2022

A Tool for Biometric Interpretation of Forensic STR DNA Profiles

Ahmad Jamal Baroudi
ajbaroudi@mix.wvu.edu

A Tool for Biometric Interpretation of Forensic STR DNA Profiles

Ahmad Jamal Baroudi

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering/Electronic and Photonic

Jeremy Dawson, Ph.D. Chair

Natalia A. Schmid, Ph.D.

Tina Moroose

Department of Electrical Engineering

Morgantown, West Virginia
2022

Keywords: DNA Analysis, Signal processing, STR profiles, Degraded DNA, Artificial STR Generator, Alleles Match Score Matrix

Abstract

A Tool for Biometric Interpretation of Forensic STR DNA Profiles

Ahmad Jamal Baroudi

Rapid DNA biometric identification applications are becoming more essential and widely used in human identity validation processes. Despite their powerful identification capabilities, processing a sample to generate a forensic DNA profile still takes longer compared with other rapid biometric technologies. Methods used to speed up the analysis could lead to signal artifacts similar to those arising from low copy or degraded DNA samples, making the electropherogram unsuitable for forensic interpretation and analysis. The goal of this research effort is to apply biometrics and mathematical approaches to forensic STR (Short Tandem Repeat) profiles. To accomplish this goal, a multi-function software tool was developed to evaluate STR profiles in the form of electropherograms. This tool is capable of generating degraded and non-degraded STR profiles based on allele statistics from the human population using MATLAB.

The software also acts as an interface to apply a previously developed signal processing method to recover alleles in electropherograms produced from degraded DNA samples. The user interface offers the capability of visualizing and comparing those discovered peaks with the allelic ladder to confirm recovery or a rejection. The software is demonstrated on both artificial and real degraded STR electropherograms, indicating a higher allele recovery rate when compared with commercial GeneMapper IDx software. Finally, the software produces a match score based on the number of matching alleles when comparing two or more DNA profiles based on the number of existing and recovered allele peaks in the electropherogram.

Acknowledgements

I want to thank God “Allah,” the Almighty, for letting me through all the difficulties. I have experienced your guidance day by day. Thank you for all your blessings throughout my research work to complete the thesis successfully.

Firstly, to my caring, loving, and supportive wife, Sara: my deepest gratitude. Your encouragement when the times got rough is much appreciated and duly noted. My completion of this project could not have been accomplished without your continuous prayers, support, and understanding when undertaking my research and writing my thesis. My heartfelt thanks. I am also extremely grateful to my parents for their love, prayers, caring, and sacrifices in educating and preparing me for my future. Also, I want to extend my gratitude to my siblings for their endless encouragement, support, and motivation during my research and writing.

I want to acknowledge and give my warmest thanks to my supervisor Jeremy Dawson for allowing me to do research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity, and motivation have deeply inspired me. He has taught me the methodology to carry out the research and present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. I am incredibly grateful for what he has offered me. Besides my advisor, I would like to thank the rest of my committee members, Tina morose and Dr. Natalia Schmid, for their encouragement and insightful comments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Rapid DNA is described as a fully automated or hands-free process in developing a Core STR profile for CODIS from a reference sample in less than 90 minutes. This consists of automated DNA detection, separation, amplification, extraction, and analysis (allele calling) and has no requirement for human intervention once the sample has been collected and placed into the instrument. In 2010 the FBI established a program office to facilitate the development and integration of rapid DNA technology used by law enforcement, also known as the Rapid DNA Program Office. The program office works to ensure the coordinated development of this new technology among federal agencies such as the Department of Defense, the National Institute of Standards and Technology, the National Institute of Justice, and other federal agencies. The program office also facilitates the effective and efficient integration of Rapid DNA in the booking environment with state and local law enforcement agencies and state bureaus of identification through the FBI's Criminal Justice Information Service Division Advisory Policy Board.

Both forensic science and biometrics can be used to identify a person. The difference is how they are used and what is needed to apply them. In general forensic science is usually invoked after an event, whereas biometrics is typically used before the occurrence of the event; Due to the nature of the biometric systems, the biological traits that will be used are already known, Whereas, in forensics, the sample is meticulously extracted from a crime scene. The quality of evidence data obtained is typically lower, also in forensic cases, verbal reasoning is crucial. Computational efficiency is essential in biometric applications because recognition decisions in biometric systems must be rendered in real-time; Forensics does not require real-time recognition.

Every human possesses a unique genome made up of distinctive characteristics and qualities such as hair and eye colors, blood type, skin color as well as height, and weight.DNA can be used to distinguish one person from another. Genomic DNA can be found in a hair follicle, a blood sample, or skin. DNA contains genes. Genes have multiple forms located in the same position (or genetic locus) on a chromosome. A variant form of a gene is also known as an allele. Humans have two alleles at each genetic locus, one inherited from each parent. Each set of alleles depicts the genotype of a specific person. There are two types of genotypes; homozygous, when there are two identical alleles at a locus, and heterozygous, when the two alleles are different.

Multiple loci creates a DNA profile known for an individual. STR profiles identify people from each other by containing alleles at a locus; this makes a very distinctive pattern That will be different among individuals.

STRs are described as repeated segments of DNA that are typically 2-16 base pairs in length scattered throughout our genome. There is one STR allele at each locus coming from the father and one from the mother. In the human population, the number of repeats of each STR at each locus varies. The variability in repetitions makes STR DNA testing extremely valuable as a human DNA identification tool.

The result of many biometrics applications does not often require a 100 percent match, especially since it is comparing the data to an existing database. A biometric match can consist of a score that designates the similarity between the reference template and the sample. Typically, this match should never be identical; due to subtle changes over time and errors in data collection, freshly gathered samples will inevitably vary somewhat from the reference template. Given this information, a match with 70 percent of degraded data would be closer to a positive match in the database. Whereas a 40 percent match would leave you with many possibilities, making this useless and less valuable., If all 20 alleles are not present, applying signal processing will amplify the signal and improve the amount of information there. Then performing a matching analysis will get some idea of a match score. The obtained data can be used to identify a person solely. The final results will be presented to the person who makes the decision.

The FBI officially launched a nationwide DNA database on October 13, 1998. At the end of 2003, it was named the Combined DNA index system or CODIS and contained over 1.5

million STR (short tandem repeat) profiles [26]. This connected all 50 states with a DNA profile database like the FBI fingerprint database, revolutionizing the ability to use DNA profile information in linking crime scene evidence to perpetrators. [27] When testing the core loci of only 13 CODIS, the average random match probability is rarer than 1 in a trillion among unrelated individuals. This probability was based on the calculation of one million samples. Although this is an infrequent possibility compared to the world's population, there is still a chance that a match might occur knowing that the world's current population is a little over seven and a half billion.

Research in this area has rapidly improved over the last decade, leading to a powerful way to differentiate new loci in the human genome. This is done by using expanded DNA testing kits, increasing the discriminating power of DNA analysis.

1.2 Problem statement

There are several barriers to the use of DNA as a biometric. One issue is processing time. However, rapid DNA systems have been developed that can process DNA samples in 1.5hrs or less, and whole genome approaches may become even faster [28]. However, as with benchtop DNA analysis, these systems are designed to produce results with 'pristine' buccal samples, with called allele count dropping rapidly as the quality of the sample degrades due to either environmental exposure or a low starting amount of genetic material (low-copy samples). While these are significant issues in forensic applications of DNA analysis, biometrics applications of DNA as a unique identifier could benefit from partial information extracted from low-quality STR profiles that come from degraded and low-copy samples. This 'all or nothing' paradigm of forensic DNA analysis extremely limits the use of partial or incomplete profiles that could have significant utility in biometric DNA applications. A degraded DNA sample cannot identify a person if it is missing multiple alleles and/or loci. DNA's degradation can happen as a result of environmental damage prior to DNA extraction and analysis, leading to partial or incomplete DNA profiles. Incomplete DNA profiles have diminished uniqueness, especially in large datasets. Generally, these degraded samples will not be used in the identification process. There have been multiple attempts to match the degraded profiles with existing profiles in databases [29] [30][31]. And other attempts have been made using chemical techniques [32], but there have been no attempts to generate and recover STR profiles using FBI population records and probability ratios.

1.3 Electropherogram creation

The section below gives an overview of the DNA analysis process, from DNA sample collection to creating a forensic DNA profile in the form of an electropherogram signal. In the later sections of this introduction, we will discuss the challenges identified in this high-level overview.

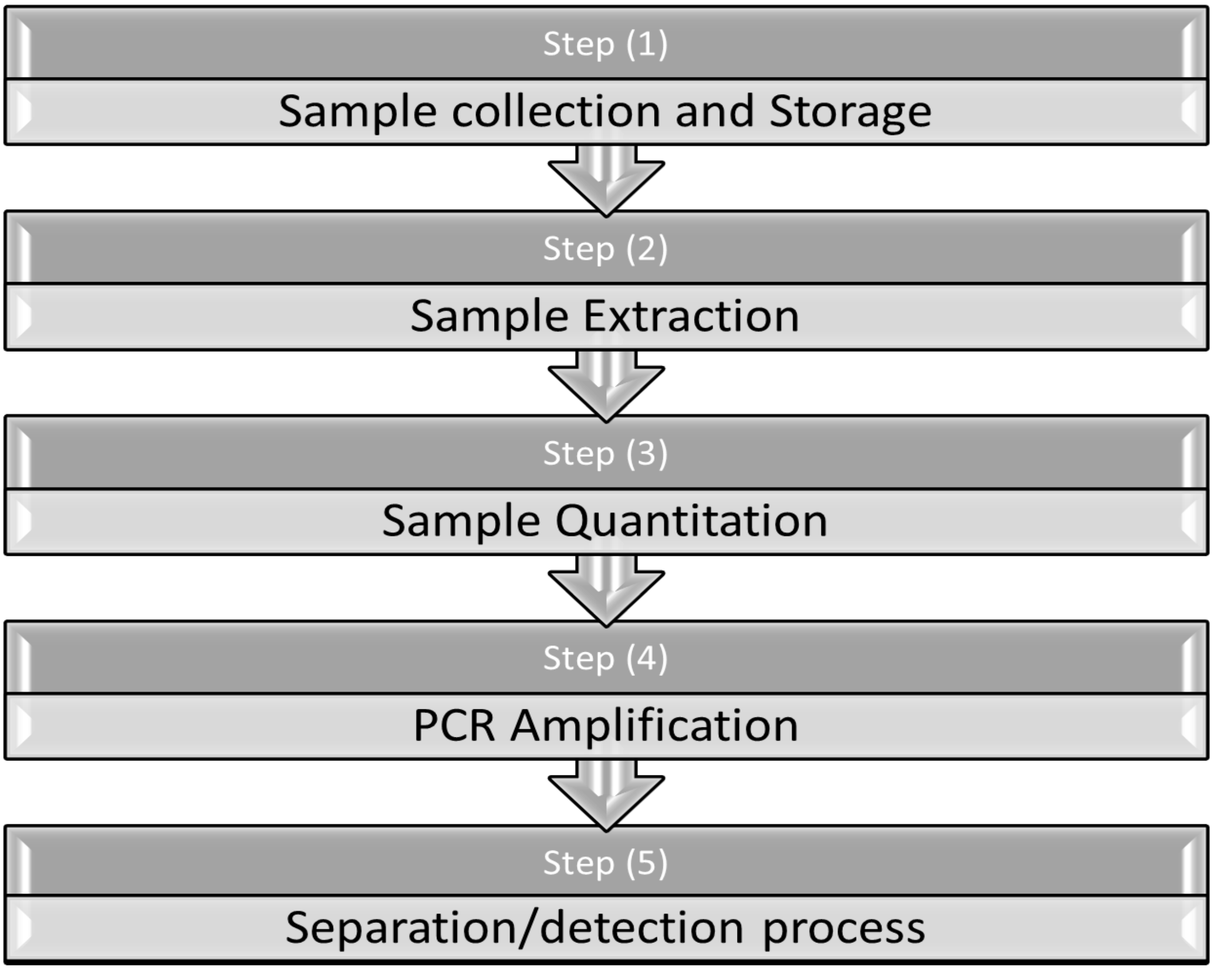


Figure 1.1: The overall block diagram shows the process to create STR DNA profile in form of electropherogram signal.

1.3.1 Sample collection and Storage

Collecting an appropriate sample is essential to the final DNA sample quality (see Figure 1.2) [1]. A sample can be collected from varies sources. A Buccal, or cheek swab, contains a large amount of viable DNA, while crime scene samples may contain degraded DNA due to humidity, temperature, or other environmental conditions.

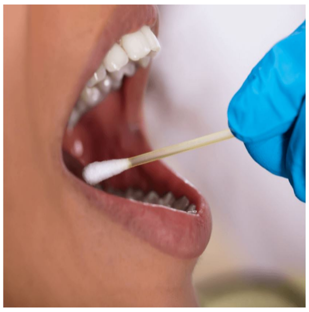


Figure 1.2: Collecting DNA sample using Buccal Swab method [1]

1.3.2 Sample Extraction

After the collection of cells containing DNA, the DNA must be extracted before further processing can occur. The extraction method separates and isolates the DNA from other parts of the cell, such as the other proteins, nucleus, etc. (see Figure 1.3) [2] It is also during this process when the different loci and short tandem repeat (STR) regions of the genomic DNA, which are critical to forensic DNA analysis, are isolated

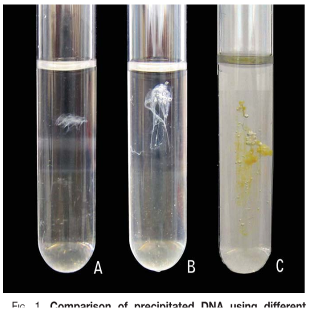


Figure 1.3: DNA extraction method through laboratories process [2]

1.3.3 Sample Quantitation

After the DNA is separated from the other cell material, the amount of DNA must be measured before it is amplified (see Figure 1.4) [3]. This process, known as quantitation, determines how much DNA is present in the starting sample so that it is not over or under-amplified in the following amplification process.

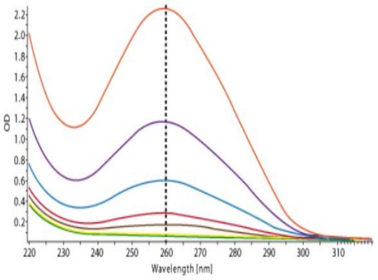


Figure 1.4: Wavelength method is used to determine DNA quantitation [3]

1.3.4 PCR Amplification

The starting amount of DNA in a given sample, even a buccal swab, is often very small. Therefore, a process known as the Polymerase Cycle Reaction (PCR) is used to make exact copies of the various STR regions isolated in the previous step. The process involves thermal cycles (i.e., heating and cooling) that 'unzip' the DNA double helix and create an exact replica using complementary nucleic acids. The number of cycles is determined by the DNA quantity available in the sample. It is also during this process where STR fragments are labeled with fluorescence-emitting molecules called fluorophores which will allow the STRs of varying sizes to be detected after separation. (see Figure 1.5) [4]

Figure 1.5: Shows Bio-Rad real-time PCR detection systems with multiplex PCR capability [4]

1.3.5 Separation/detection process

After amplification, the DNA is injected into a microscale capillary tube containing a sieving polymer.(see Figure 1.6) [5] A high voltage potential is applied to the capillary from one end to the other. Because the fragments of DNA have a polarity, the voltage induces their travel through the capillary, with the sieving polymer causing fragments of different size to travel at different speeds. DNA fragments of the same size, such as those for a specific allele, group together as they travel. This process allows the different loci and alleles to be spatially separated along the length of the channel, making it easier to detect the different colored fluorescent labels attached to these DNA fragments in the next step.

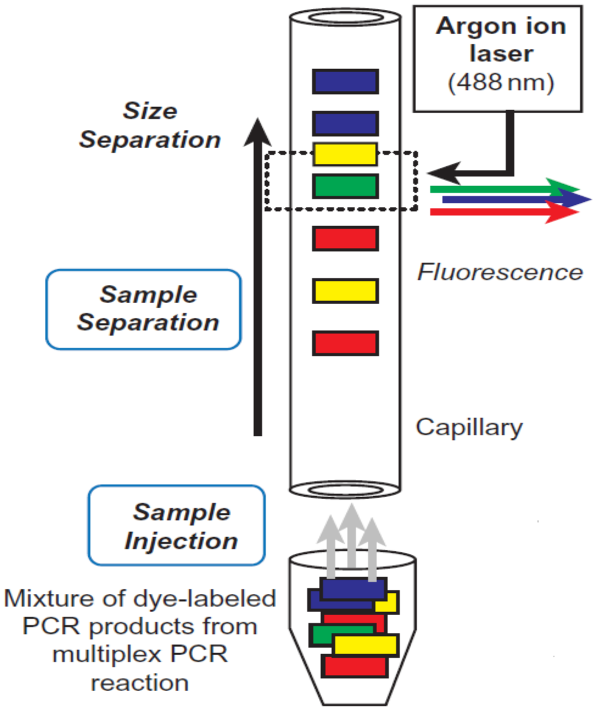


Figure 1.6: Schematic demonstration of the separation and detection of STR alleles [5]

A photodetector is located at the end of the capillary. As the separated groups of DNA pass near the detector, they are illuminated by a laser, causing the fluorophores to emit photons. This emission intensity, proportional to the amount of DNA of a given size present, is measured, causing an allele peak in the resulting electropherogram (y-axis is relative fluorescence intensity).

