JOURNAL LA LIFESCI

VOL. 05, ISSUE 01 (049-063), 2024
DOI: 10.37899/journallalifesci.v5i1.1171

DNA Profiling, Bioinformatics and Databases in Forensics: Human Identification Purposes

Andi Nur Sakina Tri Meilana[1,2], Elza Ibrahim Auerkari[1,2]

[1]*Division of Forensic Odontology, Department of Oral Biology, Faculty of Dentistry, Universitas Indonesia, Jakarta-Indonesia*
[2]*Department of Oral Biology, Faculty of Dentistry, Universitas Indonesia, Jakarta-Indonesia*

**Corresponding Author: Elza Ibrahim Auerkari*
Email: elza.ibrahim@ui.ac.id

Article Info

Article history:
Received 5 February 2024
Received in revised from 28 February 2024
Accepted 13 March 2024

Keywords:
DNA Profiling
DNA Markers
DNA Interpretation
Bioinformatics
Forensic DNA Databases

Abstract

Individual identification is an essential element in establishing truth the forensic investigation process, be it in criminal, medico-legal, or mass disasters case. When ante-mortem data are not available, the only thing that can be used is identification trough DNA. Alec Jeffrey, a geneticist (1984), found that certain regions of DNA contain repetitive sequences and the number of repetitions in individuals differs from one another. This identification method known as DNA profiling. DNA profiling is described as an important and significant discovery in forensics and has been considered as the standard in modern human identification. Until now, the recommended DNA profiling method must be based on the PCR to analyze degraded DNA and short-sized DNA (Short Tandem Repeats) through PCR amplification. In profiling, the result DNA amplification are analyzed through genetic markers (DNA Markers) and then matched into the database or compared with the collected reference samples. In the human genome, the genetic markers most frequently used in forensics are autosomal STR, Y-STR, mtDNA, X-STR, SNPs, and Amelogenin. Accomplishment of proving the suitability of DNA profiles, an application of bioinformatics in forensics is carried out. Bioinformatics is a combination of molecular biology and computer informatics that aims to manage and analyze data and store biological (genetic) information. DNA database is an application of bioinformatics in the form of special software which has become an important tool for biologists and forensics. This genetic database will predict the similarities between one DNA profile and another.

Introduction

Individual identification is an essential element in establishing the truth in the forensic investigation process, be it in criminal, medico-legal, or mass disaster cases (Bukyya et al., 2021). Approximately 9.2% of the remaining bodies of individuals in several countries are unidentified, making this a global humanitarian crisis. The world has witnessed a series of natural and non-natural disasters killing many individuals. In some situations, where the bodies have been fragmented, charred, or decomposed, the identification process becomes challenging. Dental identification plays an important role in such forensic cases as teeth and jaws are resistant to various changes and extreme temperature conditions. Therefore, an identification procedure will usually involve a combination and cooperation between investigators, dentists, and pathologists (Heathfield et al., 2021; Manjunath et al., 2011).

In situations where ante-mortem data that should be used in the identification process is unavailable, the only method of detection that can be used is DNA identification (Manjunath et al., 2011). Biological materials such as blood, semen, saliva, hair will be collected as much

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

as possible from the individual or from the scene, then analyzed and compared with available databases or matched with reference samples. In some instances, teeth and bones are often the only source of DNA available when the body has been damaged or degraded. When compared, teeth is considered a better source of DNA than bone due to its composition and position within the jawbone and the enamel sheath and alveolar bone matrix providing a protective barrier that keeps DNA viable.

Now DNA profiling is deemed as the standard in human identification in this modern era and is considered as the most important and significant invention in the forensic world after the fingerprinting method (Heathfield et al., 2021; Ramlal et al., 2017; Machado & Granja, 2020). The development of DNA-related studies for individual identification depends on a broad zone between genes called the "DNA Non-Coding" region. This zone specifies certain chemical sequences that are considered unique to each individual, resulting in a "genetic fingerprint" or what is known as a DNA profile. A genetic fingerprint can clarify whether different samples come from the same individual or not, and can also describe a person's biological relationship as the DNA owned by each individual is unique, with the exception of identical twins. In DNA profiling, several steps should be taken. Starting from sample extraction to interpretation of a DNA profil. This procedure must be based on the PCR method in order to analyze degraded and short DNA (Machado & Granja, 2020).

One of the most controversial issues in the use of forensic DNA evidence is how to estimate the likelihood that two DNA profiles match each other. To ensure this successfully, the application of bioinformatics in forensics is carried out. Bioinformatics is an application of molecular biology, and computer informatics which aims to manage and analyze data and store biological (genetic) information. DNA databases are applications of bioinformatics which is a specialized software that has become an important toolkit for biologists and is becoming increasingly popular among forensics. This database can be applied in the analysis of genetic diseases, genetic fingerprinting in forensic cases, or genealogy. The combination of DNA profiling and database matching has proven to be a very effective forensic kit in proving forensic cases. In the matching, the results of DNA profiles from biological sources that have been analyzed will be inputted into the forensic database to determine the hypothesis of the DNA sample in the form of whether there is a match or not and how significant the match is (Bianchi & Liò 2007; Gauthier et al., 2019; Tan et al., 2022).

DNA as a Source of Forensic Genetics

Deoxyribonucleic acid (DNA) is described as the “blue print of life” because it contains all the information about living things. DNA consists of two parallel spiral strands called nucleotides and forms a double helix along the length of the chromosomes. Each nucleotide is composed of 3 smaller chemical components: a pentose sugar molecule, a phosphate group and a nitrogenous base. Nitrogenous bases are an important identifying part of nucleotides and have different shapes from one nucleotide to another. Consists of adenine (A), thymine (T), guanine (G), and cytosine (C). Adenine and Guanine are referred to as purine bases, while Thymine and Cytosine are referred to pyrimidine bases. Adenine pairs with Thymine, and Cytosine pairs with Guanine. The sequences of this nitrogen bases will differentiate between one individual and another. DNA is an important genetic resource in forensics because it is identical in all individual cells, chemically stable, and biologically carries genetic instructions and is a medium where instructions will be transmitted to the next generation, so it is said that DNA must be inherited from parents (Bukyya et al., 2021; Angers et al., 2019).

