified by those means (Garbus 2020). Investigators reached dead end upon dead end and did not have any other scientific recourse to assist in discovering the culprit. Eventually, majority of these crimes stopped, and the cases went cold.

Many years later, investigators looked at the case of EARONS with fresh eyes and new scientific methods to find the killer. Instead of searching for an individual, investigators decided to search for the killer's family (Guerrini 2018). They inserted one of the DNA profiles from a crime scene into a site called GEDmatch under a fake name (Guerrini 2018). GEDmatch is an online service that collects DNA profiles for genetic genealogy research. It allows for individuals to input their own DNA to be analyzed and interpreted (Kennett 2019). By November 2018, when this DNA was submitted, there were about one million people in the database (Kennett 2019). GEDmatch generated a partial match between the DNA from the scene and a potential distant relative (Guerrini 2018).

With this new information, investigators were able to circle back to the classic police work and investigate distant relatives to the person whose DNA profile partially matched the sample from a crime scene. Investigators then created a family tree to their best ability (Mateen 2020). During their investigation, they were able to exclude family members through alibis, witness descriptions, and further investigating. They were able to rule out individuals based on age, sex, race, and where they lived. Eventually, they narrowed down the search to one possible lead, that being Joseph James DeAngelo, third cousin to the DNA match on GEDmatch (Mateen 2020).

Investigators collected his DNA from an object that he discarded under surveillance (Guerrini 2018). They used surreptitious collection of DNA. This is when agencies can collect DNA without a warrant when that DNA has been voluntarily discarded and is therefore "no reasonable exception of privacy" (Kennett 2019). This surreptitious DNA profile gathered from DeAngelo matched that of a DNA profile from a past evidentiary item (Mateen 2020).

It was very apparent how serious these crimes were and how badly not only local police but federal investigators wanted to find the man who did this. In 2016, the FBI shared some more information about the perpetrator and a $50,000 reward (Morford 2018). Then, in 2018 there was the familial match on GEDmatch. After investigators zeroed in on Joseph DeAngelo, they arrested him on April 24, 2018. He was charged with eight counts of first-degree murder with special circumstances. Then, on May 10th, he was charged with four additional counts of first-degree murder. Due to the statute of limitations, the time has expired for many counts of the rapes and burglaries for him to be convicted on these crimes (Morford 2018). Joseph DeAngelo was then arraigned on August 23, 2018. On June 29th, 2019 Joseph DeAngelo pleaded guilty to thirteen counts of first degree murder and special circumstances. Finally, on

August 21st, 2020 he received multiple consecutive life sentences without the possibility of parole (Morford 2018).

The "Golden State Killer" case was a turning point in forensics. It opened the doors of forensics to forensic genealogy (Garbus 2020). If it wasn't for the fear the Joseph DeAngelo created among women, families, and the whole state of California, the field of forensic genetics may not be as advanced in this field as it is today. Because of the torture and horror these victims experienced, it really pushed investigators to not give up on finding this man.

Genetic genealogy can be used to describe the use of DNA and genealogical research to form conclusions about relationships (Kennett 2019). This can be done on Y-chromosomes, mitochondrial, or autosomal DNA. DNA profiles can be tested, sequenced, and compared to other profiles. The hope is to find a certain partial match and predict the possible relationship based off the amount of DNA shared (Kennett 2019). It is easier to predict a closer relative like a second cousin, rather than a more distant relationship due to other factors (Kennett 2019).

One main difference between forensic and genetic DNA testing is the use of autosomal DNA. Forensic testing primarily used autosomal STRs while direct to consumer genealogy sites use autosomal SNPs to analyze DNA (Kennett 2019). This does further explain the fact that forensic tests are more likely to be performed on degraded DNA, while DNA for these genealogy sites are likely in prime condition. If a law enforcement agency wanted to access genealogy databases, they would have to re-run their DNA sample on the SNP chip to achieve a comparable profile.

Autosomal DNA analysis is widely used for familial DNA analysis. Partial matches of autosomal loci can help to identify relationships or narrow down potential individuals (Mateen

2020). The loci that are looked at are the loci that show variability between individuals and are on regions of chromosomes that are polymorphic (Mateen 2020). For example, many of these loci would be selected from a non-coding region, not from a region that works in relation to essential proteins in the body.