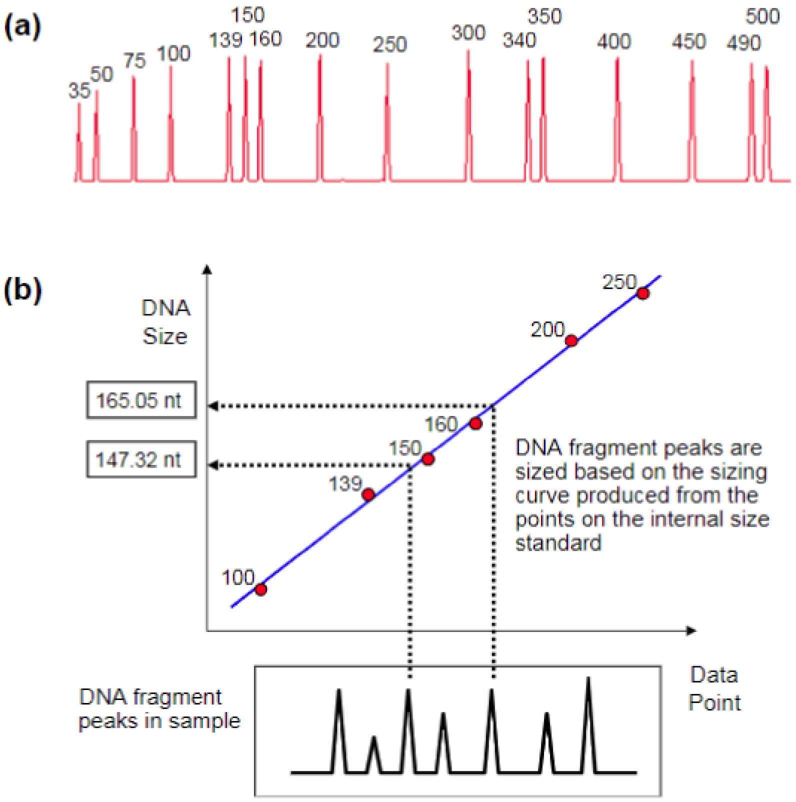


Figure 1.7: Peak sizing with DNA fragment analysis. (a) An internal size standard is analyzed along with the DNA sample and used to calibrate the peak data points to their DNA size (b) This standard is labeled with a different color fluorescent dye so that it can be spectrally distinguished from the STR alleles [6]

The electropherogram has heterozygous or homozygous allele peaks for each DNA loci of interest. The time it takes for each allele packet to travel through the capillary and create a peak in the electropherogram when measured at the detector is proportional to the size of the STR fragments in the packet.(See Figre 1.7) [6] This collection of peaks scattered in time along the x-axis is compared to a sizing standard that is simultaneously passed through the capillary with the sample to correlate time with allele size (x-axis is base pair size, which is proportional to time). In addition, the electropherogram also has an ‘allelic ladder’ standard that ensures that the allele peak positions correlate with those expressed in the human population.(see Figure 1.8) [7]

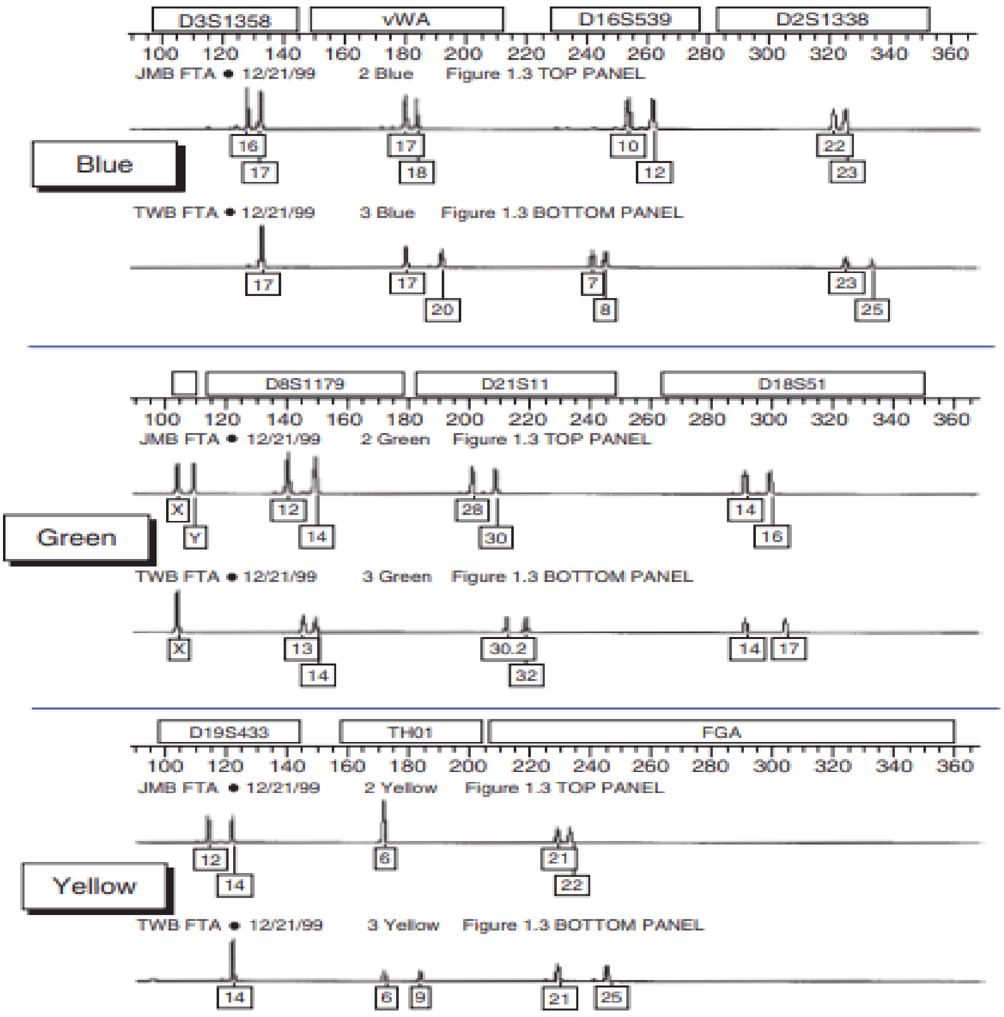


Figure 1.8: Genotype results on the two samples obtained with AmpFlSTR SGM Plus STR kit amplification and Genotyper 2.5 analysis. [7]

1.4 Issues associated with DNA profiling

1.4.1 Sample degradation

Degraded DNA comes from an alleled source of DNA such as skin, hair, blood, etc. After DNA becomes degraded when its molecules are randomly broken into smaller pieces, due to environmental exposure. The RFLP technique used high molecular weight DNA molecules need to be present in a sample to detect large VNTR (variable number of tandem repeats) alleles (eg., 20 000 bp). To evaluate the quality of a DNA sample, an ethidium-bromide stained agarose “yield-gel”. In some cases, a ”yield gel” or ethidium-bromide-stained agarose can be used to evaluate the quality of a DNA sample. Relative to an appropriate close molecular mass marker, usually high-quality genomic, high molecular weight DNA runs as a relatively tight band of approximately 20,000bp. Whereas degraded DNA looks like a smear of much less in size than 20,000bp. Figure (1.10a) [9]

Commonly high molecular weight, high-quality genomic DNA runs as a comparatively tight band of approximately 20 000 bp base pair relative to an appropriate molecular weight marker. A degraded DNA sample will appear as a smear of DNA that is much less than 20 000 bp in size. Contemporary PCR methods like multiplex STR typing are great, because small amounts of DNA can be amplified to a level at which can be detected. It’s now possible to analyze less than 1 ng of DNA using multiplex PCR amplification of STR alleles, in contrast with 100ng or more that might have been needed using RFLP several years ago. The more degraded a DNA sample becomes, the more breaks happen in the template, and less and fewer DNA molecules contain the full length that is needed in PCR amplification. [33]

Degraded DNA samples have better results when smaller short tandem repeat (STR) alleles are used. With restriction fragment length polymorphism (RFLP) and older technologies, DNA samples that were severely degraded would have been almost impossible to analyze. A high molecular weight or relatively high molecular mass, DNA molecules needed to be intact to detect large strands of (variable number of tandem repeat) VNTR alleles that contain a whole locus for one of the unique loci required (Figure 1.9) [8]

A few short years ago, 100ng or more of DNA was required with RFLP to be analyzed, but now less than 1ng of DNA is sufficient with the multiplex PCR amplification of STR alleles. The newer modern-day methods like multiplex STR typing, and PCR methods, are proving

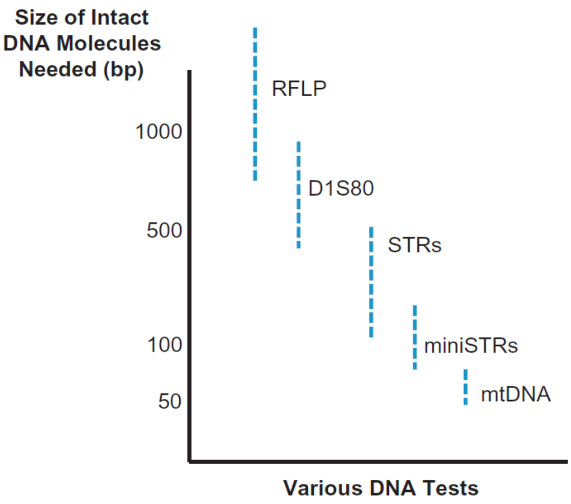


Figure 1.9: Clarification of DNA fragment sizes for different DNA tests [8]

to be powerful because they only need a minuscule amount of DNA for amplification to the level where they can be detected. This brings a new challenge yet of avoiding contamination of the samples by the crime scene technician or police. Officers collecting the biological evidence must be caution because of tests' sensitivity to a low copy of DNA. For amplification to occur, the DNA template must be intact between the primers and where the two primers bind so that PCR amplification may occur. Primer extension will halt at the break in a template, and PCR will not be successful if the DNA strand surrounding the STR region and serving as a template strand is not intact. As a DNA sample becomes more degraded, more breaks will occur in the template, and less DNA molecules contain the entire length of loci needed for the PCR amplification. [27]

When STR loci can be amplified, there is a greater chance of the STR primers finding some intact DNA strands for amplification. Also, since both alleles in a heterozygous individual are similar in size, it is less likely to have alleles drop out during preferential amplification of the smaller allele, therefore making the narrow size range of STR alleles is beneficial to the analysis of degraded DNA samples.

The results of several experiments show that there is an inverse relationship in degraded DNA samples, successful PCR amplification, and the size of the locus, for samples being obtained from a mass disaster or a crime scene [34][35][36] STR loci with larger- sized amplicons are the first to drop out of a DNA profile when significantly degraded DNA samples are amplified in multiplex amplifications, loci such as CSF1PO and Penta D or FGA and

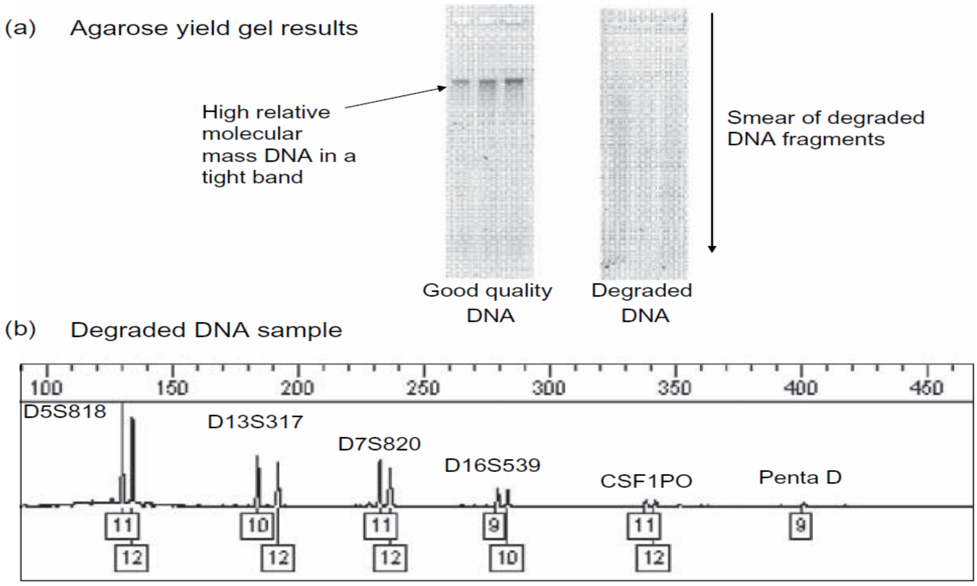


Figure 1.10: Degraded DNA impact results [9]

D18S51. (Figure 1.10b) [9]

During one of the first studies of degraded DNA samples, the Forensic Science Service demonstrated the value of multiplex STR analysis, they obtained and successfully typed a majority of 73 duplicate pathological samples from the Waco Branch Davidian fire with four STR markers [34]. On all the samples where alleles were scored, the observation showed no allele dropout and obtained concordant results. The statement also showed a correlation between successful typing at a locus and the average length of the alleles at that locus. The VWA locus containing alleles ranging from 130bp to 169bp had 115 successful amplification, while the FES/FPS locus containing alleles in the size range of 212bp to 240bp only showed 91 successful amplifications. As part of the Waco identification program, 24 examples were examined, and amelogenin amplicons (106bp or 112bp) were obtained on all of them. Failing first were loci with larger alleles [34].

Multiplex STR systems are far superior for analysis of degraded DNA samples over the DNA markers that were previously used. STRs are less prone to having dropout alleles than VNTR systems (AmpFLPs) such as D1S80 and are more sensitive than single-locus probe RFLP methods, and more discriminating than some other PCR-based typing methods. Ones such as AmpliType PolyMarker and HLA-DQA1 lead to STR profiles, either due to inhibition or degradation. Figure (1.11) [10]

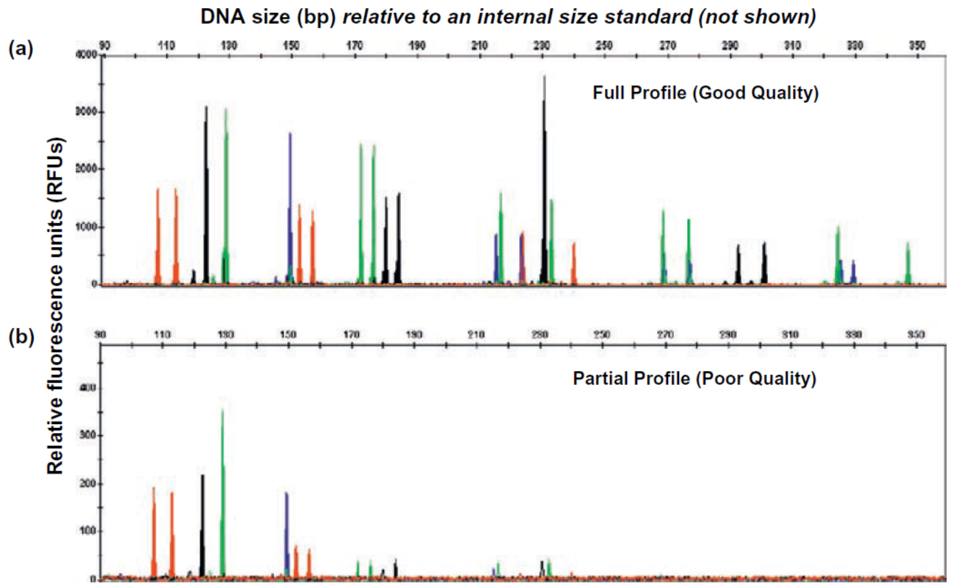


Figure 1.11: DNA profiles from different qualities, but same biological sources, in comparison [10]

1.4.2 Mixtures

Mixtures arise when there is DNA from two or more individuals present in a sample. Fluorescent measurements paired with PCR sensitivity have advanced technologies more sensitive. This has dramatically improved the ability to see minor components in the DNA profile of mixed samples.[29] the presence of a mixture can now be discovered when there are notable differences in allele intensities or when three or more alleles are observed at multiple loci in a short tandem repeat (STR) profile. [37]

When two or more individuals are present in a one sample, it is called a mixture. These “mixtures” can be challenging to interpret and clarify without substantial experience and extensive training. Compared to what was accessible with RFLP methods just a few years ago, the potential to detect small components in the DNA profile of mixed samples has advanced dramatically. Furthermore, the unproven side of statistical calculations for understanding mixture has been studied more thoroughly [38]

There can be several indications when deciding if a mixture is present. For example, the loci will display more than two peaks in the expected allele size range. Also, there will be a peak height imbalance between heterozygous alleles at a locus. Another indication is that the stutter will appear abnormally high (e.g., 15-20 percent). The Forensic Science Service has extensively examined mixture interpretation [39][40]. When three or more prominent peaks are present at one or more loci, the mixture can usually first be identified. Due to the possible genotype combinations, a sample with DNA from two different sources at a single locus can display one, two, three, or even four peaks.

If a mixed stain with two donors shares one or more alleles, the alleles may be "masked", therefore making it hard to decipher the contributing genotypes. For instance, two people with genotypes 23,24 and 24,24 at the FGA locus, by taking the ratio 1:1, will give us a percentage of 1:3 for the 23:24 peak sections. A large peak can be found considering stutter products and no other data with this specific occurrence. Although, this sample could be analyzed accordingly into its components by studying the STR profiles at other loci with unshared alleles, i.e., three or four peaks per locus. In a simulated mixture analysis by the Forensic Science Service, including 120,000 individual STR profiles from their Caucasian database, scientists attempted to see if masking would occur at every locus in a multiplex [39]. It was discovered that a more significant amount of the artificial mixtures displayed 15-22 peaks throughout a six-plex STR marker multiplex. In a mix of two heterozygous people without overlapping alleles at six STRs, the most significant amount would be 24 peaks. Consequently, in this instance with random people, simple mixtures are easily identified by the appearance of three or more alleles at various loci. Only four cases have been noted, with one or two alleles present at each locus in the six-plex, in over 212,000 pairwise comparisons. They could be designated mixtures due to peak imbalances [39]. The following steps are shown in Figure 1.12 [11] and have been taken and applied in an example to show how to interpret a mixture. In Figure 1.13 [12], two types of DNA could be seen, one female and the other male, typically seen in sexual assault investigations. The STR markers for the mixture have been separated into three panels based on their dye label to make each STR locus easier to visualize. Showing is a presence of more than two peaks at most of the loci, as D3S1358 contains four peaks, and VWA has three. Also shown here is the imbalance of the X and Y alleles of the amelogenin sex-typing marker.

It was not likely that there were more than two contributors to the mixture, as there were not more than four peaks at any one locus. Using universal designations to track possible allele combinations, the called alleles had been labeled, with letters 'A', 'B,' 'C,' and 'D.'

By studying the loci with four peaks, a ratio of individuals contributing to the mixture can be estimated. Figure 1.14 [13] shows labeled peak areas from the green panel STR data. As observed at both D18S51 and D21S11, four peaks are present. It can be assumed that

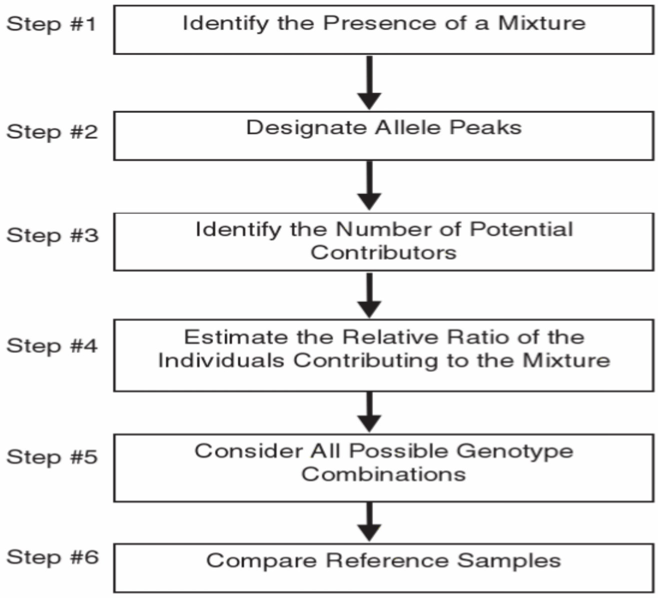

Step #1
Identify the Presence of a Mixture
Step #2
Designate Allele Peaks
Step #3
Identify the Number of Potential Contributors
Step #4
Estimate the Relative Ratio of the Individuals Contributing to the Mixture
Step #5
Consider All Possible Genotype Combinations
Step #6
Compare Reference Samples

Figure 1.12: Interpretation of Mixture [11]

AD and BC are the best possible combination of alleles to explain the data because A and D alleles and B and C alleles are similar in the peak areas in D21S11. Just like A and C alleles have similar peak areas in D18S51 and can be grouped, showing the best possible combination of alleles at this locus is BD and AC. Then by dividing the amount of larger alleles (B and C) by the amount of smaller alleles (A and D), you will get a mixture ratio of about 2:1 for D21S11.

Therefore, the most significant contributor has about twice as much DNA as the smaller contributor in the mixture. Applying the same equation, the mixture ratio for D18S51 is about 2:1 for the most significant contributor. Due to the imbalances in heterozygote peak areas and stutter products. Calculating these ratios at loci with four noticeable alleles is easier than with one or more shared alleles. With D8S1179, there are three visible peaks, and at least one of the peaks stands for an allele from both the minor and major contributors. Every possible mix of alleles has to be thoroughly examined to decide which one best fits the data.

Shown in Figure 1.14 [13] are the presumed peak patterns for each possible mixture combination of combination 2:1, including three peaks.