In human cells, DNA is found in two locations: the nucleus and mitochondria. Nuclear DNA is the result of recombination from both parents, while mitochondrial DNA (mtDNA) is only

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

inherited by the mother. In the nucleus, DNA is organized into 23 pairs of chromosomes, 22 chromosomes are called autosomes, chromosomes that carry individual traits or characteristics and 1 pair of chromosomes called sex chromosomes carry sex characteristics (XX female and XY male). Each chromosome is numbered from 1-22 and has different lengths of long and short arms (example, chromosome 1 is a long chain with 200 million nucleotides, while chromosome 22 have about 50 million nucleotides). One pair of chromosomes (homologous genes) can be identical or different, and each form at a particular locus is called an allele, each of which is inherited one allele from the mother and one allele from the father. When the two alleles are the same it is called homozygous and when the two alleles are different it is called heterozygous. Nuclear DNA has about 3.2 billion base pairs and 2 copies per cell. While mt-DNA is circular in shape, consisting of about 16 thousand base pairs and 100 copies per cell.[9,10] Based on its structure and function, the human genome is classified into several types: (a) Coding Regions: Regions in DNA that encode and regulate protein synthesis, contain important information for cells to make proteins. Humans have about 20,000-25,000 genes and 1.5-2% of the genome is involved in protein coding; (b) Non-coding Regions : Regions in DNA that are not involved in protein synthesis (Approximately 23.5% of the genome) and is usually involved in gene regulation including promoters, receptors, and signalling polyadenylation; (c) Extragenic Region : About 75% of the genome is extragenic, 50% consists of repetitive DNA (consisting of 3 different types, namely satellite, minisatellite, and microsatellite DNA) and 45% interspersed repeats (short, long, long terminal, and DNA transposons) (Bukyya et al., 2021).

It is clear that not all DNA molecules encode amino acids protein, but only about 1,5% of the genome. In DNA to RNA changes, RNA splicing occurs to remove any nucleotide sequences that are not used in the formation of proteins. At this stage, the intron segment (non-coding region) is cut in the mRNA chain, while the exon segment (coding region) is a sequence of nucleotides in mRNA that will be forwarded for protein formation (Xiang-Dong. 2014). Non-coding regions have high polymorphism and hypervariable loci so that often used in forensic purposes. There are 3 main varieties of polymorphisms present in the genome: (a) Minisatellite (Variable Number Tandem Rep eat - VNTR): A polymorphism based VNTR locus that is about 20-100 repeated base pairs with 100 x repetition; (b) Microsatellite (Short Tandem Repat - STR): STR locus based on polymorphism which has a length shorter than minisatellite, which is about 2-8 base pairs with 2-20 x repetitions; (c) Single Nucleotide Polymorphism (SNPs): SNPs loci are loci where there is variance in individuals at certain positions in the genome. SNPs are classified and differentiated based on a single base change. Occurs approximately every 1000 base paisr (Bader, 2020).

History of Forensic DNA

In 1944, Oswald Avery defined the role of DNA as the carrier of inherited traits (generational transfer). Then in 1980, David Botstein and colleagues first exploited the small variations found between individuals at the genetic level as markers to construct human gene maps. At that time the first polymorphic locus was reported. This particular type of variation is referred to as restriction fragment length polymorphism (RFLP). Furthermore, in 1983, an important development in forensic genetics came with Kary Mullis, a chemist who discovered the PCR (Polymerase Chain Reaction) method that could amplify specific regions of DNA. In 1984 while looking for disease markers in DNA, Alec Jeffreys found that certain regions of DNA contained sequences that were repeated and the number of repetitions differed in individuals. Then the development of a method based on these findings was continued, resulting in a method

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

known as DNA typing, DNA fingerprinting, or DNA profiling (Bukyya et al., 2021; Rudin & Inman, 2001).

DNA Profiling

DNA profiling is a methodological step used to analyze and identify the genetic information contained in individual cells.[13]The main purposes of DNA profiling is to identify victims, associate body parts, and identify criminals. One of the bases for DNA profiling is based on the well-known fact that about 99.9% of DNA sequences between individuals are the same, and the other 0.1% are unique. The probability of two individuals having the same DNA profile is about 1 in 594.1 trillion. The non-coding regions of the human genome are said to be particularly suitable in forensic DNA profiling due to the informative and variability in individual identification. Most DNA profiling techniques target repetitive sequence regions and basically DNA profiling in forensic includes the following steps (Butler, 2011): (1) DNA sample collection; (2) DNA sample preparation; (3) Extraction of genetic material; (4) DNA quantitation; (5) Determination of target DNA (via marker DNA kit) and amplification of the DNA sequence region using PCR; (6) Separation of DNA fragments based on size via capillary electrophoresis method; (7) Data interpretation; (8) After the DNA profile is made, a comparison is made between the DNA sample profile (unknown) and the known reference DNA profile or compared to a DNA database search (Nwawuba et al., 2020; Butler, 2011).

The first DNA profiling method is Restriction Fragment Length Polymorphism (RFLP) which analyzes variations in the amount of length in homologous DNA sequences (Variable Number Tandem Repeat/ VNTR). This method involves using restriction enzymes to slice DNA around the VNTR region. Although highly polymorphic, this method has proven difficult to standardize interpretation between different individuals and different laboratories. Many attempts have been made, yet there is still uncertainty in the uniformity of interpretation of DNA analysis results. Therefore, the current recommended DNA profiling method should be based on the PCR method in order to analyze short-tandem repeat DNA through PCR amplification (Bader, 2016; Dai & Long, 2015).

In forensic purposes, the types of DNA samples are categorized into 2, which include target DNA (evidence) whose source is unknown and comparison DNA (reference) whose source is known such as samples obtained from first-degree relatives or personal belongings that have been confirmed (Tan et al., 2022; Butler, 2011). After the collection of biological material, in order to perform analysis, DNA must be extracted from the sample. *DNA Extraction* is the process of separating cells from a specific substrate. DNA extraction was first performed by Friedrich Miescher in 1869. Afterward, scientists continuously developed various extraction methods that were easier, cost-effective, faster, and yielded better results (Bukyya et al. 2021; Angers et al., 2019). Generally, extraction methods are carried out by organic methods (phenol), chelex method, and FTA paper (Butler, 2009). Of these extraction methods, the chelex method is the most commonly used method in many forensic laboratories because the price is more economical, and the process is faster than other methods (Bader, 2016).