Genetic genealogy goes hand in hand with familial DNA testing. This strategy utilizes biological family members' DNA to identify an unknown individual (Mateen 2020). Because relatives will share a larger portion of genetic information, the unknown individual can be determined. By finding very similar matches to DNA, investigators can identify a familial relationship between an evidentiary sample and an individual and then investigate that family tree to find a related suspect (Mateen 2020). The more alleles shared, the closer in relation the two individuals are. By selecting 20 core loci, systems can find partial or full matches to solve the individuality of a sample. For example, parent and child will share more loci than child and grandparent (Mateen 2020).

The main goal is to be able to identify the relationship between and find a closely related individual to the perpetrator at a crime scene (Mateen 2020). While familial DNA analysis is not necessarily the "end all be all" of solving crimes, the breakthroughs from it are more than useful. It can better be called a "narrowing down technique" than a "match technique" (Mateen 2020).

Similar to the technologies used to identify relationships from parent to child, strategies can be used to identify relationships such as half-siblings or third cousins (Hill 2013). All these relationships can be combined and put together to form pedigrees and family trees. Better discrimination can be made between relationships by combining the information of shared regions at individual loci. Wrights Relationship is one method used to measure the degree of a

relationship between individuals (Hill 2013). For example, the relationship between uncle-nephew and half-sibling may appear to have the same Wrights relationship of 0.25 (Hill 2013). However, on the pedigree, each set of relationships would have a different visual appearance on a family tree. Potential pedigrees can be hypothesized with the qualitative description of the position, number, and length of shared regions on chromosomes (Hill 2013).

The position of shared segments is studied by looking at whether they include chromosome ends or not (Hill 2013). It was predicted that a single region will only share both ends if the relationship is close and will not share at either end if the relationship is distant (Hill 2013). As for the number of shared segments, more distant relatives will share fewer and smaller regions of DNA (Hill 2013). Lastly, the length of the segments that are shared is proportional to amount of genome that is shared. However, the degree to which the segment lengths are distributed is dependent on the pedigree relationships (Hill 2013).

Recently, direct-to-consumer genetic genealogy databases have been more widely used. This occurs in two aspects: any regular civilian just curious about their DNA; and for law enforcement agencies to identify unknown individuals through familial DNA (Kennett 2019). These databases can be used to identify missing persons, victims of mass disasters, or suspect in cold cases (Kennett 2019). Because each commercial genetic genealogy company uses different parameters for identifying individuals as identical by descent, the relationship predictions can vary from company to company (Kennett 2019).

GEDmatch, which was used in the Golden State Killer case, is a private website that was originally started by two genealogists looking to pursue their hobby (Kennett 2019). This website does not do DNA tests of its own but instead collects DNA profiles from many different

databases. GEDmatch compares the DNA profiles collected from other sites and combines them into a common ground database (Kennett 2019). For GEDmatch, there are three options when individuals are submitting DNA kits: public; research; and private mode (Kennett 2019). Public mode is intended for individuals who weren't genealogists to identify familial matches. Research mode is meant for law enforcement agencies so that investigators can see the matches, but it is not visible to civilians (Kennett 2019).

The privacy policy in GEDmatch is not very clear and alludes to the possibility that the site and information in the site can be accessed and subject to "unexpected users" (Kennett 2019). After the publicity behind the finding of Joseph DeAngelo, GEDmatch updated their policies so that law enforcement could only use their database to search for a perpetrator of a violent crime or to identify a deceased individual (Kennett 2019). At first, they also updated their terms and conditions so that users could choose to opt-out of the data index search by law enforcement (Katsanis 2020). After a different incident where GEDmatch matched the profile of a sexual assault to a minor, GEDmatch then updated its terms and conditions so that users would have to explicitly opt-in to the index search (Katsanis 2020).

In 2019, FamilyTreeDNA, another genetic genealogy database, announced that they will be working with FBI. They would allow the FBI to upload DNA profiles just as ordinary users would (Kennett 2019). Even after backlash about the fact that law enforcement was already using FamilyTreeDNA without the company's knowledge, the company actually advertised that by people submitting their DNA, criminals can be identified and caught (Kennett 2019).