Figure 1.15 [14] gives data for D8S1179, showing AC and BC allele combinations, presenting BC as the major contributor. Hence, allele C (or 14) is shared in the case in D8S1179, the major contributors genotype is 12,14, and the minor is 10,14. In this mixture, the male

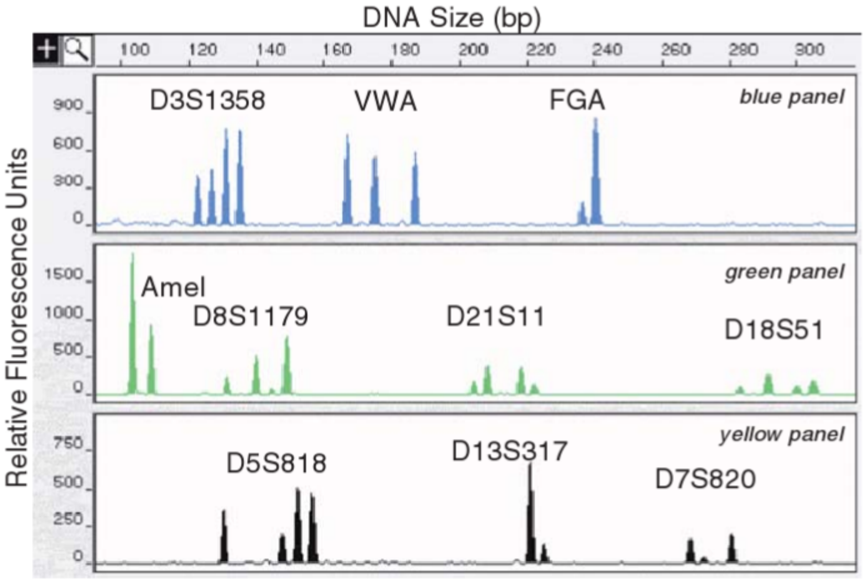


Figure 1.13: Typical result in a forensic case involving mixed DNA (male/female) samples [12]

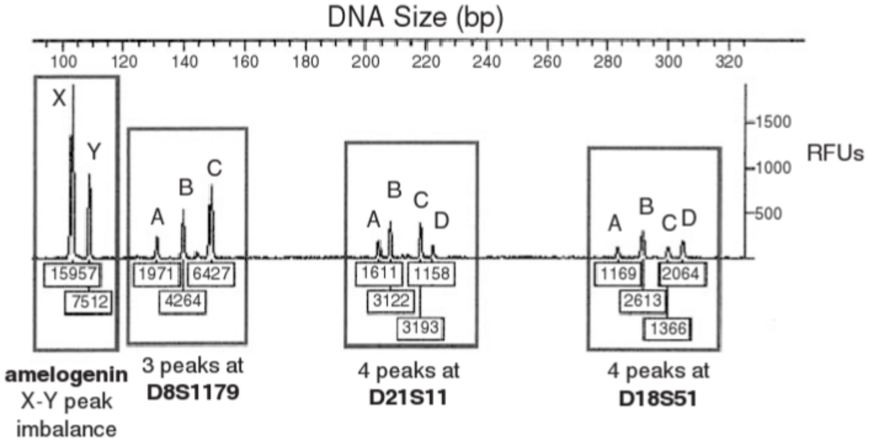


Figure 1.14: Green panel data, peak areas of example mixture [13]

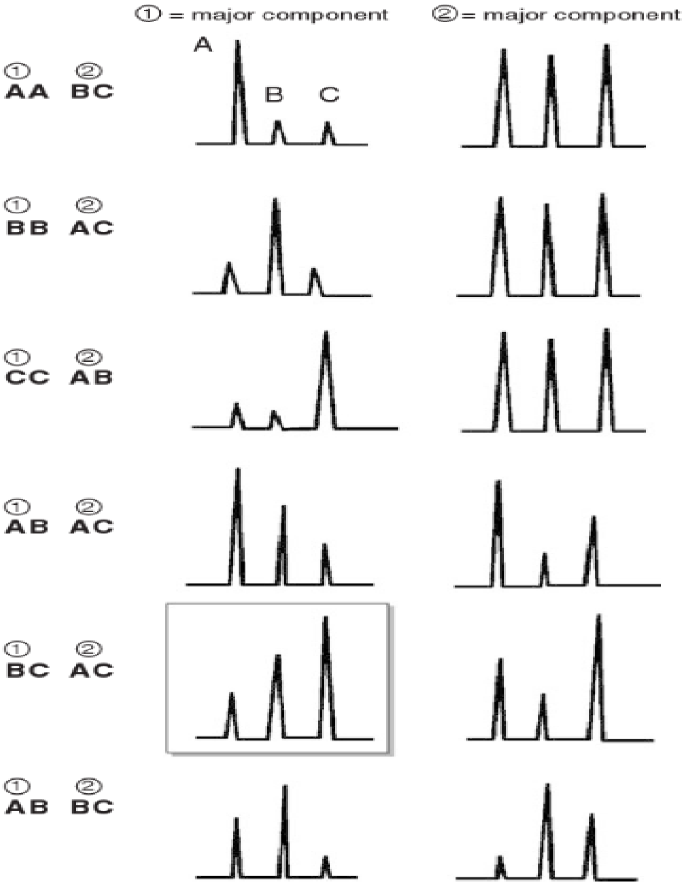


Figure 1.15: possible peak profiles for mixture combination 2:1, including three peaks [14]

DNA was the major contributor, at two times that of the female. Therefore, it was possible to see that the major contributor was the male and that alleles 14 and 17 in D18S51, 30 and 32.2 in D21S11, and 12 and 14 in D8S1179 belonged to him. The genotype profile of the major and minor contributors can be distinguished by processing all the loci in the way it has been shown above.

- **Conclusion of mixtures interpretation:**

STR typing procedures have shown to be an excellent way to differentiate components of mixed samples; these types are often seen in many forensic investigations. Although evidence sometimes contain multiple stains, not all of them are mixtures. Also, if there are various samples for testing, the easiest ones to decipher are the ones that should be tested first [41]. As suggested by Petr Gill, from the Forensic Science Service, a mixture should not be interpreted unless necessary. His lab [42] studied all mixture STR profiles over four years.

Only 6.7 % or 163 mixed profiles were solved of 2424 total samples obtained from 1547 criminal cases. Because it is easier to either include or exclude a suspect's DNA profile from a crime scene mixture profile, some labs don't even decipher the genotypes possibilities. A suspect cannot be removed as a contributor to a crime scene stain if all the alleles from that suspect's DNA are present in the crime scene mixture, just as the alleles of a victim's DNA profile can be eliminated from the mixture profile to simplify it, thus making it easier to decipher the perpetrator's DNA profile.

1.4.3 Low-copy-number DNA analysis

Using only a few cells from a biological sample to try and generate a reliable STR profile is like looking for an object in the mud or deciphering the image in a fuzzy photograph. Some recovered DNA profiles are innocently left before the crime occurred and may not be associated with the crime itself [39]. At least three artifacts typically arise when LCN (low-copy number) testing is performed, (1) where an allele fails to amplify due to stochastic effects, commonly known as allele ”dropout”, (2) allele ”drop-in” when additional alleles are often observed from sporadic contamination, and (3) enhanced stutter product (non-allelic data) amounts that are often higher than the typical 5-10 percent of the nominal allele [34]. When one of the alleles is amplified by chance in the early rounds of PCR in a preferential fashion, heterozygote peak imbalance is typically exacerbated due to stochastic PCR amplification. Allele dropout could be an extreme form of heterozygote peak imbalance.

LCN- which stands for low copy number and is probably one of the most commonly used terms, reflects that small amounts of DNA are examined. Recently some labs have begun to refer to LCN as low template DNA (LT-DNA) since some collected samples contained as little DNA as a single [43]. STR typing results have been demonstrated, and the (PCR) polymerase chain reaction is very sensitive. Positive attempts have been made to recover DNA profiles from touch evidence, encouraged by this capability. However, unless appropriate measures are taken to demonstrate the reproducibility of allele calls, this type of low copy DNA analysis can sometimes question what constitutes reliable results. Efforts have been made to strengthen reliability with low copy DNA testing, and an approach to improve DNA sensitivity are addressed here.

To obtain better results from limited biological evidence, a valuable asset is usually an improved sensitivity in the detection technique. Laboratories have applied what some term ”enhanced interrogation techniques” because of the failure of getting results with low amounts of DNA template [44]. Yet DNA testing has been successful while being applied on a single-cell level (D.N.A. Box 11.3). At the same time, it shows various strategies for improving sensitivity with low copy DNA samples (shown in Table 1.1) lists some of the advantages and disadvantages. Increasing detection sensitivity is just like turning up a radio’s volume, so it can better hear from a distance. Although, as you turn the volume up, it could distort the quality of the sound. When the sensitivity of PCR is heightened, contamination of low amounts of DNA from outside sources and the chance of gaining cells from secondary transfer typify noise that can obstruct detection of the actual signal.

Strategy	Advantage	Limitation
Increased Number of PCR cycle	Create More PCR Product	Allele drop-in possible
Post-PCR sample	Improves injection of PCR product into CE capillary	Extra expense to sample processing stochastic threshold needs to be raised to avoid homozygote designations
Increased CE injection	Improves amount of PCR product injected into CE capillary	Stochastic threshold needs to be raised to avoid false homozygote designations
Reduced volume PCR	Concentrates PCR product relatives to amount subjected to CE analysis	PCR inhibitors may be concentrated causing amplification failure; pipetting precision can be more challenging
Nested PCR	Creates more PCR product	Prone to contamination because tubes are opened to add a second round of primers and reagents
PCR enhancements (primer positions, polymerase concentration, etc.)	Creates more PCR product	Stochastic threshold needs to be raised to avoid false homozygote designations
mtDNA	Higher starting copy number per cell	Lower power of discrimination; cannot resolve individuals from same maternal lineage.

Table 1.1: Pros and cons of different strategies to solve low-copy DNA

1.5 Methods to overcome issues with DNA profiling

1.5.1 Basics of electropherogram creation

It is required that DNA testing is performed in a laboratory with equipment and dedicated facilities that meet the FBI's stringent QAS (Quality Assurance Standards) requirements. DNA is often tested at the crime scene before being analyzed in a laboratory to determine the type of biological material in question. Generally, samples are taken directly from a victim or suspect and then compared to samples collected from a crime scene. These samples are sent to a laboratory and undergo testing to determine who deposited biological material at a crime scene. This process starts with an extraction, obtains DNA from the cell, then moves on to quantitation, determining how much DNA is available. The next step is amplification, which is when multiple copies of that DNA are produced to be characterized, followed by separation, which is done to permit subsequent identification. An analysis and interpretation process compares the DNA samples qualitatively and quantitatively to a known profiles. The final step is quality assurance; this is done by reviewing analyst reports to assure technical accuracy. This process ultimately provides the analyst with a chart called an electropherogram, which gives a display of each locus tested from the genetic material present see Figure 1.16 [45] [15]

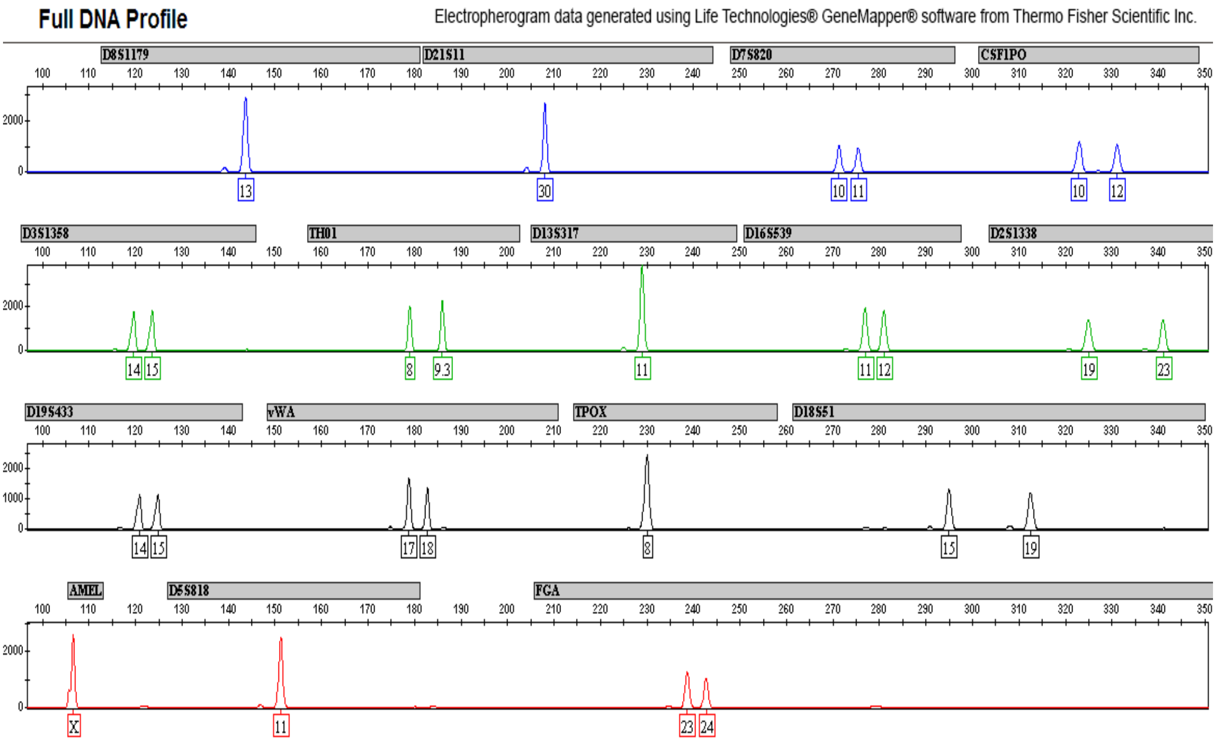


Figure 1.16: Complete profile for a random individual [15]

The number of alleles in each locus is what determines whether the person is homozygous (a pair of matching alleles) or heterozygous See Figure 1.17. [16]

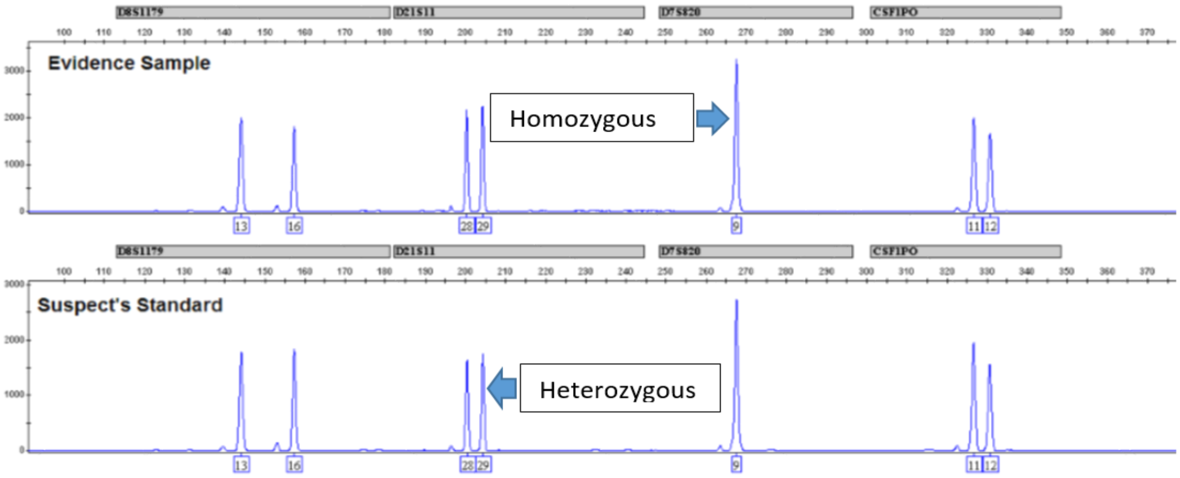


Figure 1.17: The difference between homozygous and heterozygous loci [16]

DNA molecules become degraded by randomly breaking into smaller pieces, this can happen in several ways. Environmental exposure is a significant degrading factor; other factors include exposure to water or enzymes called nucleases destroy part of the DNA.see Figure 1.18 [17].

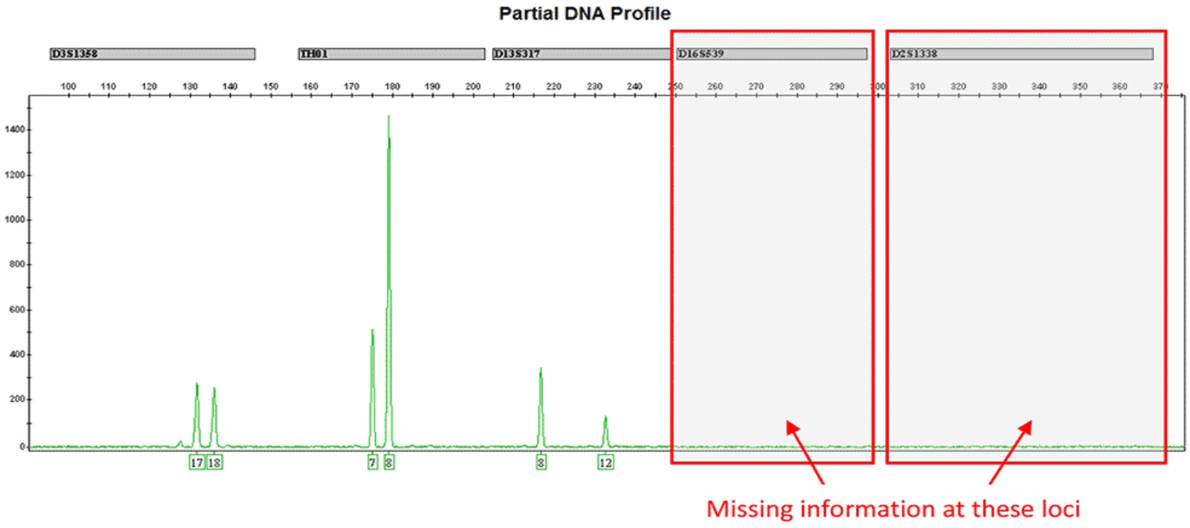


Figure 1.18: shows a green dye for individual missing information at two of its loci [17]

When any of those situations occur, the alleles are missing from the loci, which leaves us with a partial DNA profile, also known as degraded DNA [33]

1.5.2 Biology approaches

It has been a significant challenge to integrate processing DNA to obtain a ”swab- to-profile” result without user intervention. In August 2010, a report on one such integrated device from the University of Arizona Center for Applied Nano-Bioscience and Medicine and the Forensic Science Service was published [46]. A DNA cartridge with wax seals delivered reagents and samples to the necessary reaction chambers to allow PCR amplification, DNA purification, and a collection of the amplified product, resulting in less than four hours. This type of DNA separation is performed by connecting a Teflon tubing to an accompanying CE chip [47].

- **Reduction of PCR product size (mini STRs)**

Comparing established sequences for the same loci that generated longer amplicons, high and low amounts of DNA were successively typed using some newly redesigned PCR primers nearest to the STR repeat[32].

Stated in an article entitled “Less is more- length reduction of STR amplicons using redesigned primers”[32]. If the established sequences are compared to that generated longer amplicons for the same loci, very low amounts of DNA and highly degraded DNA were successfully typed, using some newly redesigned PCR primers close to the STR repeat.[32]

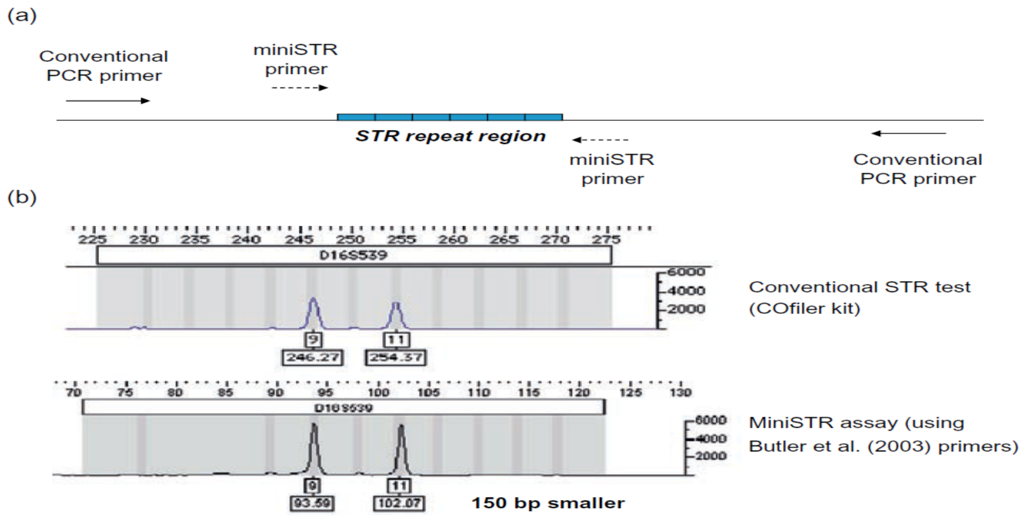


Figure 1.19: MiniSTR primers, being closer to the STR repeat region, compared with conventional PCR primers [18]

STR loci can extend in size past 400bp in commercially available kits. Most of this length comes from the flanking sequences surrounding the STR repeat of interest. To make it fit into the desired size range for a multiplex assay, PCR primers are taken away from the repeat region that imparts variability to the locus [48]

For example, of the core AAAGA, repeat the two PCR primers used for the PowerPlex 16 locus Penta D anneal 71bp upstream and 247bp downstream. Amplifying these PCR primers gives us alleles ranging from 2.2 to 17 repeats and generates amplicons in a size range of 376bp to 449bp [35]. For alleles 2.2bp to 17bp, overall PCR product sizes are reduced from 282bp to a range of 94bp to 167bp, when primers are brought to within 11bp upstream and 247bp downstream of the repeat region [49]. (Figure 1.19) [18] miniSTR in

STR's" or the size reduction principle when making smaller STR amplicons. Some loci can be smaller in size than others. Since the size aspect has been taken away, this creates several disadvantages, meaning that only a few loci can be amplified simultaneously.