The next procedure is *DNA quantitation*. Once the DNA has been isolated, a quantity and quality assessment must be performed. Determining the amount of DNA in the sample is very important and the amount of DNA expected to get maximum results is around 0.5 ng - 2.0 ng. Too much DNA can cause difficulties when interpreting the results and the procedure is time consuming, while too little DNA quantity can result in the loss of alleles required for analysis. A quantity of 1 ng of genomic DNA (1000 pg) will yield approximately 333 copies at each locus.

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

After the quantitation is done, the next step is the use of PCR (Polymerase Chain Reaction) which is a method to produce millions of specific copies of the targeted DNA sequence in a short period of time. Therefore, PCR is also often referred to as a biological photocopier (Bukyya et al., 2021; Butler, 2009; Kobilinsky et al., 2007). PCR amplifies specific regions of DNA that are planned to be analysed. This step is referred to as *DNA amplification*. In general, the DNA amplification cycle stage starts from denaturation, annealing, and extension. In the first stage, the sample is heated at around 94°C to separate the double-stranded DNA into single-stranded DNA (denaturation). This process usually lasts for 3 minutes, so that the DNA molecule becomes denatured into single-stranded DNA. Incomplete denaturation causes the DNA to renature, and the PCR process is unsuccessful. The sample is then cooled to a suitable temperature, which is around 60°C, so that the primers can bind each single-stranded segment on the DNA in the target region with its complementary base pair / amplify the template DNA in the region to be amplified (Anneal). After that, the temperature is then raised to about 72°C, where the DNA polymerase enzyme can extend or add nucleotide bases that match their respective complementary bases (extension), so that DNA replication occurs. In each cycle, the number of DNA molecules is doubled. A normal PCR cycle is between 28 - 32 repetitions. When the DNA is very low, the cycle can be increased. The overall PCR process lasts about 3 hours with each cycle taking about 5 minutes. After 30 cycles, around 1 billion copies of the target region on the initial DNA template have been generated and this PCR procedure is called amplicon (Bukyya et al., 2021; Kobilinsky et al., 2009).

The PCR products are then separated according to size through the electrophoresis method. *Electrophoresis* is a laboratory technique useful for separating DNA, RNA, or protein molecules based on their size and electrical charge. In liquid, DNA has a negative charge, so it will always move from the negative electrode (cathode) to the positive electrode (anode) when exposed to an electric field. Smaller DNA molecules will move faster than long molecules, thus separating DNA fragments by size, and molecules of the same size will stay together. Two types of electrophoresis methods can be used for DNA separation analysis, namely gel electrophoresis (GE) and capillary electrophoresis (CE) (Butler, 2009).

In gel electrophoresis, the 2 types of gels commonly used in molecular biology and forensic DNA laboratories are agarose gels, which have a fairly large pore size (used to separate larger DNA molecules such as RFLP methods; 600 bp - 23,000 bp) and polyacrylamide gels (used to obtain high-resolution separations for smaller DNA molecules such as STR alleles amplified by PCR; 100 bp - 400 bp) (Butler, 2009). The second method is capillary electrophoresis (CE), which has become a highly used electrokinetic separation method in the healthcare sector due to its speed of analysis, small sample size required, and excellent resolution and efficiency (Beauchemin, 2020). From the 90s, CE has been the modality of choice for DNA sequencing due to its high resolution and today, CE has become the main methodology used in separating and detecting STR alleles in forensic DNA laboratories worldwide. Some of the main instruments in CE are glass capillary tubes, buffer vials, and 2 electrodes connected to high voltage power. The device system consists of laser excitation, fluorescence detector, and computer, to control the injection and detect the sample. Basically, to achieve reliable DNA profiling, 3 things must be fulfilled. Spatial resolution to separate alleles that differ in size by one nucleotide, spectral resolution to separate colours from each other, and finally DNA size precision for calibration purposes (Butler, 2009; Voeten et al., 2018).

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

DNA Markers

Autosomal Short Tandem Repeats (STR)

STR is a short fragment of a repeating DNA sequence. The STR marker or known as a microsatellite is similar to VNTR, the difference is a long sequence of DNA STR is smaller about 2-6 pairs of bases (Bader, 2016). At a specific chromosomal location, a locus contains 2 DNA alleles each of which was inherited one from the father and one from the mother (Figure 1). These short fragments are found to be constantly repeating, and the number of such repetitions can identify the individual.