On the other hand, MyHeritageDNA claims to resist law enforcement access (Kennett 2019). They only allow law enforcement to use the database with a court order or other legal

documentation (Kennett 2019). However, it can hard to regulate law enforcement access when the company has no knowledge of whether an individual is uploading a file from a non-standard source (Kennett 2019). A non-standard source could be from an individual whose DNA is unwillingly collected and submitted into a database without their knowledge or consent.

However, the use of these DNA databases could lead investigators down the wrong path if the genetic match is a false positive. In 2014, law enforcement used a Y-STR test from Ancestry.com to find a perpetrator in a rape and homicide case (Katsanis 2020). They had a profile of 35 Y-STR loci from semen recovered from the crime scene, and then subpoenaed the site to compare the Y-STR profile with those in their database (Katsanis 2020). The site identified an individual who matched 34 out of 35 loci (Katsanis 2020). Law enforcement then compelled the individual to give a DNA sample, which ended up excluding him as a suspect (Katsanis 2020). This was one of very many scenarios that proved that Y-chromosome based searched led to a concerning number of false positives (Katsanis 2020). In the same year, Ancestry stopped offering Y-DNA tests and shut down the Y-STR database (Kennett 2019).

Autosomal DNA relative matching is now what is offered by the majority of the genealogy sites (Kennett 2019). Both AncestryDNA and 23andMe do not accept transfer profiles from other sites and would only allow law enforcement usage if it is demanded for legal processes. They use saliva kits so this does make it more difficult for law enforcement to find a loophole in submitting a non-standard source (Kennett 2019). AncestryDNA focuses more on the ethnicity rather than genealogy. It compares the SNPs to that of other known ancestry SNPs and can give a value from their ethnicity estimator (Mateen 2020). The site 23andMe can also provide an ancestry estimation but has additional information about the potential genetic-related health problems of an individual (Mateen 2020).

While the public is relieved that criminals such as Joseph DeAngelo were captured, the means of capture does cause some concern. The technology in forensic genealogy has advanced faster than our ability to present safeguards to protect individuals' rights (Kennett 2019). The big question is if the privacy and security of an individual is maintained in these commercial DNA databases and whether the risk involved is worth the information obtained.

It is extremely important to note the difference between regular consumer databases like 23andMe and GEDmatch, where DNA is submitted voluntarily by individuals, and the Combined DNA Index System (CODIS). CODIS is a national database held by the FBI that holds genetic profiles of individuals who committed felonies, misdemeanors, that were arrested, and DNA from crime scenes (Guerrini 2018). Civilians do not have access to CODIS and law enforcement agencies do have strict guideline for accessing DNA profiles from this database. Currently, familial DNA searching is prohibited at the national level of CODIS, but it can be done at the state level CODIS databases (Sherrill 2017).

One huge issue is that someone who did not give their DNA to a consumer database is at risk of their DNA being analyzed or further investigated based on the profiles the database contains. If someone closely related to an individual has submitted their DNA to any type of database, then a large percentage of the individual's DNA is technically also in that system (Kennett 2019). Because close relatives share larger portions of DNA, investigators would have "part of your DNA profile" (Kennett 2019). This is the same process that explains how similarities in DNA can be used to find a perpetrator's relative (Kennett 2019). What is most concerning is that individuals may not know that this portion of their DNA is in a database, since it was not submitted voluntarily by themselves.

However, there are individuals who do voluntarily give their DNA because they are genuinely curious about some aspect of their genetic history which these websites offer to explain. But that doesn't mean that they are not due the same respect and privacy that individuals who didn't put their own DNA into deserve (D. Syndercrombe Court 2018). Many sites do include a vague indication that third party vendors may access information, but like this can get lost in the privacy terms or conditions that people rarely read (Guerrini 2018).

Law enforcement officers can and do find loopholes or other ways to get around the systems set in place by these consumer companies to protect privacy. There is almost no way of knowing if a DNA profile that is being submitted by Individual A is truly the DNA of Individual A or someone else (D. Syndercrombe Court 2018). In the case of Joseph DeAngelo, law enforcement avoided all forms of suspect consent by submitting the DNA under a false name (D. Syndercrombe Court 2018).