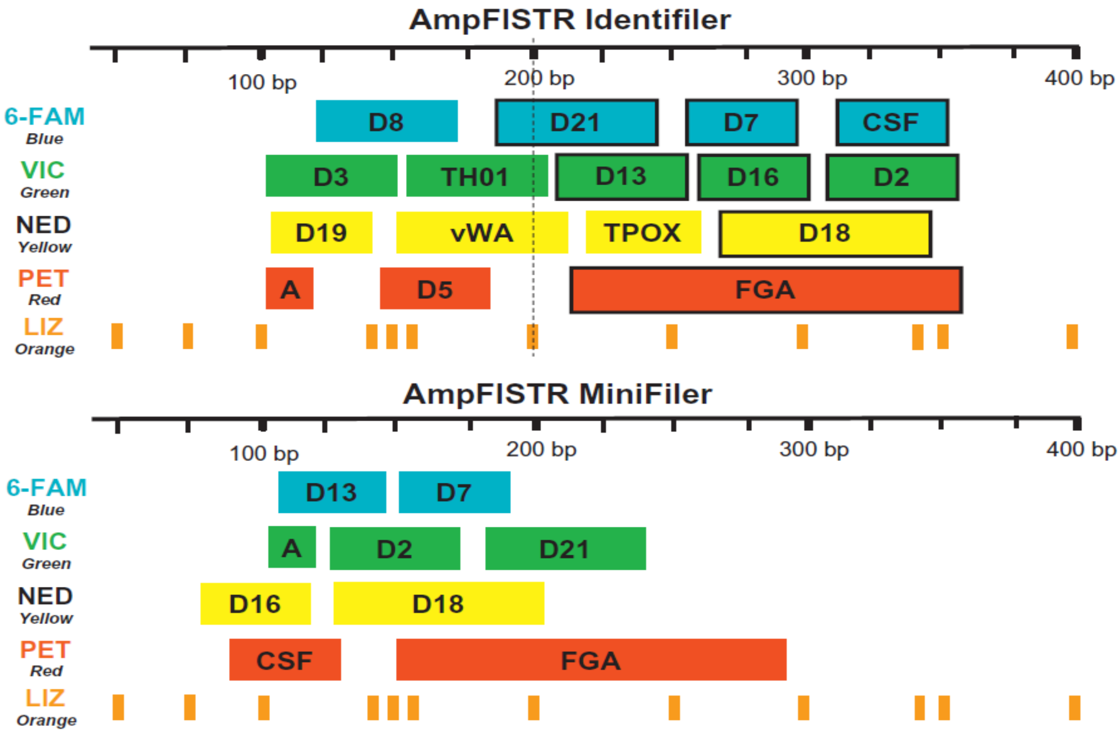


Figure 1.20: is depicting relative dye and size labels of PCR products generated with MiniFiler STR and Identifiler kits [19]

Primers are shifted away from the repeat region to create larger PCR products. Four or more loci are packed into a single dye color in large multiplex assays such as *GlobalFiler*^{TM} (PCR Amplification kit). Because all of the loci have almost the same size range of 100bp, "typically only having one locus per dye color," the "mini-plexes" created for amplifying miniSTRs have primers that are as close as possible to the repeat region [49].

Applied Biosystems managed to put eight miniSTRs and amelogenin into their single amplification *MiniFiler*^{TM} kit by using mobility modifiers to adjust the electrophoretically ob-served PCR product sizes [50] (Figure 1.20) [19] To verify that allele dropout from primer binding site mutations is rare or non-existent, concordance studies must be performed because different PCR primers are used with miniSTRs compared to conventional STR megaplexes [51]. This is done by examining the genotyping results and comparing them to see if they are the same between the primers sets [27]. Sometimes, a deletion, insertion, and even a point mutation can occur outside a miniSTR primer binding site in the flanking region. This can lead to undetectable and problematic differences in a heterozygous allele call [49][52][51]. It is likely that miniSTRs will be used in the future analysis of degraded DNA, no matter the disadvantages. It can help recover information on larger loci that have been lost using conventional megaplex amplification.

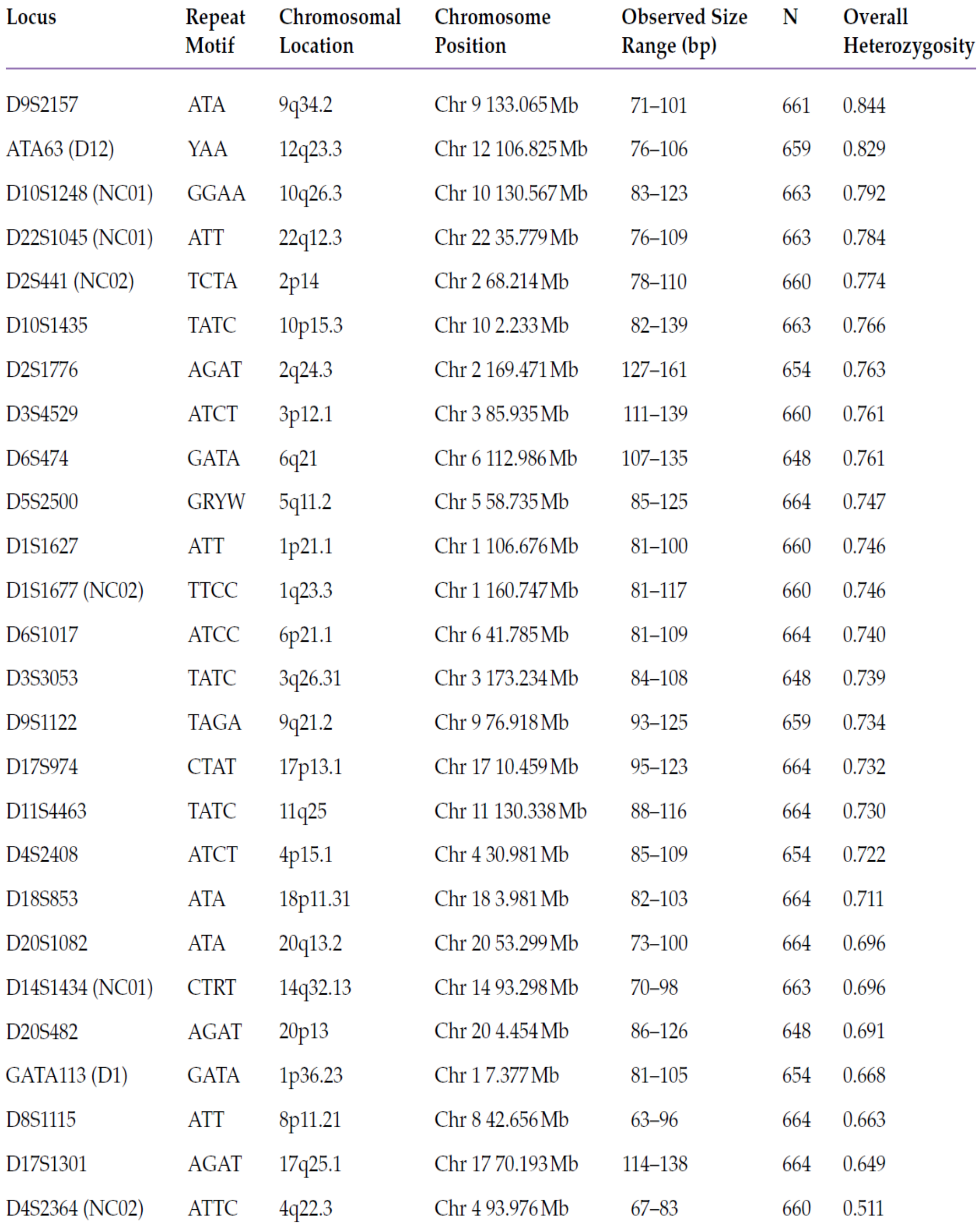
MiniSTR loci performed better than loci from a commercial STR kit, with enzyme-digested DNA [53]. Using only burned and damaged bone samples, some victim identifications from the World Trade Center were possible thanks to reduced size STR assays [54]. Reduced-size STR amplicons have been used to type a lot of DNA successfully, even telogen hair shafts, which contain very little nuclear DNA. [55][56][57].

The European DNA Profiling Group (EDNAP) published results of a study in 2006, con-taining degraded DNA samples where miniSTR primer sets were compared to conventional STR multiplex kits and an experimental single nucleotide polymorphism (SNP) assay [58].Leaders of the European community were led to advocate for miniSTR loci in future STR kits because all the miniSTR assays performed the best on degraded DNA samples [59].

- **New miniSTR Loci**

MiniSTR systems currently being used in forensic DNA typing and STR loci are being developed to focus on small size ranges and loci processing low copy samples. A battery of additional assays has been made available to assist forensic practitioners working with degraded DNA specimens.

MiniSTR systems that are currently used in forensic DNA typing and STR loci other than the CODIS markers have been developed with a focus on loci possessing few alleles and small size ranges [60][61][62]. A battery of additional assays have been made available to assist researchers and forensic practitioners who work with degraded DNA specimens. A set of 26 new miniSTR loci taken from over 900 candidate STR loci with multiple allele ranges and the capability to create PCR primers close to the flanking range were selected by scientists at the U.S. National Institute of Standards and Technology (NIST) [61][63]. While 26 of these loci were characterized in U.S. population samples (Figure 1.21), 25 of them were added to amelogenin and put in a single 26plex assay for typing reference samples [64].



Locus	Repeat Motif	Chromosomal Location	Chromosome Position	Observed Size Range (bp)	N	Overall Heterozygosity
D9S2157	ATA	9q34.2	Chr 9 133.065 Mb	71–101	661	0.844
ATA63 (D12)	YAA	12q23.3	Chr 12 106.825 Mb	76–106	659	0.829
D10S1248 (NC01)	GGAA	10q26.3	Chr 10 130.567 Mb	83–123	663	0.792
D22S1045 (NC01)	ATT	22q12.3	Chr 22 35.779 Mb	76–109	663	0.784
D2S441 (NC02)	TCTA	2p14	Chr 2 68.214 Mb	78–110	660	0.774
D10S1435	TATC	10p15.3	Chr 10 2.233 Mb	82–139	663	0.766
D2S1776	AGAT	2q24.3	Chr 2 169.471 Mb	127–161	654	0.763
D3S4529	ATCT	3p12.1	Chr 3 85.935 Mb	111–139	660	0.761
D6S474	GATA	6q21	Chr 6 112.986 Mb	107–135	648	0.761
D5S2500	GRYW	5q11.2	Chr 5 58.735 Mb	85–125	664	0.747
D1S1627	ATT	1p21.1	Chr 1 106.676 Mb	81–100	660	0.746
D1S1677 (NC02)	TTCC	1q23.3	Chr 1 160.747 Mb	81–117	660	0.746
D6S1017	ATCC	6p21.1	Chr 6 41.785 Mb	81–109	664	0.740
D3S3053	TATC	3q26.31	Chr 3 173.234 Mb	84–108	648	0.739
D9S1122	TAGA	9q21.2	Chr 9 76.918 Mb	93–125	659	0.734
D17S974	CTAT	17p13.1	Chr 17 10.459 Mb	95–123	664	0.732
D11S4463	TATC	11q25	Chr 11 130.338 Mb	88–116	664	0.730
D4S2408	ATCT	4p15.1	Chr 4 30.981 Mb	85–109	654	0.722
D18S853	ATA	18p11.31	Chr 18 3.981 Mb	82–103	664	0.711
D20S1082	ATA	20q13.2	Chr 20 53.299 Mb	73–100	664	0.696
D14S1434 (NC01)	CTRT	14q32.13	Chr 14 93.298 Mb	70–98	663	0.696
D20S482	AGAT	20p13	Chr 20 4.454 Mb	86–126	648	0.691
GATA113 (D1)	GATA	1p36.23	Chr 1 7.377 Mb	81–105	654	0.668
D8S1115	ATT	8p11.21	Chr 8 42.656 Mb	63–96	664	0.663
D17S1301	AGAT	17q25.1	Chr 17 70.193 Mb	114–138	664	0.649
D4S2364 (NC02)	ATTC	4q22.3	Chr 4 93.976 Mb	67–83	660	0.511

Figure 1.21: 26 miniSTRs Characterized at NIST

1.5.3 Neural network approaches

Recent studies have explored the possibility of using neural networks to evaluate DNA profiles. While some of them confirmed the possibility of replacing human validation with a specific type of neural network, others investigated the possibility of recovering damaged DNA samples. I have listed a few of these studies and efforts below.

- **Evaluation of STR typing by neural networks as a replacement for human reading**

Two forensic laboratory personnel typically interpret the STR capillary electrophoresis profile data independently, compare results, and resolve any discrepancies. Recent research has focused on developing a machine learning tool that classifies areas of fluorescence in raw capillary electrophoresis profile raw signal data in the same way as a human profile reader. For reading GlobalFiler^{TM} DNA profiles, FaSTR^{TM} DNA has integrated the ANN approach. A test was carried out at Forensic Science South Australia (FSSA) to determine if one of the human profile readers could be replaced by an ANN reader using the ANN feature of FaSTR^{TM} DNA. In reference profiles, FaSTR^{TM} DNA accuracy was 99.7% and was deemed high enough to be incorporated into the FSSA's reference reading workflow as a one-reader workflow.[65]

Validation work presented in this research shows the FaSTR^{TM} DNA software performs to a high standard and is suitable for supplementing or replacing existing forensic analysis software. As a result of this research, ANNs are a significant step forward for improving routine processes and applying lean thinking principles. This can be used in the majority of forensic laboratories. DNA profile analysis still has room for further innovation using machine learning algo-rithms. As an example, one could eliminate the use of ATs (and evaluate all scan points inde-pendently as a baseline, pull-up, stutter, or allele) for complete information to be extracted from an EPG. When using a semi-automated system that requires human decision-making, there is the potential to train the underlying ANN within the software, which needs further research. In such a scenario, the human reader would override decisions made by the ANN, allowing it to make better peak classifications and to learn from wrong decisions run by run, strengthening the algorithm over time. Several questions must be considered moving forward, including whether or when machine learning algorithms will completely replace human reading, as well as the risks, policy requirements, and modifications needed for existing accreditation obligations.[65]

- **Bayesian model for peak detection using (LC-MS) untargeted data**

The present study develops a Bayesian peak detection algorithm for liquid chromatography-mass spectrometry (LC-MS). With the probabilistic result, one can make a final determination about which points in a chromatogram are the result of chromatographic peaks. In contrast, some points are merely the result of noise. Probabilities contrast the traditional method, which relies on a threshold to determine a binary answer. The Bayesian peak detection presented here, on the other hand, allows the values of probability to be propagated into other preprocessing steps, potentially increasing the importance of chromatographic regions to the final results (or decreasing their significance). The present study uses the statistical overlap theory of component overlap from Davis and Giddings as a prior probability in the Bayesian formulation. [66]. It was demonstrated that the algorithm was successful in distinguishing chemical noise from actual peaks to be used with LC-MS Orbitrap data without any preprocessing.

Bayesian statistics is becoming a valuable tool in many areas of analytical chemistry. A case in point would be peak detection, where (unlike the conventional binary ”yes” or ”no” result applied by most existing algorithms), a Bayesian statistical model could offer a solution that would allow the end-user to incorporate the prior knowledge they already possess with the probabilistic outcome in order to reach an informed decision. As a result, threshold-based approaches delivering binary answers are likely to generate spurious results. Data preprocessing and filtering do not appear to be necessary with the current algorithm.

In addition, the method doesn't use any threshold. As a result, the data is just weighted with a probability score that can be propagated into other processing steps. In contrast, the data is discarded as part of the peak detection process. Using this algorithm, This method proved that the statistical overlap approach from Davis and Giddings is able to fully integrate the overlap theory, as well as being robust against errors in the parameters (noise in the signal and N/nc). Considering the computation time required to implement the proposed method for each chromatogram, the proposed method might not be feasible.

By using the approximations presented, the computation time can be reduced by a factor of 10, resulting in nearly the same results. Nevertheless, 10% error does not equate to 10% of features, but to 10% of possible data points. Feature detection does not significantly differ if a 10% loss of information is distributed across different regions in 2D space, because the user can make the final decision based on the cluster of probabilistic results. As a competitive feature detection algorithm, Bayesian 2D peak detection can be used because of its sensitivity to detect especially low-abundant peaks in raw LC-MS data, which are typically more biologically/chemically significant. [67]

1.5.4 Prior accomplishments by Dr. Dawson's research group

The endeavor and difficulty of finding the location of STR peaks were formulated as a problem of anticipating the shape of signals using differentiation. A peak-type electropherogram signal was provided in STR analysis data; in its first derivative, the location of the maximum could be computed as the zero-crossing points. Positive peaks are detected in the smoothed first derivative by looking for zero-crossing points. Discrimination is based on an adaptive amplitude threshold of the loci peak amplitude. This identifies the commercial STR electropherogram analysis software's inability to detect drop-out STR peaks in degraded DNA samples. The systems work on a fixed ”quality” threshold set by the forensic lab. The software does not call (i.e., measure) an allele peak below the threshold. A more accurate model of the noise in the input DNA signal has been developed by improving the quality of the signal before its peak detection. To allow reconstruction of the signal based on characteristics of a pristine DNA sample, the modeled noise from the degraded DNA sample has been subtracted to obtain a short tandem repeat (STR) analysis involving degraded DNA samples. Lastly, an automated scheme for quality enhancement and assessment for DNA signatures has been created. [68].

1.6 Motivation and Impact

Every day in the US, the number of unsolved crime cases increases. Many cases could easily be solved if the collected DNA samples could be processed. Unfortunately, many samples are not collected or stored correctly. Whatever the reason is, it leaves millions of degraded DNA samples sitting in evidence rooms, as well as millions of unsolved cases. With the developed system, some of these samples may be useful again. Those can be rerun through the database system and possibly connect a suspect with a crime committed years ago. On the other hand, it would be easier to access a device, pay a bill, or even send money worldwide with today's technology. And all of this can be done instantly by using biometrics techniques. It has been a decade of fast technological advances, and what used to take hours to verify can now be done in milliseconds. In the world of biometrics, a DNA sample could be one of the most potent ways of identifying a person. In today's world, you can change eye color and modify many identifiable traits in the human body, but one thing that cannot be altered is a person's DNA. If a database of DNA profiles for each human is established and linked to their personal information and government identification, there would be no way to misrepresent this person, which decreases the crime rate, improves society, eliminates identity theft, and makes it a problem of the past.

The goal of this thesis is to develop a way to artificially generate STR degraded and non degraded profiles based on 2017 FBI population genetics [69] using signal processing techniques developed into the graphic user interface.

The research goal can be accomplished by completing the following tasks:

Task A: Generation of artificial electropherograms based on population genetics

To have realistic and efficient results, a giant database is needed to run peaks finding developed method through. This will help to ensure that the conclusion represents the best available choice. Since it is hard to obtain real profiles from existing labs and there are few available through any other sources, the study uses an STR generator to find real STR peaks. The developed STR profile generator can produce many artificial profiles using the latest probability loci chance from the FBI database. This will give the used approach higher efficiency, with a realistic rate of 100 percent of obtained data ratio, while assisting us in the recovery of missing alleles. By running the profile generator, discovered are may apply developed approaches to recover partial profiles.