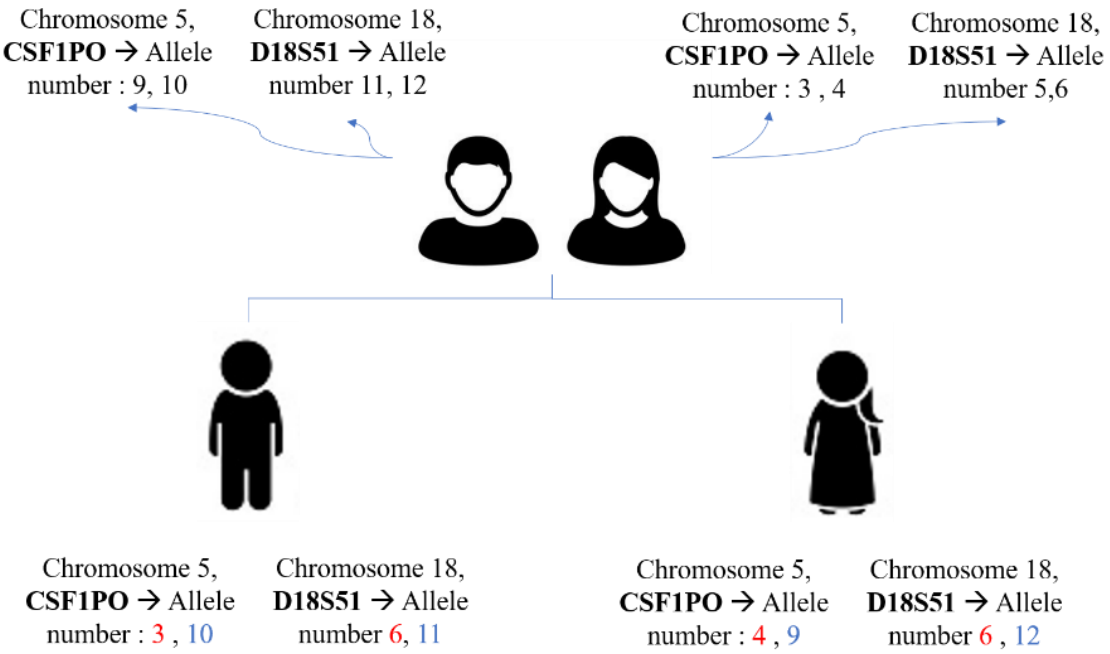

Figure 1. Illustration of parental and children STR inheritance

The genetic autosomal marker of STR was discovered in 1980 and is the most common DNA marker used in forensic DNA profiling. These markers are useful in molecular diagnosis, population studies, identity testing, paternity, and kinship cases. STR is highly polymorphic, therefore, the possibility that two individuals have the same number of repetitions on the DNA profile is extremely rare. For example, at the same locus, the nucleotide repetition sequence is CTAG. Although the sequence of nucleotide repetitions is similar, the number of repetitions of each individual will be different (Nwawuba et al., 2020; Stanley et al., 2020; Dash et al., 2021).

To obtain information and the creation of a valid database, a set of STR markers is proposed for analysis. The FBI laboratory in the United States (1997), selected 13 sets of STR loci (CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11) to be used as these markers. The results of the analysis of a set of STR loci are required in the US to upload a DNA profile to the National database. In 2011, the FBI announced that they would expand the STR core locus set to 20 sets of loci to improve international compatibility. It is also discussed in CODIS (The Combined DNA Index System) and the expansion of CODIS, the European Union (ESS), and INTERPOL (Dash et al., 2021; Butler & Hill, 2012).

Scientists basically unused their own STR tests. Usually, they will choose to use a commercial kit that has controlled quality. The kit will be equipped with an allelic ladder, sample positive

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

control, mixed reagents, as well as a primer targeting the specific locus to be amplified (target DNA). The kit makes it possible to detect more than 20 loci in a single PCR reaction and be amplified simultaneously. This is *referred to as multiplexing*. Here are some commonly used types of kits such as Identifier, Powerplex, AmpFISTR, Qiagen, etc. Currently, various world forensic labs generally use commercial kits consisting of 24 autosomal STR loci (CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D2S1338, D19S433, Penta D, Penta E, D1S1656, D2S441, D10S1248,D12S391, D22S1045, D6S1043, SE33) and amelogenin**.** [17,23] Each of these loci has definition, for example STR loci D5S818 and DYS19. In this case 'D' is DNA, 5 is for chromosome 5 and Y for chromosome Y, S means Single copy sequence, and the last number indicates the order in which the marker was found and categorized for that particular chromosome (Butler, 2011).

Y-STR (Paternal Inheritance)

The polymorphic region on the human Y chromosome provides a genetic marker in the form of a lineage called a single haplotype from father to son. A haplotype is a repetition of a short tandem (STR) of an allele on a Y chromosome. Parental traits of this marker can be useful in some forensic cases such as sexual assault, missing persons, identification of disaster victims, and kinship analysis. On the Y chromosome map, humans have 2 pseudoautosomal regions (PARs; PAR 1 & PAR 2) with short arm in the form of Yp and long arms in the form of Yq. The two are separated by a centromere. PARs are located at the telomere end of the chromosome and are regions of chromosomes that are recombined with the X chromosome. The large region that remains or does not recombine with the X chromosome contains about 70 genes in the male-specific region (MSY). This region is used in the determination of patrilineal (Syndercombe Court, 2021). A study recommended that the use of 9 core loci as markers of this DNA referred to as Minimal Haplotype Locus (MHL) including DYS19, DYS385 (a), and (b), DYS389 (I and II), DYS390, DYS391, DYS392, and DYS393. In 2003, the use of two additional loci, DYS438 and DYS439, was recommended by the Scientific Working Group on DNA Analysis Methods (SWGDAM). These two-locus groups combined are referred to as "SWGDAM core loci" (Bader, 2016).

mtDNA

mtDNA is an important marker in forensics because mitochondrial DNA has a high number of copies. There are about 100 copies of mtDNA in each cell, making this type of marker more sensitive than the STR autosome. mtDNA is descended from the maternal lineage. Human mitochondrial DNA is maternally inherited, therefore, each individual in the same maternal lineage has an identical mitochondrial DNA type (McCord et al., 2018).

X-STR

In the identification and analysis of some complex genetic relationships, the use of autosomal STR markers will probably result in uncertain or less effective conclusions. With the genetic transmission mode and the unique inheritance pattern of the X chromosome, the X-STR marker can be a necessary complement. The development of X-STR as well as the corresponding multiplex amplification system has successfully proved a complex case of individual kinship. For example, the case of mother-son, father-daughter, and mother-daughter relationships may be resolved more easily when X-STR is also used. X-STR can investigate cases that cannot be resolved by autosomal STR, such as the case of half-siblings and the relationship between grandfather and grandson (Zhang et al., 2021; Xiao et al., 2021). These markers use multiplex 19 X-STR fluorescent dyes namely DXS10079, DXS101, DXS10135, DXS10162, DXS6795,

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

DXS6800, DXS6803, DXS6807, DXS6809, DXS6810, DXS7133, DXS7423, DXS981, DXS9902, DXS9907, GATA165B12, GATA172D05, GATA31E08 and HPRTB along with amelogenin. The system is validated in accordance with the guidelines issued by SWGDAM (Jia et al., 2021).

Amelogenin

Amelogenin is a gene found on both the X and Y chromosomes so it is a marker used in determining sex. Sex determination depends on the presence of 2 homologs of amelogenin, namely AMEL X and AMEL Y which show polymorphism. This gene differs both in size and order to provide sex differentiation. AMEL X has deletions of 6 pairs of bases that contribute to the size difference between these 2 homologous genes. The amelogenin assay used a common set of primers and produced fragments with the length of each about 106 bp each for AMELX and 112 bp for AMEL Y (Dash et al., 2020).