One other disadvantage of any system like this is that there could be misidentification. Things like adoption or misattributed parentage can alter the genetic genealogy and therefore misattribute alleles into a pedigree (Kennett 2019). This could lead to incorrect conclusions of parents, siblings, or even suspects. Although there is always further investigation after a lead, this process can breach the privacy and affect the livelihood of an innocent individual (Kennett 2019). One prime example of this is also from the Joseph DeAngelo case. When investigators found a distant relative of the perpetrators of these violent crimes of California, they found ten potential distant kin who shared genetic alleles with the perpetrator (Katsanis 2020). They then narrowed this down to two male relatives and ended up zeroing in on a different man first (Guerrini 2018). They investigated an elderly man in Oregon whose daughter uploaded his DNA into Y-search who proved to not be the perpetrator (Guerrini 2018).

Another common worry is about the inclusion of minors in familial searches. While sites like GEDmatch may have created limitations for individuals under 13 years old, there are no parameters set in place for individuals between 13 and 18 years old (Kennett 2019). People may be inserting the DNA of young children or minors into these databases as a makeshift paternity test. A further issue is if investigators include the DNA profiles of these minors when looking for a match for perpetrators (Kennett 2019).

A survey was taken to understand perspectives on police access to genetic genealogy websites and information gathered from these genealogy databases (Guerrini 2018). Out of the 1,587 participants, 79% supported the use of police searches of genetic websites to identify a relative from genetics (Guerrini 2018). Participants were supportive when these activities were done to find perpetrators of violent crimes (80% support), crimes against children (78% support), or missing persons (77% support) (Guerrini 2018). There was notably far less support to use these methods to identify perpetrators of nonviolent crimes, at only 39% support (Guerrini 2018).

However, the survey's demographics were not thoroughly explained, including whether or not the respondents used or submitted DNA into any of the consumer genealogy databases. The respondents of this survey may have also had bias due to their young age, relation to someone victim to a crime, or relation to a member of law enforcement (Guerrini 2018). There were many limitations to this study, but still provides useful information for future research aims. More studies and surveys do need to be made for a more accurate representation of the whole population

Another ethical issue is concerning the disadvantage that may be presented to minorities. Racial and ethnic minorities are already overrepresented in the CODIS database (Sherrill 2017). So, in this database for example, when familial tests are being carried out, it is disproportional because the search database does not meet the true population (Sherrill 2017). Therefore, there may be more matches in CODIS for a member of this minority because the search parameters are centered about a population filled with these minorities. Further ethics conversations need to be had and made known to companies and individuals who do decide to submit to DNA testing.

With time, comes more and more advancements in DNA technology and forensic science. Time and efforts are what made the capture of the Golden State Killer possible. Investigators inputted the DNA profile from a crime scene into a genealogy site and found matches. Through familial DNA testing, investigators found a family member in that database, and could therefore eventually find Joseph DeAngelo. As technology keeps improving and more knowledge is shared within the field of forensic science, there will be more and more advancements. Privacy issues will continue to be dissected as well as the ethical procedures regarding the use and search of DNA profiles within direct-to-consumer databases and CODIS.

CHAPTER 5: FUTURE OF FORENSIC GENETICS

The future of forensic genetics is endless. Based off the advancements that have been made in the past few years and with the improving technology, the conversation will focus more on the human genome and individuals that contribute to DNA samples.

Recently, one advancement has been the ability of DNA to predict visible characteristics and biogeographical ancestry from unknown samples (Palencia-Madrid 2020). This forensic DNA phenotyping is the prediction of the phenotype, or physical appearance, through determining externally visual characteristics from DNA (Palencia-Madrid 2020).

When comparing DNA phenotyping to STR loci analysis, there are some key differences. DNA phenotyping is especially useful when DNA identification isn't possible because the perpetrator lacks a reference sample, and relatives lack a reference sample in a DNA database. Regular STR analysis can only yield identification by a direct comparison of a crime scene DNA profile and the DNA profile of a suspect (Schneider 2019). The goal of DNA phenotyping is to simply narrow down the number of potential crime scene trace donors based off physical characteristics (Schneider 2019). This can be compared to familial DNA searching by saying that it is an investigative tool rather than a direct means of identification (Schneider 2019). Then, for any type of individual identification or probability markers to be used in court, STR profiling would be used (Schneider 2019).