Task B: Match score for degraded profiles

The developed software can represent all the recovery methods and matching scores that the WVU engineering group has achieved to move the research methods and the obtained results to the next level. It is always helpful to know how much of the degraded profile checking. The used database assists us in finding the closest possible match by comparing the degraded profile to the NIST (National Institute of Standards and Technology) artificial database to see if there are any matches by the ratio in the whole database. This excludes a lot of profiles existing in the database, which enhances the conclusion. This is done best by generating match score matrices. Then compare each allele in the degraded profile to all profiles in the existing database, followed by scoring each locus by either zero or one, then adding all loci matching numbers together to obtain a profile score. Finally, by comparing them to each other, which include sorting them in descending order to find the closest match once we get "38", double the number of loci included in the 20 CODIS identifying DNA mapping. It can be considered a perfect match, meaning, The DNA profile sample we were looking for is found, which can be used once it develops more as one of the biometric matching techniques.

Task C: Apply peak detection and signal recovery enhancement

After enhancing the quality of the STR profile signal, A script has been written to detect the peaks in any signal. The GUI (Graphic user interface) can display the height and size for each allele in the DNA profile enhancing program, calculating the quality ratio and denoising for any signal—finally, adding and combining all those scripts into one GUI for a better user experience. The new system shows all degraded alleles in any STR profile signal by enhancing and denoising them. In the end, this gave a whole and complete method to recover the STR profile, which can be used to study and recover any degraded STR profile data.

This research contains two advanced tools that will vastly open the door to validate and assist all future biometrics research. First, the STR generator tool will help create an unlimited number of STR profiles which can help to expedite the validation process to any new signal processing methodology that needs to be proven. Creating and degrading an artificial profile with a different degradation level can lead to a more robust result-driven method with unlimited samples to test on, which can solve a challenging problem research.

Secondly, the enhancement tool can impact biometrics applications differently by taking a swab from an individual to create an STR profile. This individual who is initially trying to cross a border or seek asylum in a foreign country and comparing his DNA profile to an international database with worldwide criminal records, can in result, expedite and redirect an investigation to an individual in the right direction and will eliminate the possibility of acceptance in case of a possible match and provide validation for his records.

1.7 Overview of thesis

This section provides a short hierarchy of signal processing for degraded DNA as a biometric module its pros and cons. Also, the assessment of used algorithms and implementation of the various user interfaces are discussed. The rest of the research is ordered as follows:

- Chapter 2: indicates a summary of the basic terminology related to DNA typing, describing the biological laboratory processes which are used to recover a degraded DNA sample, in additional for an overview of used signal processing methodology.

- Chapter 3: describes the method behind developing STR profile generator based on FBI population statistics of 2017. It explains in details how we can form a data set of STR profiles starting from just alleles probabilities.

- Chapter 4: contains all the techniques that have been used to degrade the artificial STR profile. Also discussed are added noise characteristics with examples for some common values.

- Chapter 5: contains all the mathematical calculations and signal processing theories that have been used to analyze STR profiles and perform signal processing to degraded alleles to provide a partial or complete recovery for the tested profile.

- Chapter 6: summarizes the conclusions from all results and discuss possible future extensions to this work scoop

Chapter 2

Theory

2.1 Describing DNA typing

DNA sample processing has been developed throughout the years, evolving the ability of forensic science to match criminals with cases. The number of closed cases increases with criminals behind bars while more wrongfully accused are freed using the power of hidden forensic techniques.

A summary of the biological process is clarified in figure (2.1) [20]. After collecting all sample materials from a crime scene or individuals, DNA gets extracted from its biological sources and is then quantified to calculate the measurement of recovered DNA. Certain regions of captured DNA are multiplied with PCR (Polymerase Chain Reaction). By using commercial kits, concurrent PCR of 20+ short tandem repeat (STR) markers can be commonly enabled. STR alleles are illustrated in correspondence to PCR amplification, where it remains by measurements utilizing capillary electrophoresis and statistics scanning software. A statistical illustration evaluates the exception of the alleles from created DNA profiles, which might be a mixture or individual based on the data origin.

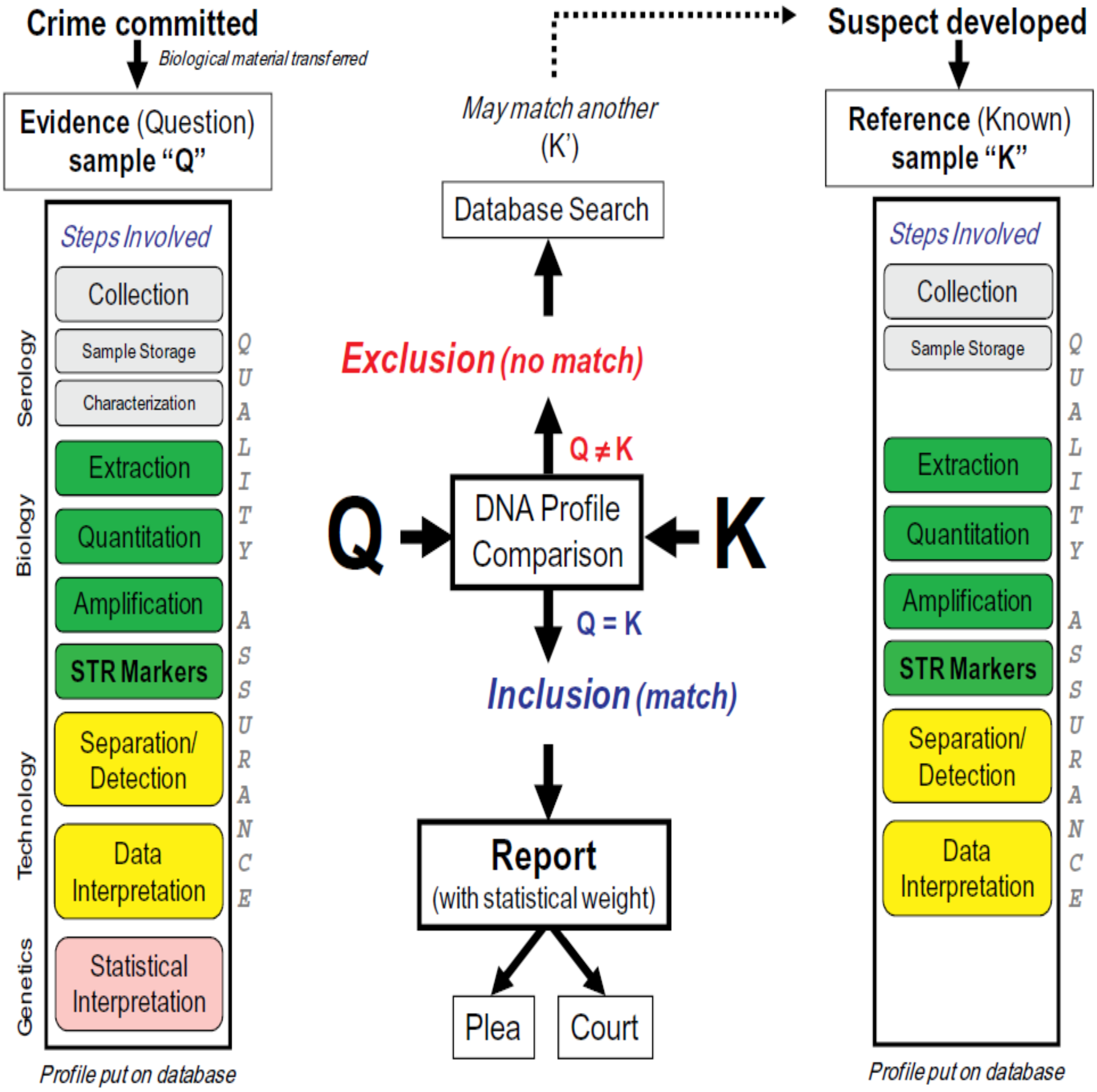


Figure 2.1: Overview of steps involved in DNA testing [20]

2.2 Rapid DNA analysis

Living in a quick, rapid era where any type of technology would be almost obsolete in five or ten years makes it hard to foretell the future of forensic DNA typing a decade from now. It results from the parallel efforts in all sciences, and our case: biotechnology developments tools for forensic DNA analysts. Forensic DNA typing techniques have been improved rapidly over the past two decades but settled down into two major concepts: (1) short tandem repeat (STR) typing (2) capillary electrophoresis detection. The innovations have changed almost every single aspect related to DNA profile typing. There's always a huge desire to achieve faster DNA analysis, greater detection accuracy, and more powerful discrimination capabilities. Below, we will review some ongoing efforts with expert systems for DNA profile analysis.

2.2.1 RapidHIT 200 Rapid DNA Profiling System

IntregenXTM developed RapidHITTM 200 Human DNA Identification System (See Figure 2.2) [21] . STR-based Human Identification (HID) is a first-of-its-kind fully automated system created by IntegenX. Microfluidic and STR-based chemistry are integrated for the success of the IntegenX system. Up to eight buccal swab samples are loaded into disposable cartridges of reagents, and then the system initiates sample processing without further user interaction. In about 90 minutes, the system extracts DNA from the samples, amplifies it, separates it by electrophoretic gradient, and analyzes it using software to generate full DNA profiles. Data for the U.S. market are stored in CODIS-compatible format within the embedded GeneMarker HID® software (SoftGenetics®, LLC). If necessary, additional data formats are available [70]. NDIS law enforcement booking stations will be able to use the RapidHIT$^{\mathrm{TM}}$ ID DNA Booking System v1.0 starting July 1, 2021 [71].

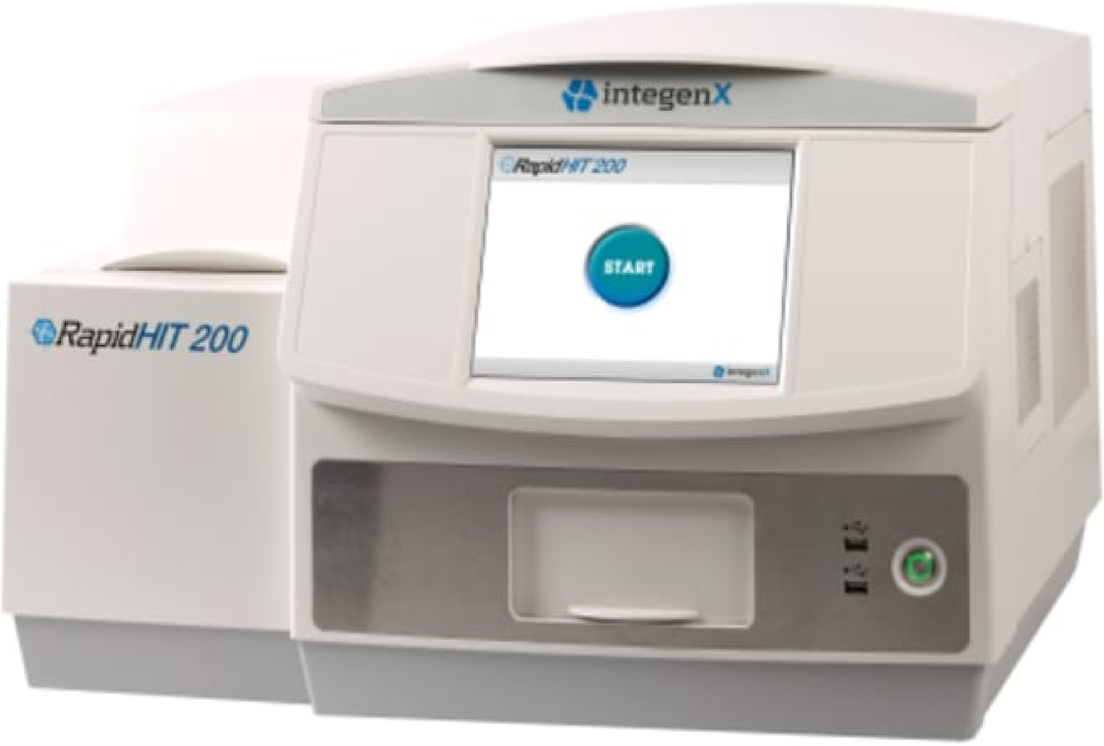


Figure 2.2: RapidHITTM 200 instrument by IntegenX [21]

2.2.2 ANDE Rapid DNA instrument

A CODIS Core Loci STR profile can be generated from a buccal swab using Rapid DNA by ANDE's completely automated (hands-free) process. Without human intervention, "swab in - profile out" involves automatic extraction, amplification, separation, detection, and allele calling.

ANDE received FBI approval for the National DNA Index System (NDIS) on June 4, 2018, DNAscan 6C Rapid DNA Analysis System. As a result of this approval, NDIS accredited laboratories can use the ANDE system to analyze DNA samples and search resulting ANDE DNA IDsTM against the FBI's Combined DNA Index System (CODIS) without manual interpretation and technical review.

Under the Rapid DNA Act of 2017, DNA will be taken from arrestees in police booking stations with the intent of identifying arrestees who are wanted in connection with rapes, murders, and other crimes while they are still in custody (instead of releasing them without testing for DNA evidence as it currently is). The rapid identification of suspects through

DNA testing of arrestees can dramatically reduce violent crime by identifying repeat offenders. ANDE is the first Rapid DNA System to receive NDIS approval for use in police booking stations imposed by the Rapid DNA Act of 2017 [72] (Figure 2.3) [22]. NDIS law enforcement booking stations will be able to use the ANDE 6C Series G starting February 1, 2021 [71].

Figure 2.3: The ANDE Rapid DNA instrument - FBI NDIS Approved [22]

2.2.3 Summary of rapid DNA technology uses

In addition to being essential in forensics, rapid DNA technology is critical to immigration controls. As of May 2019, ICE and U.S. CBP (Customs and Border Protection) piloted a program with Massachusetts-based ANDE to expose' 'family unit fraud," which involves asylum-seekers posing illegally as biologically related by traveling with their children. An expansion contract worth $5.2 million was awarded to Bode Cellmark Forensics, Inc. in June 2019. Test refusal may be considered in the amnesty conditions, despite the fact the tests are voluntary. Privacy advocates question whether the program is genuinely voluntary, uses a narrow definition of a family, and whether it could lead to errors.[[73]. The US Department of Defense and related intelligence agencies have supported the development of rapid DNA systems that can be applied to forward field operations. In addition to developing DNA profiles from buccal swabs, the RapidHIT and ANDE systems can also analyze tiny amounts of samples, like the residue left on a glass after a suspect had used it. According to reports, the US military has tested several DNA-based systems to combat terrorism globally. [74].

2.3 Types of signal non-idealities (Drop-out, Drop-in, Stutter)

Getting results from low amounts of template DNA still has several challenges. However, STR typing results have been obtained from as little as a single buccal cell using fluorescent multiplexes. An underlying scientific obstacle of stochastic amplification can be faced when attempting to produce results with a low copy of DNA because of a random amplification of alleles. When a weakened DNA template-to-primer-to-polymerase ratio exists, the stochastic effects occur. However, sample enrichment techniques like whole genome amplification have been used [36]. Four artifacts commonly arise when enhanced detection methods are employed. (1) Allele drop-out, an allele present in the original sample, fails to amplify due to stochastic effects. See Figure 2.4. [23]

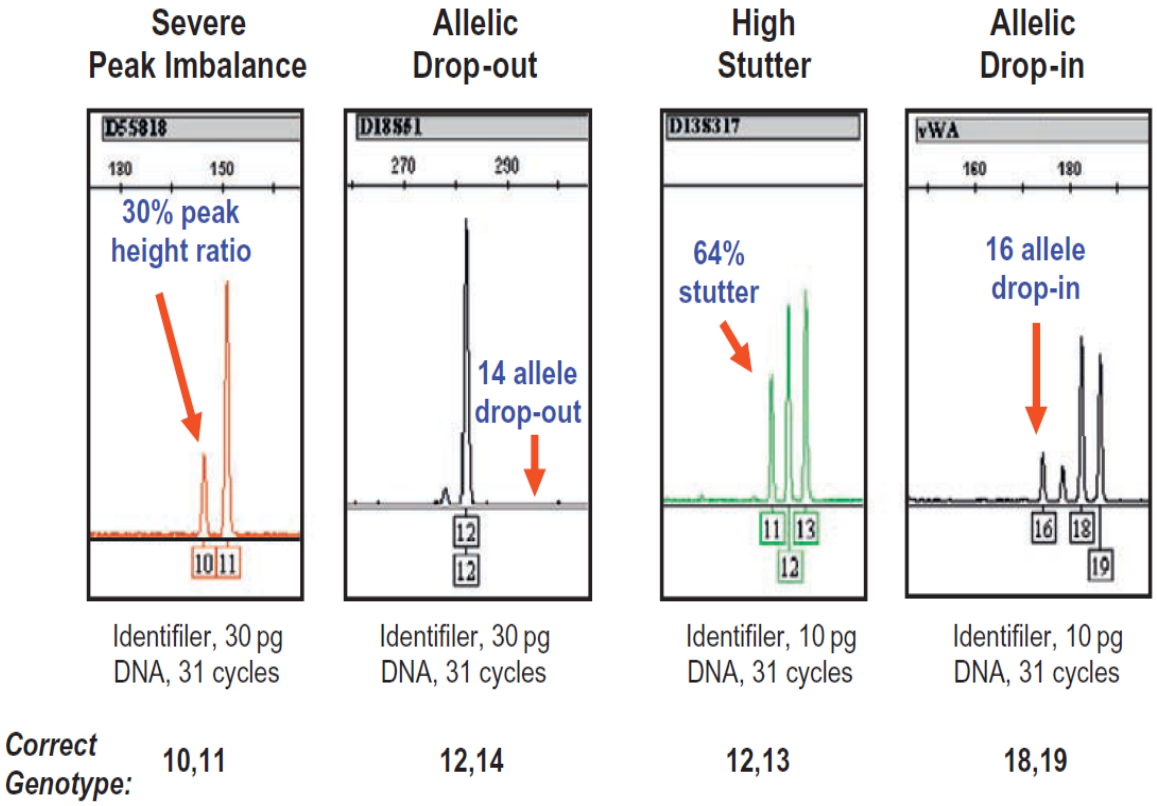


Figure 2.4: Challenging with low copy DNA samples. (2) Allele drop-in, due to sporadic contamination, more alleles are often observed in the DNA profile (3) Heterozygote peak imbalance, because of stochastic PCR amplification it is often exacerbated and (4) Increased stutter, when the stutter products are higher than the normal 5% to 10% of the symbolic allele.[23]

When testing with small amounts of DNA, stochastic effects cannot be avoided, so basically there are two thoughts on how to work with these types of samples: (1) try to confine the impact of stochastic variation by more testing and cautious expedition guidelines based on affirming studies, or (2) stop testing or decipher the data before it becomes low enough to be in the stochastic realm [44] Those who support the first method usually increase their procedure responsiveness, like enlarging the number of PCR cycles. They may get as many possible ones from the limited sample. While the other approach usually includes duplicate testing and the expansion of consensus profiles.