Single Nucleotide Polymorphism (SNPs)

The recommended marker alternative to autosomal STRs is SNPs. Single nucleotide polymorphisms are common genetic variations that occur when nucleotides occur in a certain position in different genomes between members of the same species. This marker can be used when DNA samples are subject to high degradation. SNPs happen when the sequence of DNA variations changes. For example, the DNA sequence of AAGGCTAA becomes ATGGCTAA. Although this marker is not routinely utilized by forensic labs, this type of marker has more advantages than STR, such as its smaller size that less susceptibility to degradation (Bright et al., 2020). In SNPs, very small amplicon rates are useful in analyzing degraded samples, therefore, the potential mutation rate is lower than STR. In recent times, SNP has been widely used in various health sector applications such as medical diagnostics, population genetics, and human identity testing. When all profiling systems fail in identifying individuals then SNP can be used. The limitation of the SNP is its unstable locus and the multiplexing test requirements are large. The SNP likely cannot replace the locus STR that uses the DNA database. In general, the function of SNPs in forensics is divided into: (a) SNP Individual Identity Testing; (b) SNP information related to Pedigree; (c) SNP is ancestrally informative (SNP extension is informative to the lineage from generation to generation to provide information related to the ethnicity of the individual); (d) SNP information related to phenotype (Angers et al., 2019; Bader, 2016).

DNA Profile Interpretation

During the electrophoresis process, the results of the PCR product will be recorded on a computer based on size according to the time and intensity of fluorescence at various wavelengths. The computer-processed data is then plotted and will represent a graphic of the results of the detection of DNA fragments at each locus in the form of a colored peak along the X-axis (fluorescent tags attached to the PCR primer are used as color codes on the locus). These graphics are known as electropherograms (EPG/e-grams). Afterward, the EPG will be evaluated using STR genotype software to produce a table of final results that is a DNA profile. Each commercial DNA marker kit usually provides a special software that already has an allelic ladder in it and will determine the locus in the kit as well as the allele repeat number for each locus according to their respective standards (Bright et al., 2020; Butler, 2014).

In interpreting an electropherogram, the DNA profile must be in accordance with standardized guidelines, here are some important stages that must be carried out and determined: (1) Distinguish whether the chart is peaks or noise through the analytical threshold of the CE instrument; (2) Distinguish whether the peak is a representation of alleles or artifacts through

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

a stutter threshold; (3) Distinguish whether alleles are heterozygous, homozygous, or missing alleles; (4) Distinguish whether a DNA profile is sourced from one individual or the possibility of another DNA contributor (mixture); (5) Determining whether a DNA profile can be interpreted or not (full DNA profile / partial profile); (6) Comparing/matching DNA profile results; (7) Create a final conclusion regarding a potential match, accompanied by statistical interpretation to estimate its probability (Butler, 2014; Daeid et al., 2017).

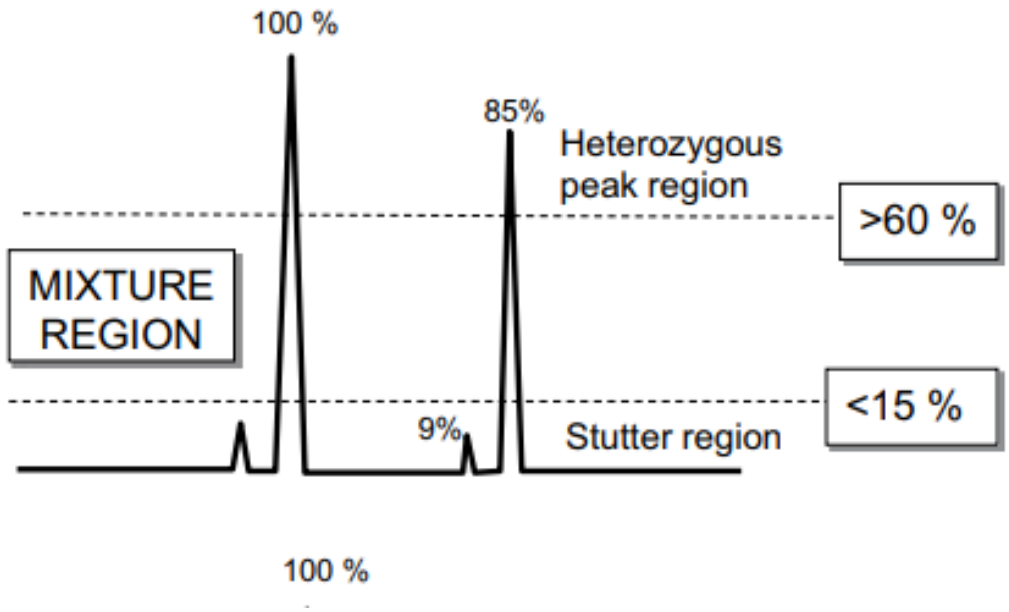

Figure 6. The relative peak height of the CE is an indicator in determining the presence of a sample mixture

If the highest peak at a locus is set at 100 %, then the heterozygous allele in the sample is at least greater by about 60%. Peak stutter is usually observed at low peaks of about <15% (Butler, 2014).

Normally, an individual's DNA profile results will show one or 2 peaks which are alleles in each locus. A locus featuring one allele indicates that at the locus, the individual inherits the same marker from both parents (homozygous). And when 2 alleles are displayed on a locus, the individual inherits a different marker (heterozygous). This interpretation process will then be repeated at all loci and compared with the results of the comparator/database DNA profile. By interpreting the DNA profile through EPG, we can determine whether the sample is of male or female origin and whether the sample comes from one individual or several individuals. In the locus of amelogenin which is the gender marker, when the sample comes from a woman, there is only one peak/allele appears. While in males will show 2 peaks with different heights each. And for other observed loci, the number of peaks will give an indication of the number of individuals contributing to a DNA profile.