VISAGE is the **VIS**able Attributes through **GE**nomics Consortium that aims to complete composite sketches of perpetrators solely based off DNA samples in the European Union (Palencia-Madrid 2020). Within the framework is the development of a panel consisting of 153 markers for predicting the physical appearance and ancestry (Palencia-Madrid 2020). The assay,

when validated, produced reliable and concordant results. It gave complete genotypes at of the SNPs and correctly genotypes 99.67% of the markers (Palencia-Madrid 2020). The most successful studies for the visible characteristics of humans were models for eye, hair, and skin color (Palencia-Madrid 2020). An additional trait that can potentially be deduced from DNA is the estimation of age (Schneider 2019).

By analyzing around 30 or more SNPs we can also estimate biogeographical ancestry (Palencia-Madrid 2020). Biogeographic ancestry does not allude to race or ethnicity, instead it explains the geographical region which an individual's biological ancestors originated (Schneider 2019). The DNA from an individual can show the genetic features that an individual inherited from their ancestors. The more genetic differences between two individuals directly correlates to the distance in geographic regions (Schneider 2019). Ancestry-informative DNA markers are specific DNA markers that are only seen in certain populations due to genetic isolation, migration, and local selection (Schneider 2019).

Ancestry-informative DNA markers can be used in three ways: Y-chromosomal, mitochondrial, or autosomal. Markers that are located on the Y-chromosome are only passed from father to son, so these markers will indicate the geographical origin of the paternal lineage whereas mitochondrial DNA markers are passed down only from the mother (Schneider 2019). In a similar matter, any mitochondrial DNA markers can indicate the ancestors coming from the maternal line (Schneider 2019). Autosomal DNA on the other hand, is passed down from both parents, and foes through recombination, meiosis, and fertilization (Schneider 2019). The autosomal markers are therefore half from the father and half from the mother (Schneider 2019).

DNA phenotyping can also be used in criminal investigations where there may be few or no eyewitnesses (Schneider 2019). Because we can still assign probability values to the phenotypic traits, it can be extremely useful to pinpoint potential perpetrators. For instance, a 95% probability of blue eyes can be much more reliable than a 75% probability of the same phenotypic trait.

As before, when there are more improvements in forensic DNA technology, there arise ethical concerns regarding these advancements. First and foremost is the invasion of privacy of people who may have contributed to a DNA sample, but more concerning may be those who did not actually contribute that DNA sample but fall under the phenotypic description generated. One of the greater risks is that phenotypic DNA will be discriminatory against minority groups (Schneider 2019). Appropriate measures need to be taken to ensure that police investigations use this data proportionately and transparently (Schneider 2019).

Another area of focus for the future field is Promethease. Promethease is a literature retrieval service that can be used to obtain health risk reports. In the case of Joseph DeAngelo, his genotype data was uploaded to Promethease. It is usually only used in investigators for reports about eye color and propensity to baldness (Kennett 2019). We can apply this same theory to conclude that this can also be done to look at a suspects propensity for various diseases as well (Kennett 2019). Promethease can actually generate wellness, health, or trait reports on personal genetic data or provide links to scientific literature that is relevant to that data (Guerrini 2018).

While new ideas and inventions are useful, the improvement of already used methods are just as important. Many methods are new, so forensic scientists are still working on ways to be more time and cost efficient.

While each DNA comparison technique is useful, like STR an SNPs, scientists have not yet found a common ground of using both. Once both methods can be studied collectively, as opposed to competitively, the chance of error can be reduced (Mateen 2020). With the release of SWGDAMs guidelines for the validation of probabilistic genotyping only being released in 2016, the validation of STRmix is still new (Bright 2016). Specificity, precision, accuracy, and reproducibility especially, are all aspects of STRmix that can be and will be improved as more time and experiments are conducted.

In probabilistic genotyping, there is a teetering line between random match probability and likelihood ratios. Both methods have their perks, but one method is not yet proved to be better than the other. With either method, there is the hope to add more correction factors and have more accurate results (Mateen 2020). Forensic scientists must be aware of the progress and limitations of every statistical approach to assist in future endeavors regarding forensic genetics.