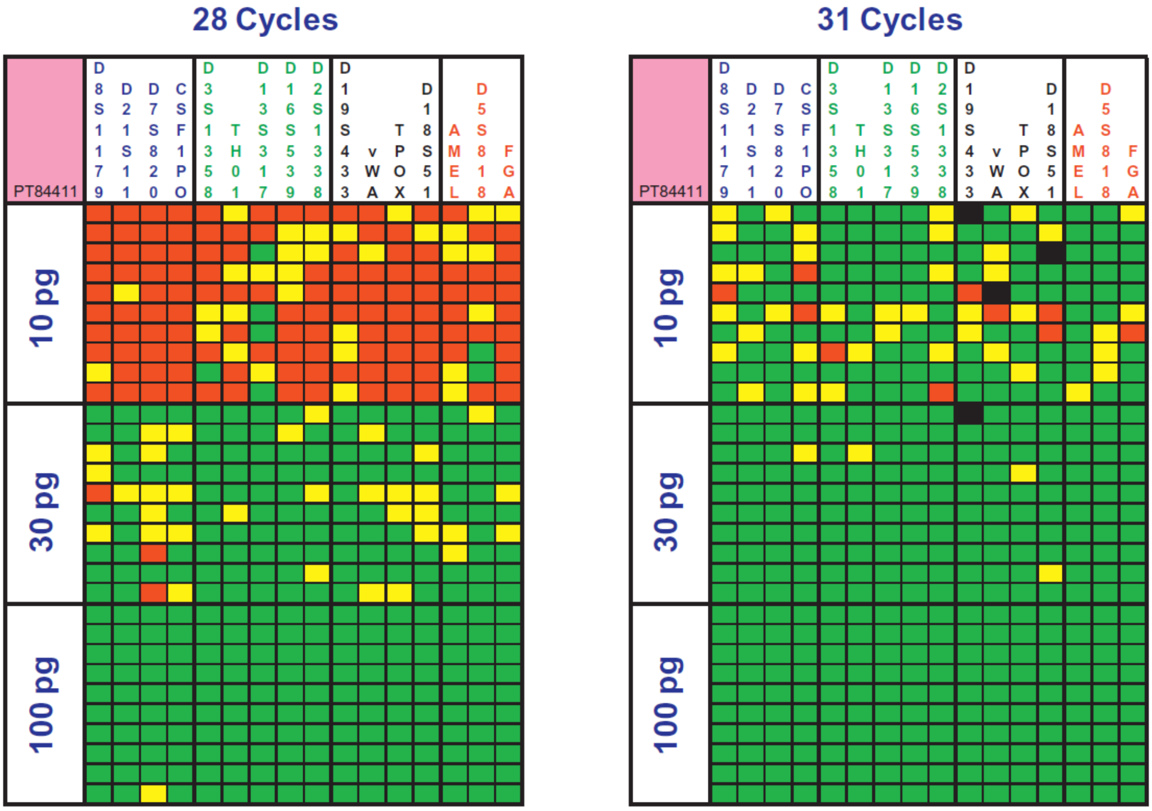


Figure 2.5: Summary of Performance and sensitivity for Identifiler with 28 and 31 cycle data [24]

In Figure 2.5 [24] the results of one sample tested using the Identifiler kit at 28 cycles and 31 cycles can be seen. By using a larger number of cycles (31 cycles) more correct genotypes can be obtained, as shown by the green squares in the figure. The success rate for a correct heterozygous call is shown to have improved with the three extra PCR cycles from 60% (290/480 possible) to 88% (423/480 possible). The most substantial improvement from locus drop-out to correct genotype can be seen at the 10pg level as shown in the figure with red-to-green squares, where full genotype recovery was enhanced from 4% (7/160) to 68% (108/160). When using 31 cycles, the three DNA amounts tested show that the allelic drop-out amounts fell from 14% (65/480) to 9% (43/480). Also, the locus drop-out lowered from 26% (125/480) to 2% (10/480), showing that the overall success rate and sensitivity are enhanced by increasing the number of cycles. Unfortunately, allele drop-in occurred in four cases while using a more significant number of PCR cycles, as shown by the black squares in the figure, while none occurred when using a lower number of cycles. The probability of allele drop-in reveals the significance of duplicate amplifications growth of consensus profiles elude miscalls when employing increased interrogation techniques. Performing ten amplifications cannot be done if the sample recovered DNA is restricted to a limited amount of amplification. Furthermore, if a larger DNA sample were available, the example would most likely be used in a single high copy DNA analysis rather than dividing it into multiple samples. When doing forensic casework, low amounts of DNA template retrieved from evidentiary items and DNA degradation or PCR inhibition can complicate elements. It can be tough to retrieve the correct profile of the first contributors to the mixture, the stochastic effects become worse, and the single components will be even smaller in size when we are observing mixtures at low DNA amounts. Statistical procedures that account for allele dropout become an option when a complete profile cannot be retrieved, even by using replication and consensus.

2.4 DNA profile generator (Data source)

To prove any new biotechnology method or computer program, you need to run it and test it against proven legacy methods. This will provide a standard ground accuracy that can later be a globally used solution. For this reason and during this research, the most significant difficulty was providing enough data from random samples to support the used technique. This is why a new software is invented, which most labs will be using soon. It is called (DNA profile generator).

This computer software has been developed in Matlab; its main idea is to generate all possible situations in humanity. This's based on population genetics taken from the NIST DNA database.

FBI population database release of 2017 is used. This release has nine races. Each of those races differs from the others in two major aspects. Firstly, the locus that might exist in that race, and secondly, the alleles that might be found. The final DNA profiles are generated using $GlobalFiler^{TM}$ chemical kits, and the script is built to match that. Therefore, only the locus in that kit will be shown.

There are 87 possible allele sizes per locus. Each locus should have two alleles. One to represent each parent. According to population genetics, each allele has a different probability of appearing in the profile. The script will randomly choose from available options. This depends on its chance and will fill out the loci with two alleles. Then it moves to the following loci until it picks up all alleles and builds an artificial DNA profile completely extracted from those probabilities. At the same time, it might match the DNA profile of someone who lives somewhere else.

2.5 Signal processing methodology

Measuring peaks in a signal and their measurement of positions, heights, widths, or areas are common aspects of scientific data processing. In this technique, we take advantage of the fact that a peak's first derivative has a downward crossover at its maximum [75]. In actual experimental signals, however, there will be a lot of false zero crossings as a result of random noise. This problem is avoided by first smoothing the signal's first derivative, then looking for zero crossings in the downward direction and selecting only those whose slope exceeds a certain minimum. This property is called the "slope threshold") at a point where the original signal exceeds a certain minimum (called the "amplitude threshold"). The smooth width, slope threshold, and amplitude threshold can be carefully adjusted so that the filter detects only the desired peak and ignores peaks that are too narrow, too wide, or too small.

Further, this approach can be extended to estimate each peak's location, height, and width in the vicinity of a zero-crossing by fitting a segment of the unsmoothed signal to a least-squares curve. Consequently, even if heavy smoothing of the peaks is necessary to discriminate from noise, peak parameter estimation by curve fitting is not adversely affected by smoothing. The effect of random noise in the signal is reduced by curve fitting over multiple points in the peaks. This technique can be measure peak heights and positions. However, measurements of peak widths and areas are more accurate when peaks are Gaussian shaped.

The functions locate the positive peaks in a noisy data set, perform a least-squares curve-fit on the upper half of the peak, then compute the peak's position, height, and width. We define FitWidth as the number of points around each peak top fitted in the script (6th input argument). In addition to the peak number and the estimated position, height, width, and area of each peak, the other arguments will return a list (in matrix P). The program can detect and curve-fit over 2,800 peaks per second in very large signals. A signal with multiple points in each peak will find this helpful, rather than spikes with only one or two points.

2.5.1 Fitting Gaussian peaks:

The natural logarithm transformation is used to convert a positive Gaussian peak to the form that can be fit by polynomial curve fitting. [76] which has the fundamental functional form $exp(-x)^2$,), into a parabola of the form $-X^2$, Which can be fit with a second order polynomial (quadratic) function:

$$\gamma = \alpha + bx + c(\varkappa)^2 \tag{2.1}$$

The equation for a Gaussian peak is:

$$y = \mathbf{h} * exp(-((x - \mathbf{P})/(1/(2 * sqrt(ln(2))) * \mathbf{w}))^2)) \tag{2.2}$$

where **h** is the peak height,**P** is the x-axis location of the peak maximum, **w** is the full width of the peak at half-maximum. The natural **log** of y can be shown to be:

$$log(\mathbf{h}) - \left(4 * \mathbf{P}^2 log(2)\right) / \mathbf{w}^2 + \left(8\mathbf{P} * log(2)\right) / \mathbf{w}^2 - \left(4 * x^2 log(2)\right) / \mathbf{w}^2 \tag{2.3}$$

Which is a quadratic form in the independent variable x because it is the sum of x^2, x, and constant terms. Expressing each of the peak parameters **h**, **p**, and **w** in terms of the three quadratic coefficients, equation to calculate all peaks parameters. [75] will show the peak (height, maximum position, and width) can be calculated from the three quadratic coefficients **a**, **b**, and **c**; it's a classic "3 unknowns in 3 equations" problem. The peak height is given by:

$$exp\left(\mathbf{ac} * (\mathbf{b}/(2 * \mathbf{c}))^2\right) \tag{2.4}$$

The peak position by

$$-\mathbf{b}/(2 * \mathbf{c}) \tag{2.5}$$

The peak half-width by:

$$2.35482/\left(sqrt(2) * sqrt(-\mathbf{c})\right) \tag{2.6}$$

This is called "Caruana's Algorithm" [77]. The area under the Gaussian peak of height "height" and full width at half maximum "width" can be shown to be:

$$1.064467 * height * width \tag{2.7}$$

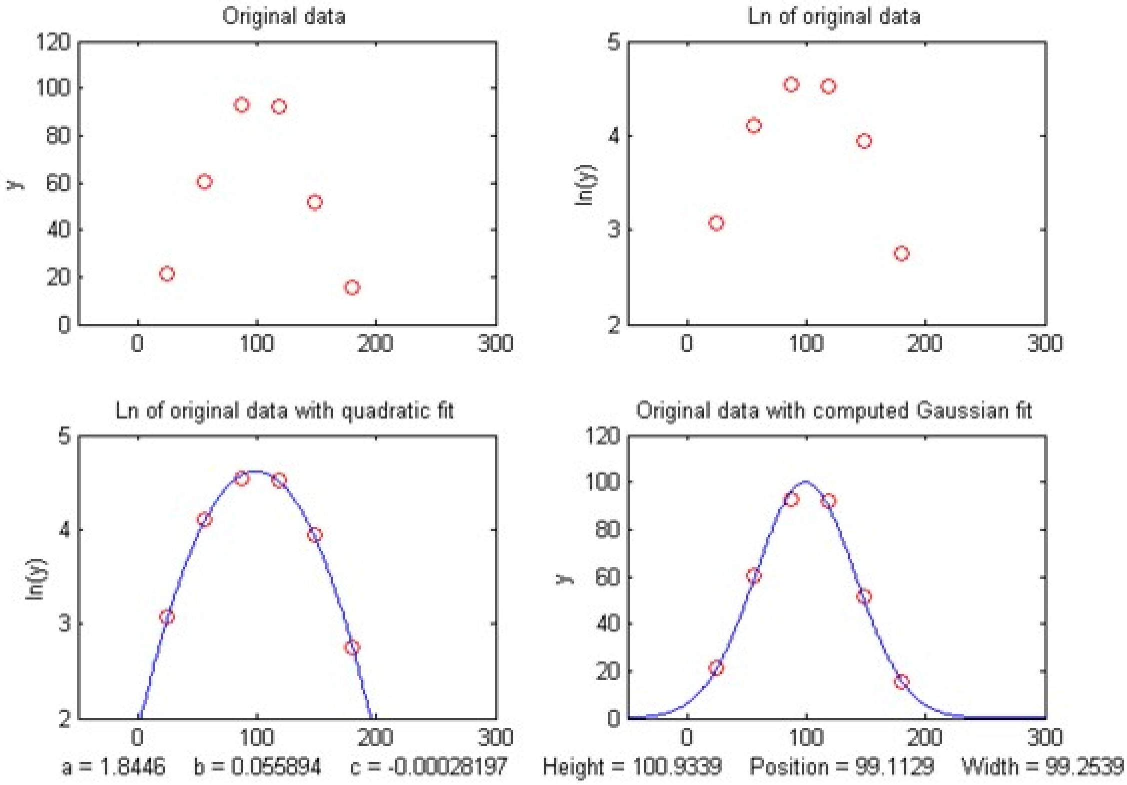


Figure 2.6: Shows how to calculate peaks parameters using Gaussian fit[25]

One advantage of this type of Gaussian curve fitting, as opposed to simple visual estimation, is illustrated in the figure (2.6). The signal is a Gaussian peak with a true peak height of 100 units, a true peak position of 100 units, and a true half-width of 100 units, but it is sparsely sampled only every 31 units on the x-axis.

Table below data points, shown by the red points in the upper left, has only 6 data points on the peak itself. If we were to take the maximum of those 6 points (the 3rd point from the left, with x=87, y=95) as the peak maximum, we'd get only a rough approximation to the true values of peak position (100) and height (100). If we were to take the distance between the 2nd the 5th data points as the peak width, we'd get 3*31=93, compared to the true value of 100. If we were to attempt to estimate the area under the peak from those measurements, we would get $1.064467 * 95 * 93 = 9404.6$, much lower than the theoretical width of $1.064467 * height * width = 10644.67$.

X	**Y**
25	21.56
56	60.3
87	93.16
118	92.27
149	51.71
180	15.65

Table 2.1: Original data points with Computed Gaussian fit

However, taking the natural log of the data (upper right) produces a parabola that can be fit with a quadratic least-squares fit (shown by the blue line in the lower-left panel Figure(2.6))[25]. From the three coefficients of the quadratic fit, we can calculate much more accurate values of the Gaussian peak parameters, shown at the bottom of the figure ($height$ = 100.93; $position$ = 99.11; $width$ = 99.25; $area$ = 10663). The panel in the lower right shows the resulting Gaussian fit (in blue) displayed with the original data (red points). The accuracy of those peak parameters (about 1% in this example) is limited only by the noise in the data. In order for this method to work properly, the data set must not contain any zeros or negative points; if the signal-to-noise ratio is very poor, it may be useful to skip those points or to pre-smooth the data slightly to reduce this problem. Moreover, the original Gaussian peak signal must be a single isolated peak with a zero baseline, that is, must tend to zero far from the peak center. In practice, this means that any non-zero baseline must be subtracted from the data set before applying this method.

2.5.2 Slope method and math details:

The least-squares best fit for an x,y data set can be computed using only basic arithmetic. Here are the relevant equations for computing the slope and intercept of the first-order best-fit equation:

$$y = intercept + slope * x \tag{2.8}$$

as well as the predicted standard deviation of the slope and intercept, and the coefficient of determination, R^2, which is an indicator of the "goodness of fit".

(R^2 is 1.0000 if the fit is perfect and less than that if the fit is imperfect).

n = number of x,y data points

$$sumx = \sum x \tag{2.9}$$

$$sumy = \sum y \tag{2.10}$$

$$sumxy = \sum x * y \tag{2.11}$$

$$sumx^2 = \sum x * x \tag{2.12}$$

$$meanx = \frac{sumx}{n} \tag{2.13}$$

$$meany = \frac{sumy}{n} \tag{2.14}$$

$$\textbf{slope} = \frac{n * sumxy - sumz * sumy}{n * sumx^2 - sumx * sumx} \tag{2.15}$$

$$\textbf{intercept} = meany - (slope * meanx) \tag{2.16}$$

$$ssy = \sum (y - meany)^2 \tag{2.17}$$

$$ssr = \sum (y - intercept - slope * x)^2 \tag{2.18}$$

$$R2 = 1 - (\frac{ssr}{ssy}) \tag{2.19}$$

$$\textbf{Standard Deviation of the slope} = SQRT\left(\frac{ssr}{(n-2)}\right) * SQRT\left(\frac{n}{(n * sumx^2 - sumx * sumx)}\right) \tag{2.20}$$

$$\textbf{Standard Deviation of the intercept} = SQRT\left(\frac{ssr}{(n-2)}\right) * \\ SQRT\left(\frac{sumx^2}{(n * sumx^2 - sumx * sumx)}\right) \quad (2.21)$$

(In these equations, Σ represents summation; for example, Σx means the sum of all the x values, and Σx $* y$ means the sum of all the x * y products, etc).

The last two lines predict the standard deviation of the slope and the intercept, based only on that data sample, assuming that the deviations from the line are random and normally distributed. These are estimates of the variability of slopes and intercepts you are likely to get if you repeated the data measurements over and over multiple times under the same conditions, assuming that the deviations from the straight line are due to random variability and not systematic error caused by non-linearity. If the deviations are random, they will be slightly different from time to time, causing the slope and intercept to vary from measurement to measurement, with a standard deviation predicted by these last two equations.[78]

The reliability of these standard deviation estimates depends on assumption of random deviations and also on the number of data points in the curve fit; they improve with the square root of the number of points.

Chapter 3

STR Generator

To verify the Peak Finding method, two different types of data were considered: 1) real data from an actual rapid DNA instrument, and 2) artificial STR data. In Case 2, a software tool was developed to create an unlimited number of STR profiles based on population genetics statistics [69]. The enhancement tool was tested on artificial STR profiles and samples from the FBI National Institute of Science and Technology (NIST) database to ensure credibility. Creating each STR profile included four main steps: 1) creating the main structure for the artificial signal, 2) degrading the sample based on an exponential function, 3) adding a random noise signal to it, and 4) creating several profiles based on user entry (Figure 3.1). The following sections discuss how each step was accomplished, the difficulties along the way, and how these difficulties were solved.

3.1 Overall STR profile structure

The FBI's population genetics release of 2017 was used to build each profile, as shown in Appendix A, for the blue dye channel. This release contained all alleles in each locus and its probability was registered from collected samples. A system was built for each locus to identify all possible outcomes. The outcomes for each locus were then listed in a table for a special script to randomly choose alleles from each table depending on its weight (probability), as shown in Appendix B. The GlobalFiler chemical kit was used as a reference while developing the script. The script identified all alleles that may appear during alleles calling progress in four different dyes (blue, green, yellow, and red). It was designed to select an allele and generate one dye before moving on to the next allele. Profiles were generated via the script in the following order: blue, green, yellow, then red. Each dye had the following alleles:

- Blue dye: D3S1358, vWA, D16S539, CSF1PO, TPOX
- Green dye: Yindel, AMEL, D8S1179, D21S11, D18S51, DYS391
- Yellow dye: D2S441, D19S433, TH01, FGA
- Red dye: D22S1045, D5S818, D13S317, D7S820, SE33

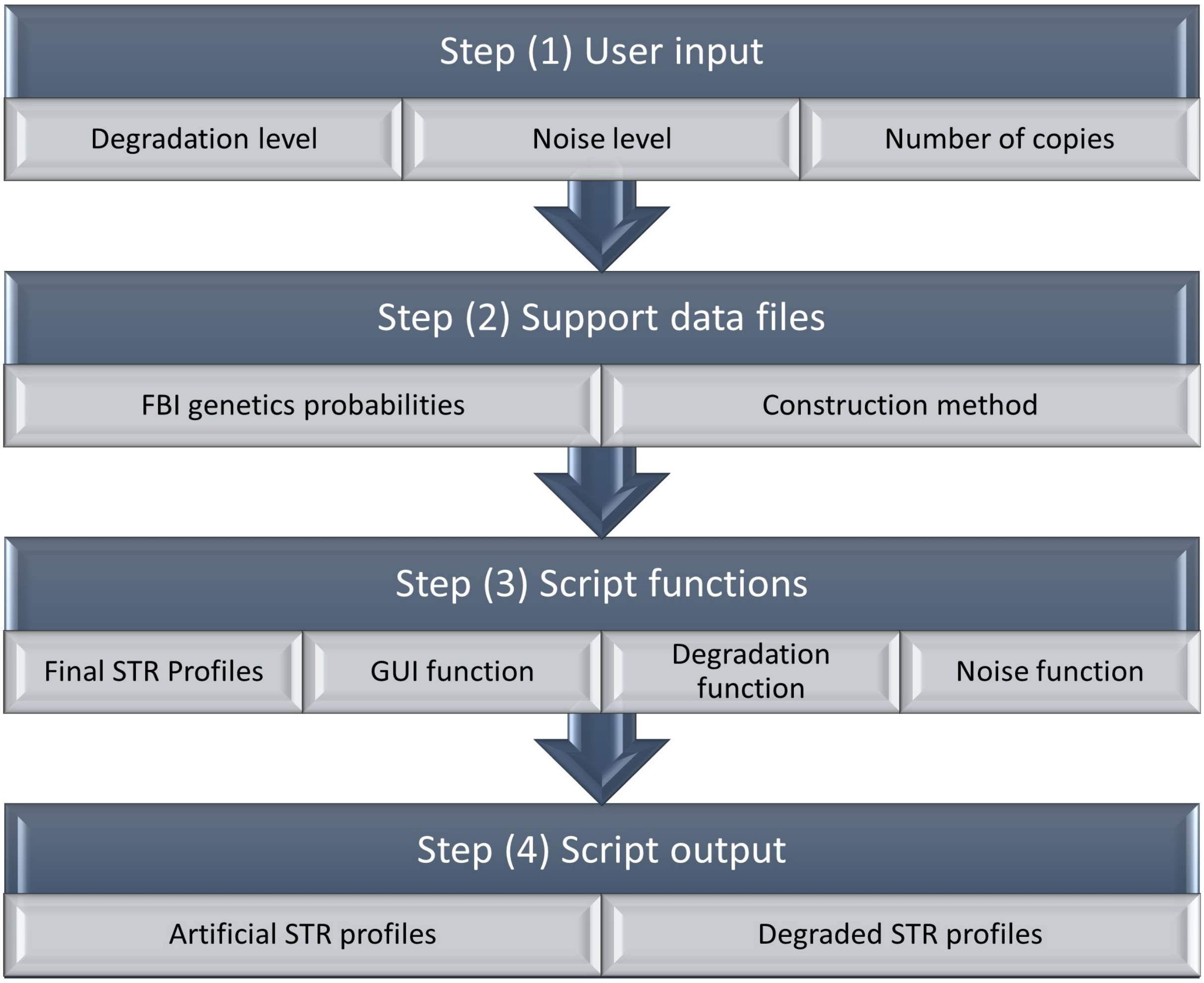


Figure 3.1: The overall block diagram shows step by step functionality of the STR profile generator.