When a DNA profile is obtained from a known sample, the number of DNA samples can be ascertained to be optimum, so the interpretation of the DNA profile may be easily and simple. However, when DNA profiles are processed from crime scene samples (unknown source) where the amount of DNA may be low, degraded, or even consist of many contributors, it will be difficult to conduct interpretation. This is often a challenge for analysts (Puch-Solis & Polis, 2021). *'Mixed profiles'* are often found in many forensic cases. It is common in a locus of DNA profile results will be found more than 2 alleles, so it is difficult to determine which genotype

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

contributed to a sample. Things will become more complicated when the DNA sample is low or degraded, as it will be susceptible to some stochastic effects. Therefore, it is difficult to distinguish between peaks that arise due to minor contributors or as a result of DNA samples that are not optimal (Gill et al., 2015). In interpreting complex DNA profiles, scientists usually often observe small peaks that are close to the baseline because there are difficulties in analyzing true alleles and artifacts, and there will be difficulties in determining the number of DNA contributors in a sample. When DNA degrades, it will become smaller pieces, and will probably result in a low peak height or even disappear. There are several stochastic effects that are often found in a DNA profile: (a) Locus drop-out: the entire locus failed to be amplified; (b) Drop-out alleles: Undetected alleles. This can occur as a result of one of the allele pairs on the locus failing to amplify to a detectable level; (c) Drop-in alleles: The addition of alleles to the locus due to sample contamination; (d) Stutters peak: Stutters is a peak with a low height or often referred to as a false peak (artifacts) (Daied et al., 2021; Gill et al., 2015).

All these effects can occur and increase when the amount of DNA is not optimal/degraded. Generally, the profile of a degraded DNA will display an overall decreased peak height with increasing molecular weight resulting in an allele failing to be detected/broken ("dropping-out") at a higher molecular weight locus. Thus, when this is found in a DNA profile result, the profile result will be considered a partial profile (Daied et al., 2021).

In interpreting a mixed DNA profile, special software can be used such as GeneMapper, DNAmixtures, Lab Retriever, STRmix, Lrmix, and TrueAllele software. [33,34] Scientifically, a mixed profile is observed when EPG shows 3 or more peaks at more than 2 loci. In cases of sexual abuse, the victim's profile can be taken separately and will be deducted with a mixed DNA profile. But it may be difficult to be done in other cases. A more complex statistical approach is needed in some cases. So that the mixed profile is usually less capable to be a piece of usable evidence. However, mixed profiles are inevitable in a DNA profile database, so the requirement for a profile may not contain more than 2 individuals. However, under the rules of the international exchange of DNA profile information under *Prum*, mixed profiles are not allowed (Angers et al., 2019).

The conclusion of the interpretation of DNA profiles in forensic cases is divided into 3: (1) Inclusion; Match: When the reference DNA profile is consistent with the DNA evidence profile, thus the individuals cannot be excluded as DNA contributors. Peaks in the compared DNA profiles have similar genotypes and there are no unexplained differences; (2) Exclusion; Non-Match: When the DNA profiles are both inconsistent, thus the individual is excluded as a possible contributor to the DNA sample evidence. On profile interpretation, genotype comparisons show differences in profiles described as samples coming from different sources; (3) Inconclusive this means that the results of DNA profiles cannot be inferred/interpreted / inconclusive. This can be due to various reasons such as the quality and quantity of DNA to mixed DNA of some individuals. These three conclusions are determined by the number of probabilities of each of them (Bright et al., 2020; Daied et al., 2017).

Bioinformation And Forensic Database

Bioinformatics is an application of molecular biology and computer informatics which aims to manage and analyze data and store biological (genetic) information, built on the basis of statistics and computational principles. Bioinformatics helps in the identification of mass disasters which must identify parent-child relationships and other kinship relationships. Bioinformatics can affect forensic statistics and found the statistical significance of the DNA profile (Bianchi & Liò, 2007). DNA database is one of the applications of bioinformatics in the form of a special software that has become an important toolkit for biologists and is becoming

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

increasingly popular among forensics. The database can estimate the probability of match or mismatch between two DNA profiles.

DNA database functions in forensics: (a) Deal with serious crimes such as murder, sexual assault, and assault; (b) Identifying potential criminals and linking crime scenes as part of an incident; (c) Identifying missing persons or unidentified bodies; (d) Release wrongly convicted individuals; (e) Deal with gang crimes, such as robbery and vehicle theft; (f) Deal with international crimes such as people smuggling, terrorism, and drugs

Forensic DNA databases in most countries generally contain 2 profiles: the reference profile of the individual, perpetrator / suspect (known source) and the profile of the crime scene (unknown source). Other categories of individuals that can also be included in the database are volunteers (can be witnesses or relatives of missing persons), missing persons and unknown deceased persons. The crime scene database should include samples of all types of biological material. In database searches, forensic profiles are searched to determine if there is an association between DNA profiles and perpetrators. To facilitate the search, assessment and retrieval of DNA profiles in the database, several requirements must be met, including: (a) DNA profiles must be associated with unique identifiers; (b) DNA profile should be of the best possible quality and high number of loci; (c) Setting standards for uploading partial profiles, such as the minimum number of loci; (d) The inclusion of other information, including personal information.

DNA Database Classification

The DNA database is divided into 2:

Criminal database-DNA

Contains DNA profiles originating from crime scenes, perpetrators or suspects aiming to solve crime cases by matching DNA. In some cases, this database also contains a DNA profile of a person involved in the investigation of a criminal case.

Non-criminal database

Contains DNA profiles of missing persons, their relatives, unidentified bodies. The non-criminal database also aims to link different body parts and missing parts such as in cases of natural disasters, war crimes and terrorist acts. This database can sometimes be compared to the criminal database in an attempt to identify victims. Some countries separate these two databases, while others have one integrated database (Heathfield et al., 2021; Angers et al., 209).

There are 3 international standards in the creation of DNA profiling and DNA databases and several international organizations that are mandated are the International Criminal Police Organization (INTERPOL), the Federal Bureau of Investigation (FBI) and SWGDAM which is a working group formed by the FBI, and the last is the International Standard Organization (ISO). For international search, the most effective strategy is that the country sends DNA profiles of missing persons to INTERPOL. When samples are not available, DNA profiles from relatives are sent to INTERPOL after approval. In the DNA database, there is an exchange of DNA profile data with each other.

GenBank

DNA sequence information around the world is recorded in a large computer database known as GenBank. GenBank was founded by the *National Center for Biotechnology Information* (NCBI), a division of the National Library of Medicine (NLM), located in the United States.

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

In 2019, GenBank recorded more than 6.25 trillion base pairs from more than 1.6 billion nucleotide sequences. DNA sequence information is stored not only from humans but about 450,000 different species represented in GenBank. (www.ncbi.nlm.nih.gov/genbank/). When a DNA profile is entered into GenBank, none of these new accessions will replace the existing accessions, and all accesses of a DNA sequence can still be used (Sayers et al., 2019).