The use of mitochondrial DNA testing is limited because it can only be used for maternal lineage, like Y-chromosomal analysis for the paternal lineage. The combination of these two analyses is the future of forensics so that pedigrees can be created to the highest degree of certainty possible. This, combined with autosomal DNA analysis should be used cohesively so that where one method lacks, another method can make up for.

The Golden State Killer is behind bars, but it was not the only cold case that can be solved with new technologies. Forensic scientists need to continue to look back at old cases and

use new ideas to solve cases. Each day, there are more cases being solved by law enforcement because of familial DNA testing.

Direct to consumer sites such as 23andMe, GEDmatch, and AncestryDNA are becoming more and more popular as individuals are more interested in their heritage and DNA. As more DNA samples are submitted into these sites, more individuals are putting their privacy in jeopardy when there is limited protection set in place at the time. The legal, ethical, and social issues surrounding familial DNA analysis is pressing. Uniform policies and laws are needed for the continuation of usage of direct-to-consumer site to exist.

DNA has revolutionized forensic science. It holds a similar weight to fingerprints due to the individuality that follows these evidentiary items. DNA is often left at a crime scene in some aspect, whether it is intentional or not. After investigators and detectives collect this DNA from a crime scene it is later tested. DNA is then extracted, quantitated, amplified, and then analyzed. DNA profiles that are created from can help to identify the individual that contributed to that sample. This can be done by comparing STRs to the federal database of CODIS. If this has no results, a more recent method would be to compare DNA profiles to find a close familial relative which can lead to the individual who contributed to that sample. This can be done with autosomal DNA marker, mitochondrial DNA, or the Y-chromosome. Statistical methods can then be employed to estimate the probability that the sample did or did not originate from the individual in question. These weights are important for the use of DNA in the courtroom so that the correct verdict can be made. The future of forensic science is currently centered about DNA technologies, and genealogy is a major tool in the toolbox.

Works Cited

Amorim A., Fernandes T., Taveira N.. 2019. Mitochondrial DNA in human identification: a review. Peer http://doi.org/10.7717/peerj.7314

Bille, Todd W., et al. "Comparison of the Performance of Different Models for the Interpretation of Low Level Mixed DNA Profiles." *ELECTROPHORESIS*, vol. 35, no. 21-22, 2014, pp. 3125–3133., https://doi.org/10.1002/elps.201400110.

Bright, J.-A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., & Buckleton, J. (2016). Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. *Forensic Science International: Genetics*, *23*, 226–239. https://doi.org/10.1016/j.fsigen.2016.05.007

Bright, Jo-Anne, et al. "The Variability in Likelihood Ratios Due to Different Mechanisms." *Forensic Science International: Genetics*, vol. 14, 2015, pp. 187–190., https://doi.org/10.1016/j.fsigen.2014.10.013.

Buffolino, Pasquale. "Breaking Traditions: DNA STR Interpretations." Long Island University Forensic Science Program. Probabilistic Genotyping Workshop, 8 May 2021, Brookville, Long Island University.

Buffolino, Pasquale. "Continuing With Continuous Models." Long Island University Forensic Science Program. Probabilistic Genotyping Workshop, 15 May 2021, Brookville, Long Island University.

Buffolino, Pasquale. “Population Statistics.” Long Island University Forensic Science Program. Probabilistic Genotyping Workshop, 8 May 2021, Brookville, Long Island University.

Butler, J. M., & Butler, J. M. (2010). *Fundamentals of forensic DNA typing*. Elsevier Academic Press.

Debus-Sherrill, S., & Field, M. B. (2019). Familial DNA searching- an emerging forensic investigative tool. *Science & Justice*, *59*(1), 20–28. https://doi.org/10.1016/j.scijus.2018.07.006

Dr. Barry Karger. (2011). DNA Sequencing Using capillary Electrophoresis. https://doi.org/10.2172/1013010

Garbus, L. (Director), & Wolff, E., Barry, K., Kane, M., Koury, J., Cogan, D., & Gaither, J. (Producers). (n.d.). *I'll Be Gone in the Dark* [Television series]. HBO.