3.2 Detailed Functions description:

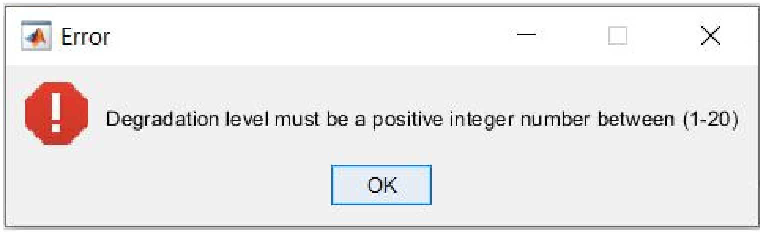
The script uses a combination of inputs a user can determine depending on why the user is generating a profile. The script will create two folders in the same location as the main file. The first folder will have a determined number of non-degraded profiles while the second folder will have the same number of artificial profiles but in a degraded form. The following sections discuss each step in detail and what choices can change the outcome of the generated profiles.

3.2.1 User input

After running the executable file, the user determines the following three inputs: the number of generated profiles, the level of degradation, and the required amount of added white Gaussian noise.

- **Degradation level:**

Users can determine the level of degradation depending on why they generate artificial profiles. This is essential to the degraded profiles outcome. This item is discussed more in detail in Chapter 4. Users can enter any number between 1 and 20. If a user tries to enter a character, a negative number, or a positive integer greater than 20, the system errors out and displays an error message (Figure 3.2).



Degradation level must be a positive integer number between (1-20)

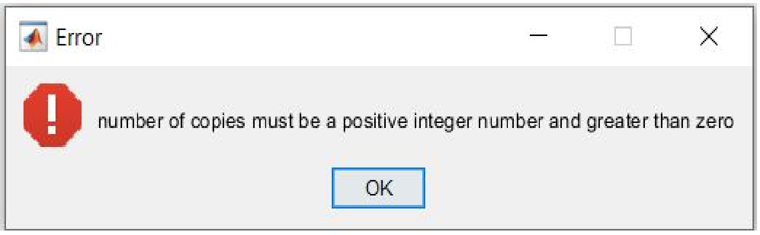
Figure 3.2: The error message when entering an invalid value.

- **Number of artificial profiles:**

Users can determine the total number of artificial profiles based on the number of needed profiles to perform a complete validation for their research purposes. When executing the final STR function, it goes into a “FOR” loop. This “FOR” loop repeats the entire profile generation process to create a different STR profile than the one that has been just created.

The number of artificial profiles is determined by the user at the beginning of the executing process. Choosing an input value equal to “1” will generate two profiles: one non-degraded profile in a .MAT file and an exact matching degraded profile in another file. The non-degraded file will be generated in the “STR sample” folder, while the degraded one will be created in the “degraded STR sample” folder. As another example, choosing a value of “5” will result in five different STR profiles in the “STR sample” folder and another five degraded profiles in the “degraded STR sample” folder.

The user can enter any positive integer to determine the number of copies needed. Failing to enter a positive number greater than zero will cause the script to error out, thus displaying an error message (Figure 3.3).



number of copies must be a positive integer number and greater than zero

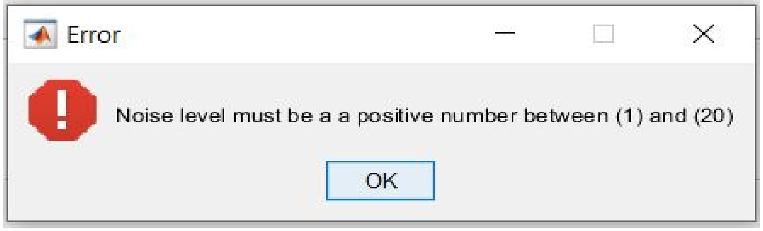
Figure 3.3: The error message when entering an invalid number or character.

- **Noise level:**

White Gaussian noise is chosen due to its similarity with the noise signal that comes from the instrument during the separation/detection process. The user has to add a number between “one” and “twenty”. When smaller numbers are added, the script is designed to add more noise to the artificial STR profile. “One” is very noisy, while “twenty” add the least noise—almost no noise—to the signal. Adding positive numbers outside of the range 1–20, negative integers, and/or characters will lead the graphic user interface (GUI) to promote an error message as shown in Figure (3.4). This function is being described in detail in Chapter 4. If the user enters no input to all the three required fields, all error messages will pop up at once.

3.2.2 Supporting data files

As shown in the block diagram in Figure (3.1), the script will use supported data files (FBI genetics probabilities and construction method) to construct the electropherogram signal. Both files are used each time the script is creating a locus.



Noise level must be a a positive number between (1) and (20)

Figure 3.4: The error message when entering an invalid noise level or character.

- **FBI genetic probabilities**

The table in Appendix A shows that each allele in a certain locus has a different probability to appear in the locus. This probability also differs from one race to another; some races have a higher probability of having a specific allele in their profile while others may not have this allele at all in their profiles. The FBI genetics probability release of 2017 that was used in this study is a database of all races alleles probabilities.

The script starts from the first locus located on the blue dye with the smallest base pair and moves to the second locus with a bigger base-pair allele. Each locus has two alleles: one from the mother and one from the father. In the demonstration of the blue dye in this study, the first allele in the chemical used kit was D3S1358. This allele could be either heterozygous or homozygous. It randomly depended on the weight of its probability in the FBI population genetics release considering the entire data set. Therefore, the script is designed to weigh in the possibility of that allele and to choose the alleles accordingly.

The software was developed using MATLAB, which uses a random selector instruction (see Equation 3.1) to weigh in the possibility of that allele. The entire alleles table from the FBI genetic release of 2017 was inserted as a reference file for this script/instruction to select from (See Appendix B). The system has all alleles, and the possibility of appearing in each locus is hardcoded as a file called (weights) inside the script main source. Once the script executes, it randomly selects the first allele for the first locus; after that, the script moves on and chooses the second allele for that locus. A very high probability for a specific allele in a particular locus means that the profile has a homozygous locus. In this case, the script adds both alleles' heights together, making the allele's height in that specific locus double that of heterozygous alleles in the same locus. If the software chooses a different allele as the second allele for that locus, this results in heterozygous alleles for that particular locus with normal height.

$$datasample(AAD8S1179(:,1),1,'Weights',AAD8S1179(:,3)); \tag{3.1}$$

Considering the locus “D3S1358” as one example in a dataset of one million samples, the allele “14” appeared 87,400 times while allele 15 appeared 9,200 times. Data set can be updated in the FBI genetics probabilities file inside the script source to allow the user to include bigger data set with more accurate alleles probability if needed.

After completing the first locus in the blue dye, the script moves to the next locus and chooses the second locus alleles. Similarly, the second locus alleles might be homozygous or heterozygous. Once both alleles are selected for the second locus in the blue dye, the script continues to the third locus and repeats the same process until all blue dye alleles have been chosen for all five loci the blue dye has (see Figure 3.5).

For advanced chemical kits such as GlobalFiler, the blue dye has five different loci to identify; therefore, the positive control and the ladder need to fill one less locus with different alleles possibilities for the MATLAB script to choose from.

Upon finishing the blue dye, the script moves on to the green dye and repeats the same process, then the yellow dye, and finally, the red dye. All identified dyes are then stored in one file as a separated artificial sample.

- **Allele construction method**

The slope method is used to construct alleles based on the characteristics of each locus. Each locus of the allelic ladder has its specific height, width, and size. Ascertaining those characteristics allows to determine what should be changed to create an allele that possesses the exact specifications as the one that currently exists in the ladder; therefore, the following steps are followed:

- Allele setpoints are grouped to make up one allele.
- Each slope between two following points in those alleles is recorded then compared with the same allele in the same locus with a different ladder sample to verify the most common slope between every two following points. Next, those slopes are used to identify two points for both axes of the signal.
- The most average slope values are saved in an array from which the script can choose when constructing the profile, as shown in the example in Table 3.1.

Slope	8	14.5	9	51	186.5	350	384.5	195	-119.5	-348	-365.5	-230
Y Value	0	16	45	63	165	538	1238	2007	2397	2158	1462	731
X Value	1	2	3	4	5	6	7	8	9	10	11	12

Table 3.1: All slope values for the first allele from the first locus

- After the script has chosen the alleles based on the FBI population genetics file, there are two possible outcomes: homozygosity or heterozygosity. In the case of homozygosity, the alleles are slightly different in height and size compared with those of heterozygosity. This important fact has been considered when constructing and building artificial profiles in the system by multiplying the Y value by two (as shown in Table 3.1) and recalculating the slope.

- The script starts by constructing the alleles based on a set of options available in the slope array so that the two-point slope corresponds to one of the possibilities, which provides similar appearances and characteristics for ladder alleles.

- This script then stores those two points in a temporary final sample array and continues until all the points of an allele have been chosen to match their equivalents in the ladder.

- The script calculates the height and size of the current allele and then compares it with the ladder with an error margin of 5% to ensure that the random allele points are realistically chosen and differ for each artificial sample.

3.2.3 Script main functions

- **Graphic user interface:**

To make the software more appealing and user friendly, a GUI was created to allow the user to easily enter the main three inputs inside the script. The script takes those inputs and executes the main program based on their values.

The GUI, source code, and the rest of the data files were packed in form of an executable file to make it easier to use on any computer without MATLAB being installed on it.

- **Final STR profile:**

Once the user enters all required fields and clicks on the “generate” button on the GUI interface, the GUI will first ensure that all inputs are in the correct ranges. Once that condition is met, the GUI will start executing the final STR profile function. This function works as the main backbone for all the user inputs and supported data files. This function connects all information while executing the program as follows:

1. In case the user running the executable file for the first time, the GUI automatically creates two folders to save the artificially generated profiles: the “STR Sample” folder and the “degraded STR sample” folder.

2. The function starts creating the profiles from the first locus in the blue dye as mentioned earlier; it runs Equation (3.1) which chooses two different alleles for that locus.

3. Once the code has both alleles for that specific locus, it starts constructing the allele based on the allele construction method mentioned earlier. This process results in an array of zeros except for those two constructed alleles that have been chosen. The code places data points that shape the allele within a range of +/- 5% compared with the ladder.

4. The function then goes back and executes the same equation to choose two different alleles for the second locus and repeats Step 2 above.

5. The function keeps executing Steps 2, 3, and 4 until the entire blue dye is constructed. The result will be an array of one row and 4,500 columns which will plot a 2D plot representing the blue dye.

6. Upon completion, the function moves to the next dye (green), where it starts a new array of zeros with one row and 4,500 columns. The function repeats Steps 1, 2, 3, and 4 and saves the array as a green dye.

7. The same previous process applies to the yellow and red dye which finally results in four different arrays; each one represents a separate dye.

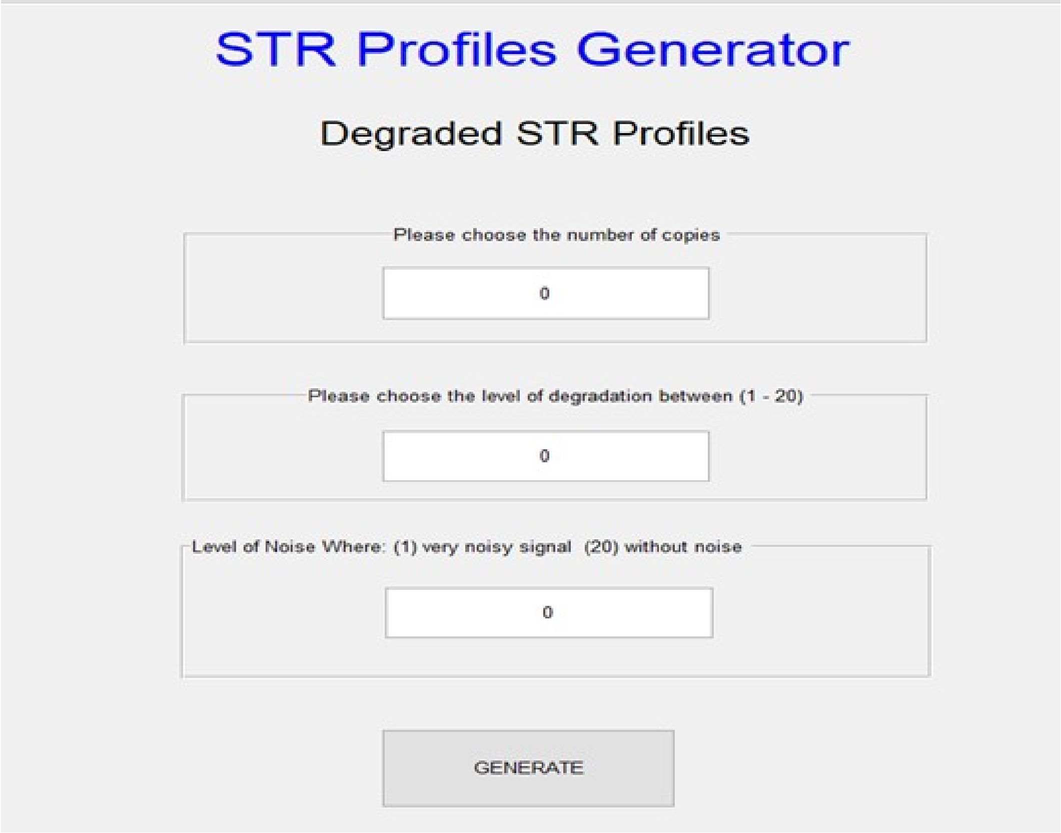
8. After completion, the degradation function occurs for each array separately. The outcome of this is another four different arrays for each array that have been constructed before degradation.

9. The noise function is then called to add noise for each constructed array (dye). The level of noise can be determined by the user's input before generating the profiles.

10. Lastly, the function saves the first undegraded four dyes that represent one sample into one .MAT file, while the other four degraded arrays are saved into another file. Each file is saved in its designated folder.

3.2.4 Script output

The last steps generate two samples: non-degraded in the "STR sample" folder and degraded in the "degraded STR sample" folder. In case of more than one requested copy, the code repeats Steps 2 to 10 until the entire requested samples are generated.



STR Profiles Generator

Degraded STR Profiles

Please choose the number of copies

0

Please choose the level of degradation between (1 - 20)

0

Level of Noise Where: (1) very noisy signal (20) without noise

0

Figure 3.5: The GUI used to generate artificial STR profiles. The user needs to enter three values only: a) the level of degradation between 1 to 20 which corresponds to very low degradation up to extreme degradation, respectively, b) the number of generated profiles that the user wants to generate at one time, and c) the noise level required to be added to the generated profiles. There is no limit on how much the user can generate at one time; the system is enhanced enough to construct up to 1000 profiles in less than one minute.

Chapter 4

Artificial STR degradation

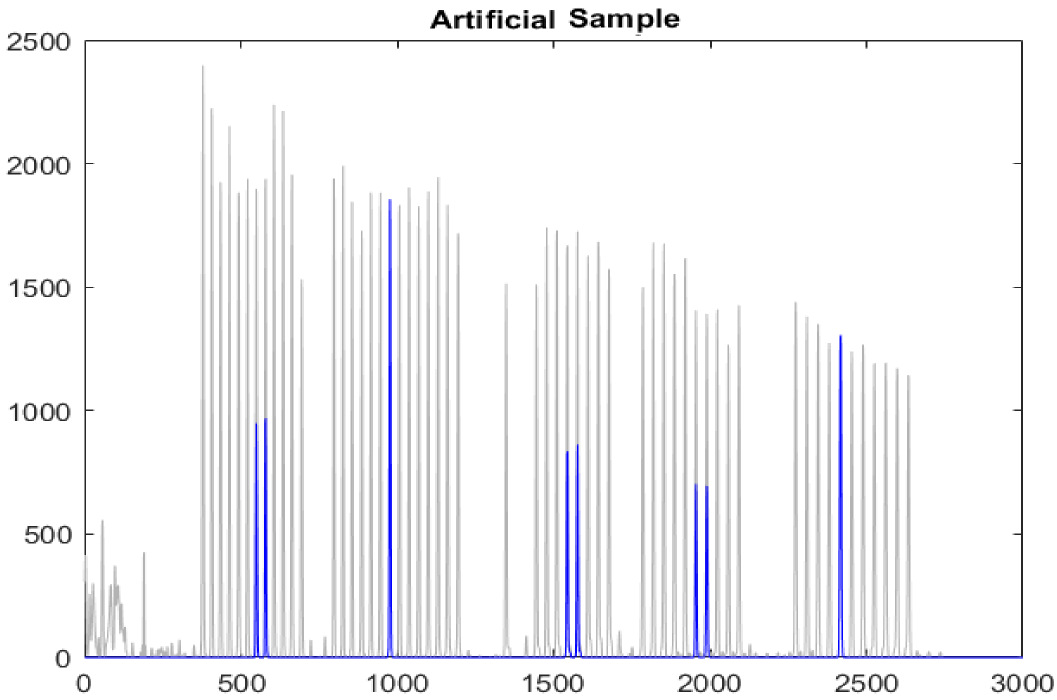
As discussed in the Theory chapter, multiple non-idealities might occur for a degraded DNA sample. Drop out and drop in challenges were addressed in this study. They mainly occur on the alleles displacement of their locations and rely heavily on a possible outcome for a specific locus compared with its ladder. The software can identify drop-out and drop-in alleles. It tests the possibility of being a true allele by comparing the location and base pair of this allele to the ladder to determine whether this might be a dropout allele or a stutter in the same fashion. If the software indicates that there is any drop-in allele, it will also test the possibility of being a true allele by comparing its parameters and characteristics to the ladder. In case the allele is located in an accurate location, the software will compare it to the rest of the finding peaks, including height, to determine whether it is a drop-in or a stutter. It will also label it with a different color to allow the user to visually differentiate the true alleles from drop-in considering the complete STR profile.

4.1 Degrade artificial STR profile

How samples are degraded was determined based on actual data from individuals. All models were collected and degraded to study the pattern those samples followed in order to determine the most common way of degradation. The longer the DNA strand is, the higher chance is for this strand to become broken and degraded. As a result, the amount of DNA decreases during the amplification, which causes longer base-pair strands to be completely dropped out. Four different sets of data (non-degraded, low degradation, medium degradation, and high degradation) were examined. The first set was the original non-degraded samples with all the alleles still present. The low degradation set was exposed to ultraviolet light for 75 μs and partially degraded. The medium degradation set was exposed to ultraviolet light for about 150 μs. Finally, the high degradation was exposed to ultraviolet light for 240 μs. When each degraded sample was compared with the non-degraded one, it was found that the degradation happened in a negative exponential pattern. Applying this theory to the artificial STR profiles that have been generated before, the degraded profile is formed by multiplying the original signal with the following equation:

$$DegradedDNAsample = e^{\omega f(\varkappa)} \tag{4.1}$$

where ω is a user input and represents the level of degradation from 1 low degradation to 20 high degradations. $\mathbf{f}(\varkappa)$ is the artificial DNA signal that needs degradation. This function allows the user to compare a non-degraded STR profile with a degraded version of itself, which also provides an excellent foundation for the enhancement tool (Figures 4.1–4.5).