CODIS

In 1989, the FBI launched the Combined Offender DNA Index System (CODIS). This system initially standardized 13 STR Loci known as CODIS Locus and there has been expanded to 20 loci STR. As of May 2015, more than 11 million individual profiles and about 630 thousand crime profiles recorded in CODIS. CODIS computer software will automatically search for indexes to match the DNA profile. The profile stored in CODIS contains the specimen identifier, sponsoring laboratory's identifier, initials or names of individuals related to the analysis, and actual DNA characteristics. The FBI has previously measured the success of the CODIS program and has proven it in several criminal cases. In CODIS, criminal history information, case-related information, and social security numbers or birth dates are not stored (Angers et al., 2019).

On October 13, 1998, the FBI officially launched the National DNA database. *The National DNA Index System (NDIS)* is part of the national-level CODIS containing DNA profiles contributed by participating federal, state and local forensic laboratories. NDIS is managed by the FBI and there is an exchange and sharing of DNA profile data. NDIS and CODIS contain more than 6.5 million STR profiles and link multiple countries. According to the 2019 FBI report, the NDIS contains more than 13 million DNA profiles of perpetrators, about 3 million prisoner profiles, and 900 thousand other forensic profiles. As of April 16, 2021, the 20 millionth DNA profile was donated to the US National DNA database via CODIS software. This is a major milestone in the number of recorded DNA profiles. Now, CODIS is installed in 203 federal, state, and local laboratories in the US to share DNA profiles with each other. CODIS software is also used by 58 other countries for their own law enforcement identification purposes.

Here are some of the National DNA Databases in the world, The UK National DNA Database (NDNAD), European Network of Forensic Science Institutes (ENFSI) (see http://www.enfsi.eu), The U.S National DNA Database, etc. When looking at DNA databases in several countries, they divided the CODIS into a 3-level schema: local, state, and national (Figure 6). For Local DNA index system (LDIS), scientists will enter the DNA profile that has been created and then it will be transmitted to the state level (SDIS) and last the national level (NDIS). Each local or state laboratory will manage the CODIS section and the FBI laboratory will manage the NDIS (Butler, 2009).

SNPs DATABASE

Once the SNPs are identified, some databases can be accessed online for genotyping interpretation. One example is Snipper *(http://mathgene.usc.es/snipper/)* which is an online tool for classifying and selecting marker sets, population groups, and algorithms. Manually, this SNP database can obtain information related to ancestry/race based on certain sequences by entering the existing information into Snipper. For example, a series of SNPs at a location would give the individual a European probability of 89.24%. In addition, based on SNP markers, it can also identify/predict individual hair and eye color. HIrisPlex *(http://hirisplex.erasmusmc.nl/)* is a software that can predict this from SNP alleles by using regression equations on population data.

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

mt-DNA DATABASE

A number of mt-DNA software databases can be accessed online for haplogroup determination. Scientists classify humans who have similarities in either the Y chromosome or mtDNA into genetic populations and refer to them as haplogroups. Phylotree is a regularly updated online phylogenetic tree of all mtDNA haplogroups. There are many mtDNA haplogroup classifiers based on Phylotree. Human Mitochondrial Database (HmtDB) is a database containing human mtDNA annotated with population data and variability data. This database offers queries on various criteria such as haplotype-defining SNPs, haplogroups, and geographic origin. Related databases and software for other mtDNA can be accessed at *(https://isogg.org/wiki/MtDNA_tools).*

There are 2 databases used for estimation, the first consists of the complete mtGenome sequence obtained from GenBank and the other is a virtual haplotype. Each with a mutation that defines a haplogroup, originating in the Phylotree. Phylotree was used to construct haplotype estimates, while the entire mtGenome sequence was used to calculate mutation stability applied to haplotyping calculations. In addition, EDNAP mtDNA Population Database (EMPOP) is a forensic mtDNA database that stores high quality and directly accessible data. EMPOP is the main hub for mtDNA analysis and provides a number of tools for analysis and interpretation *(http://empop.online/network)* (Liu & Harbison, 2018).

Conclusion

DNA profiling is a good method for forensic identification purposes. Has the ability to distinguish one individual from another, to distinguish gender, family lineage, even individual ancestry. Bioinformatics in forensics has recently greatly facilitated biologists and forensics in identifying individuals in both criminal and non-criminal cases. DNA database is an application of bioinformatics that acts as data management and analysis and storage of biological (genetic) information which is built on the basis of statistics and computational principles. With the development of forensic bioinformatics, it will greatly facilitate the matching of DNA profiles in the future. Forensic DNA analysis is performed worldwide. Therefore, it is very important for countries around the world to develop and compile their respective national DNA databases.

References

Angers, A., Kagkli, D. M., Oliva, L., Petrillo, M., & Raffael, B. (2019). Study on DNA Profilng Technology for Its Implementation in Central Schengen Information System.

Bader, S. (2016). *A guide to forensic DNA profiling*. John Wiley & Sons.

Beauchemin, D. (Ed.). (2020). *Sample introduction systems in ICPMS and ICPOES*. Newnes.

Bianchi, L., & Liò, P. (2007). Forensic DNA and bioinformatics. *Briefings in bioinformatics*, *8*(2), 117-128. https://doi.org/10.1093/bib/bbm006

Bright, J. A., Kelly, H., Kerr, Z., McGovern, C., Taylor, D., & Buckleton, J. S. (2020). The interpretation of forensic DNA profiles: an historical perspective. *Journal of the Royal Society of New Zealand*, *50*(2), 211-225. https://doi.org/10.1080/03036758.2019.1692044

Bukyya, J. L., Tejasvi, M. L. A., Avinash, A., P, C. H., Talwade, P., Afroz, M. M., ... & Srisha, V. (2021). DNA profiling in forensic science: A review. *Global Medical Genetics*, *8*(04), 135-143. https://doi.org/10.1055/s-0041-1728689

Butler, J. M., & Hill, C. R. (2012). Biology and genetics of new autosomal STR loci useful for forensic DNA analysis.

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

Butler, J. M. (2014). *Advanced topics in forensic DNA typing: interpretation*. Academic Press.

Butler, J. M. (2011). *Advanced topics in forensic DNA typing: methodology*. Academic press..