Guerrini, C. J., Robinson, J. O., Petersen, D., & McGuire, A. L. (2018). Should police have access to genetic genealogy databases? Capturing the Golden State Killer and other criminals using a controversial new forensic technique. *PLOS Biology*, *16*(10). https://doi.org/10.1371/journal.pbio.2006906

Hill, W. G., & White, I. M. (2013). Identification of Pedigree Relationship from Genome Sharing. *G3 Genes|Genomes|Genetics*, *3*(9), 1553–1571. https://doi.org/10.1534/g3.113.007500

Jeffreys, Alex J, et al. “Hypervariable 'Minisatellite' Regions in Human DNA.” *Nature*, vol. 314, no. 7, 7 Jan. 1985, pp. 67–73.

Katsanis, S. H. (2020). Pedigrees and Perpetrators: Uses of DNA and Genealogy in Forensic Investigations. *Annual Review of Genomics and Human Genetics*, *21*(1), 535–564. https://doi.org/10.1146/annurev-genom-111819-084213

Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human Genetics*, *136*(5), 621–635. https://doi.org/10.1007/s00439-017-1776-9

Kennett, D. (2019). Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Science International*, *301*, 107–117. https://doi.org/10.1016/j.forsciint.2019.05.016

Kling, D., & Tillmar, A. (2019). Forensic genealogy—A comparison of methods to infer distant relationships based on dense SNP data. *Forensic Science International: Genetics*, *42*, 113–124. https://doi.org/10.1016/j.fsigen.2019.06.019

Mateen, R. M., Sabar, M. F., Hussain, S., Parveen, R., & Hussain, M. (2021). Familial DNA analysis and criminal investigation: Usage, downsides and privacy concerns. *Forensic Science International*, *318*, 110576. https://doi.org/10.1016/j.forsciint.2020.110576

Morford, M., Ferguson, M. (Host). (2018). In Criminology [Audio Podcast]. Emash Digital.

Nwawuba Stanley U, Mohammed Khadija A, Adams Tajudeen B, Omusi Precious I,

Ayevbuomwan Davidson E. Forensic DNA profiling: autosomal short tandem repeat as a prominent marker in crime investigation. *Malays J Med Sci.* 2020;**27(4)**:22–35. https://doi.org/10.21315/mjms2020.27.4.3

Palencia-Madrid, Leire, et al. "Evaluation of the Visage Basic Tool for Appearance and Ancestry Prediction Using PowerSeq Chemistry on the MISEQ FGX System." *Genes*, vol. 11, no. 6, 2020, p. 708., https://doi.org/10.3390/genes11060708.

Panneerchelvam, S., & Norazmi, M. N. (2003). FORENSIC DNA PROFILING AND DATABASE. *Malaysian Journal of Medical Sciences, 10*(2).

Puch-Solis, Roberto, et al. "Evaluating Forensic DNA Profiles Using Peak Heights, Allowing for Multiple Donors, Allelic Dropout and Stutters." *Forensic Science International: Genetics*, vol. 7, no. 5, 2013, pp. 555–563., https://doi.org/10.1016/j.fsigen.2013.05.009.

Qian, X., Hou, J., Wang, Z., Ye, Y., Lang, M., Gao, T., … Hou, Y. (2017). Next Generation Sequencing Plus (NGS+) with Y-chromosomal Markers for Forensic Pedigree Searches. *Scientific Reports, 7*(1). https://doi.org/10.1038/s41598-017-11955-x

Roewer, L. (2019). Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA. *Wiley Interdisciplinary Reviews: Forensic Science*. https://doi.org/10.1002/wfs2.1336

Schneider PM, Prainsack B, Kayser M: The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. Dtsch Arztebl Int 2019; 116: 873–80. DOI: 10.3238/arztebl.2019.0873

Syndercombe Court, D. (2018). Forensic genealogy: Some serious concerns. *Forensic Science International: Genetics*, *36*, 203–204. https://doi.org/10.1016/j.fsigen.2018.07.011

Taylor, Duncan. "Using Continuous DNA Interpretation Methods to Revisit Likelihood Ratio Behaviour." *Forensic Science International: Genetics*, vol. 11, 2014, pp. 144–153., https://doi.org/10.1016/j.fsigen.2014.03.008.

Zhang, Q., Zhou, Z., Wang, L., Quan, C., Liu, Q., Tang, Z., … Wang, S. (2020). Pairwise kinship testing with a combination of STR and SNP loci. *Forensic Science International: Genetics*, *46*, 102265. https://doi.org/10.1016/j.fsigen.2020.102265