Butler, J. M. (2009). *Fundamentals of forensic DNA typing*. Academic press.

Puch-Solis, R., & Pope, S. (2021). Interpretation of DNA data within the context of UK forensic science—evaluation. *Emerging Topics in Life Sciences*, *5*(3), 405-413. https://doi.org/10.1042/ETLS20200340

Daeid, N. N., Rafferty, A., Butler, J., Chalmers, J., McVean, G., & Tully, G. (2017). Forensic DNA analysis: A primer for courts.

Dai, S., & Long, Y. (2015). Genotyping analysis using an RFLP assay. *Plant Genotyping: Methods and Protocols*, 91-99. https://doi.org/10.1007/978-1-4939-1966-6_7

Dash, H. R., Rawat, N., & Das, S. (2020). Alternatives to amelogenin markers for sex determination in humans and their forensic relevance. *Molecular biology reports*, *47*(3), 2347-2360. https://doi.org/10.1007/s11033-020-05268-y

Dash, H. R., Rawat, N., Vajpayee, K., Shrivastava, P., & Das, S. (2021). Useful autosomal STR marker sets for forensic and paternity applications in the Central Indian population. *Annals of Human Biology*, *48*(1), 37-48. https://doi.org/10.1080/03014460.2021.1877353

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2020). GenBank. *Nucleic acids research*, *48*(D1), D84-D86.

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in bioinformatics*, *20*(6), 1981-1996. https://doi.org/10.1093/bib/bby063

Gill, P., Haned, H., Bleka, O., Hansson, O., Dørum, G., & Egeland, T. (2015). Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development. *Forensic Science International: Genetics*, *18*, 100-117. https://doi.org/10.1016/j.fsigen.2015.03.014

Heathfield, L. J., Haikney, T. E., Mole, C. G., Finaughty, C., Zachou, A. M., & Gibbon, V. E. (2021). Forensic human identification: Investigation into tooth morphotype and DNA extraction methods from teeth. *Science & Justice*, *61*(4), 339-344. https://doi.org/10.1016/j.scijus.2021.05.005

Jia, J., Liu, X., Fan, Q., Fang, C., Wang, M., Zhang, J., ... & Yan, J. (2021). Development and validation of a multiplex 19 X-chromosomal short tandem repeats typing system for forensic purposes. *Scientific reports*, *11*(1), 609. https://doi.org/10.1038/s41598-020-80414-x

Kobilinsky, L. F., Levine, L., & Margolis-Nunno, H. (2007). *Forensic DNA analysis*. Infobase Publishing.

Liu, Y. Y., & Harbison, S. (2018). A review of bioinformatic methods for forensic DNA analyses. *Forensic Science International: Genetics*, *33*, 117-128. https://doi.org/10.1016/j.fsigen.2017.12.005

Machado, H., & Granja, R. (2020). *Forensic genetics in the governance of crime* (p. 114). Springer Nature. https://library.oapen.org/handle/20.500.12657/23264

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)

Manjunath, B. C., Chandrashekar, B. R., Mahesh, M., & Rani, R. V. (2011). DNA profiling and forensic dentistry–A review of the recent concepts and trends. *Journal of forensic and legal medicine*, *18*(5), 191-197. https://doi.org/10.1016/j.jflm.2011.02.005

McCord, B. R., Gauthier, Q., Cho, S., Roig, M. N., Gibson-Daw, G. C., Young, B., ... & Duncan, G. (2018). Forensic DNA analysis. *Analytical chemistry*, *91*(1), 673-688. https://doi.org/10.1021/acs.analchem.8b05318

Nwawuba, S. U., Momoh, S. M., & Nwokolo, C. C. (2020). Key DNA profiling markers for identification: A mini review. *Pharm Pharmacol Int J*, *8*(6), 337-343. DOI: 10.15406/ppij.2020.08.00315

Ramlal, G., Vevaraju, D., Vemula, A. Y., Swapna, T., & Bindu, P. H. (2017). Extrication of DNA from burnt teeth exposed to environment. *Journal of Clinical and Diagnostic Research: JCDR*, *11*(8), ZC120. https://doi.org/10.7860/JCDR/2017/26911.10525

Rudin, N., & Inman, K. (2001). *An introduction to forensic DNA analysis*. CRC press.

Stanley, U. N., Khadija, A. M., Bukola, A. T., Precious, I. O., & Davidson, E. A. (2020). Forensic DNA profiling: autosomal short tandem repeat as a prominent marker in crime investigation. *The Malaysian journal of medical sciences: MJMS*, *27*(4), 22. https://doi.org/10.21315/mjms2020.27.4.3

Syndercombe Court, D. (2021). The Y chromosome and its use in forensic DNA analysis. *Emerging topics in life sciences*, *5*(3), 427-441. https://doi.org/10.1042/ETLS20200339

Tan, W. C. D., Stasi, A., & Dhar, B. K. (2022). Forensic DNA profiling in the southern border provinces of Thailand: Ethical and regulatory issues. *Forensic Science International*, *336*, 111322. https://doi.org/10.1016/j.forsciint.2022.111322

Voeten, R. L., Ventouri, I. K., Haselberg, R., & Somsen, G. W. (2018). Capillary electrophoresis: trends and recent advances. *Analytical chemistry*, *90*(3), 1464-1481. https://doi.org/10.1021/acs.analchem.8b00015

Voeten, R. L., Ventouri, I. K., Haselberg, R., & Somsen, G. W. (2018). Capillary electrophoresis: trends and recent advances. *Analytical chemistry*, *90*(3), 1464-1481. https://doi.org/10.1093/nsr/nwu008

Xiao, C., Yang, X., Liu, H., Liu, C., Yu, Z., Chen, L., & Liu, C. (2021). Validation and forensic application of a new 19 X-STR loci multiplex system. *Legal Medicine*, *53*, 101957. https://doi.org/10.1016/j.legalmed.2021.101957

Zhang, Y., Yu, Z., Mo, X., Zhao, X., Li, W., Liu, H., ... & Sun, H. (2021). Development and validation of a new 18 X-STR typing assay for forensic applications. *Electrophoresis*, *42*(6), 766-773. https://doi.org/10.1002/elps.202000168

ISSN 2721-1304 (Print), ISSN 2721-1207 (online)