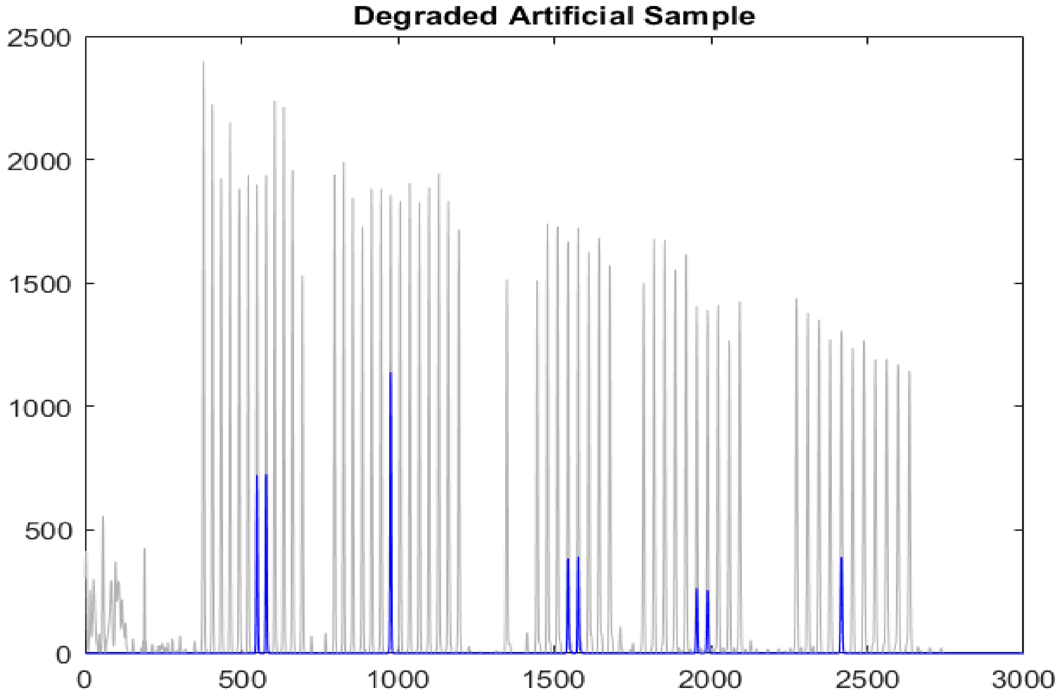
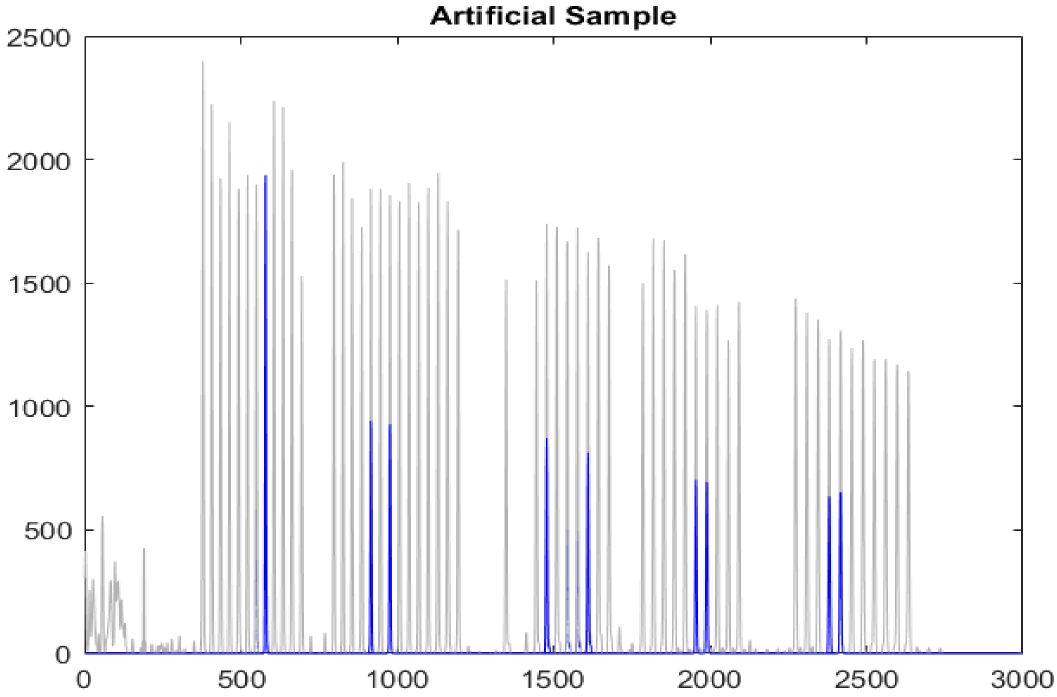


Figure 4.1: The blue dye for the same sample before and after degrading the sample with Level (5) degradation

4.2 Adding artificial noise

When building artificial profiles, it is essential to consider the noise contained in the instrument sensor during the detection process. A MATLAB instruction was used to generate a characteristic noise level similar to that measured in an actual instrument. The user can change the added noise level before saving the final profile. This noise provides another layer of verification to the enhancement tool. A false allele might come up as a drop-in allele if its amplitude is high enough to be considered as an allele. Because the developed tool can detect a false allele, it provides a good solution for the challenge researchers are currently facing. The “awgn” function used in the script adds white Gaussian noise to any signal. The white noise is added to the degraded blue dye sample with a level of “L” which is an input determined by the user. The result is then compiled in the final “blue deg dye sample” which is a 2-D array that can plot the final noisy degraded blue dye signal.



$$Blue.Deg.Dye.Noise = awgn(deg.blue.dye.Sample, L'measured'); \tag{4.2}$$

Where “deg.blue.dye.Sample” is the input degraded sample before adding noise to it, “L” is the level of generated noise, “measured” allows to generate a noise based on the standard deviation of the signal, and “Blue.Deg.Dye.Noise” is the final degraded noisy signal.Figures 4.6–4.8 show the differences between different levels of noise for one sample with and without degradation.

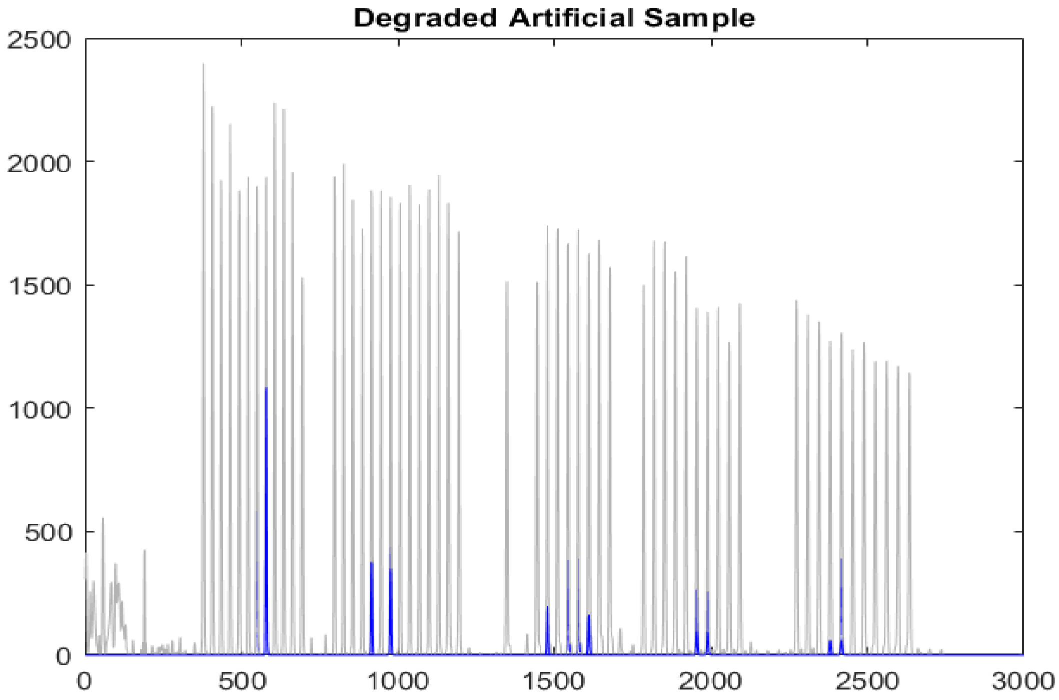


Figure 4.2: The blue dye for the same sample before and after degrading the sample with Level (10) degradation

4.3 Compile and save STR profiles

The last step is to generate profiles based on three main elements: degradation level, noise level, and the number of copies the user wants to generate. The new profiles are saved in the folders mentioned before to be used and analyzed by the enhancement tool.

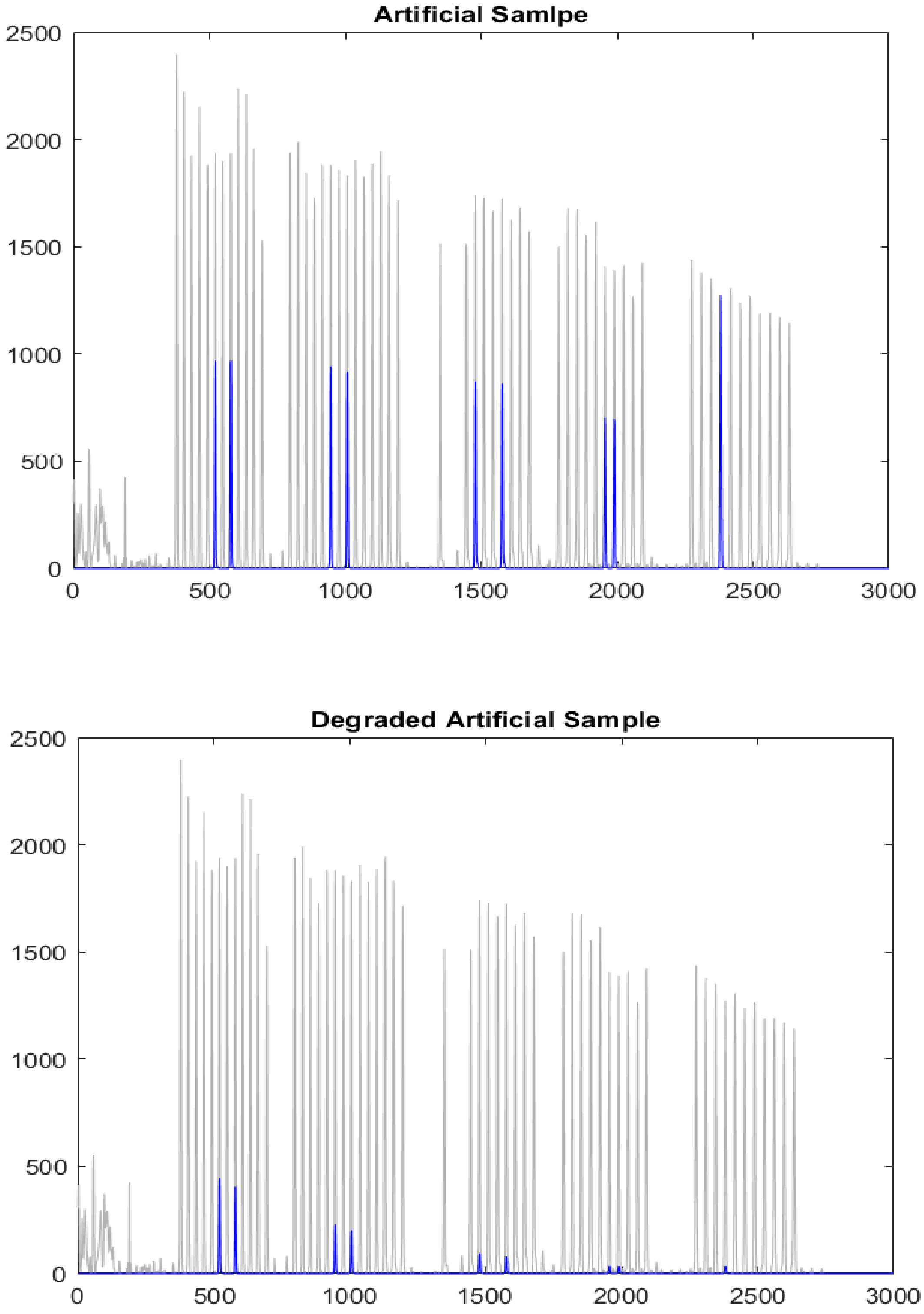


Figure 4.3: The blue dye for the same sample before and after degrading the sample with Level (15) degradation

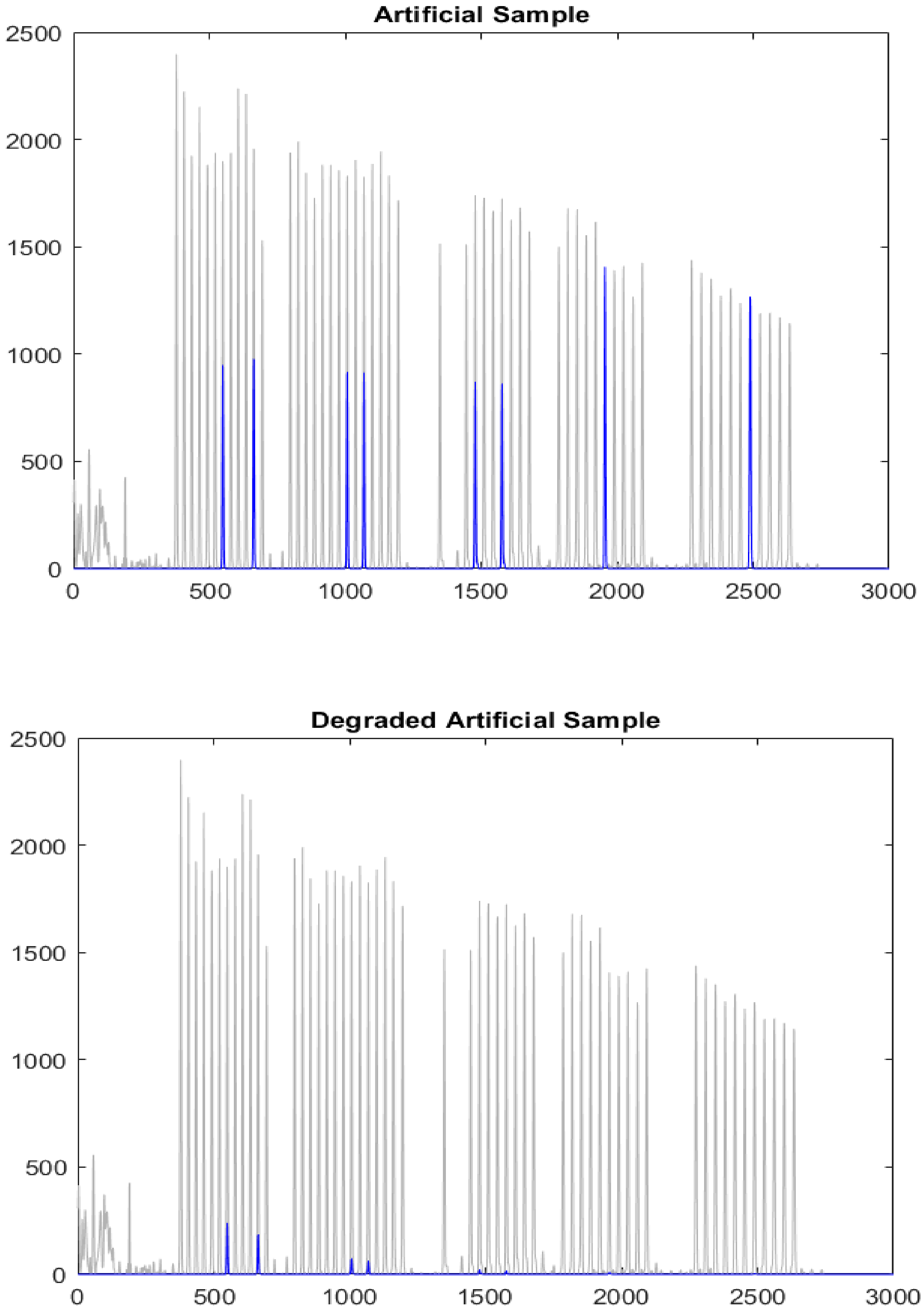


Figure 4.4: The blue dye for the same sample before and after degrading the sample with Level (15) degradation

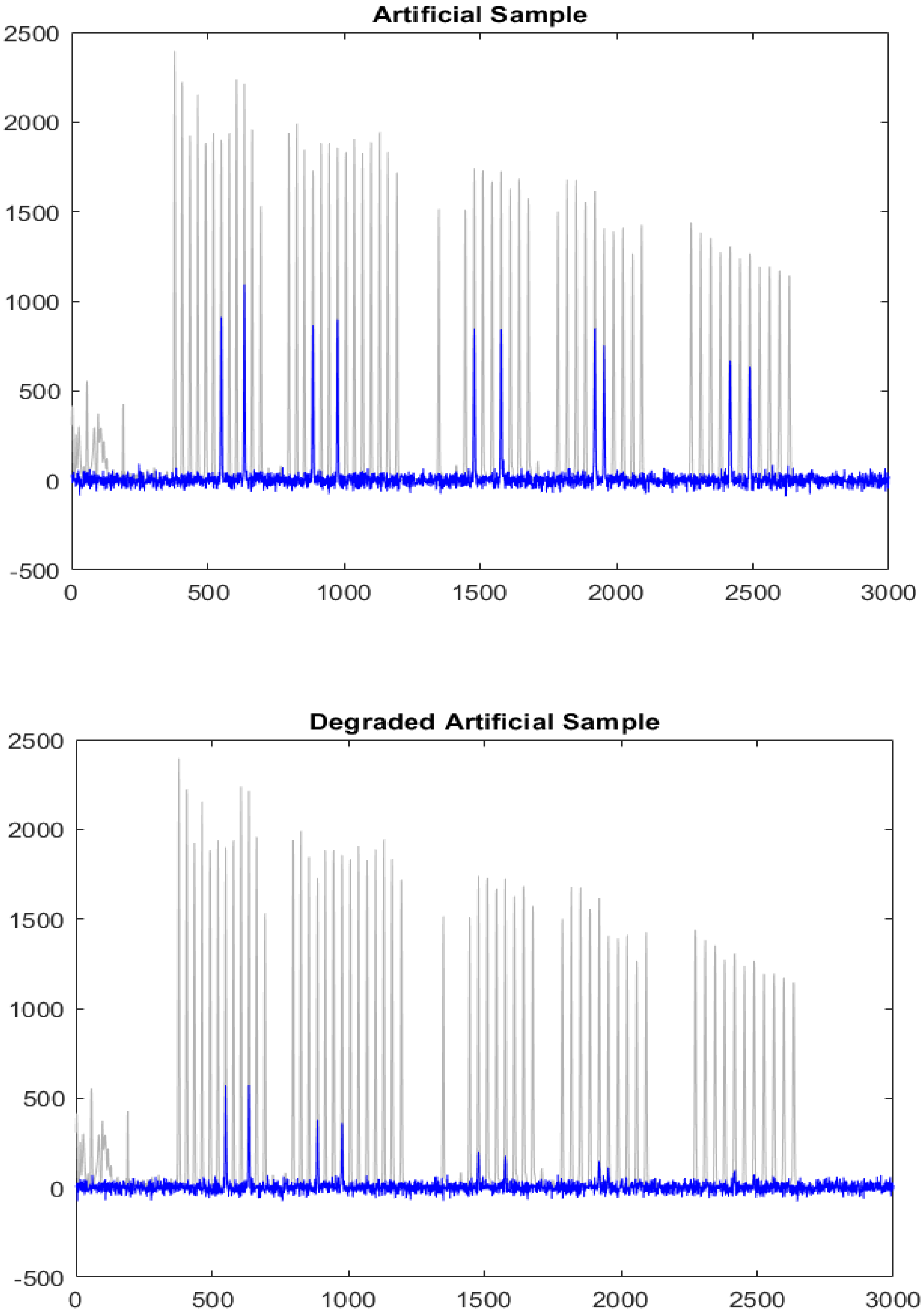


Figure 4.5: Level (3) noise added to a sample before and after the same level of degradation.

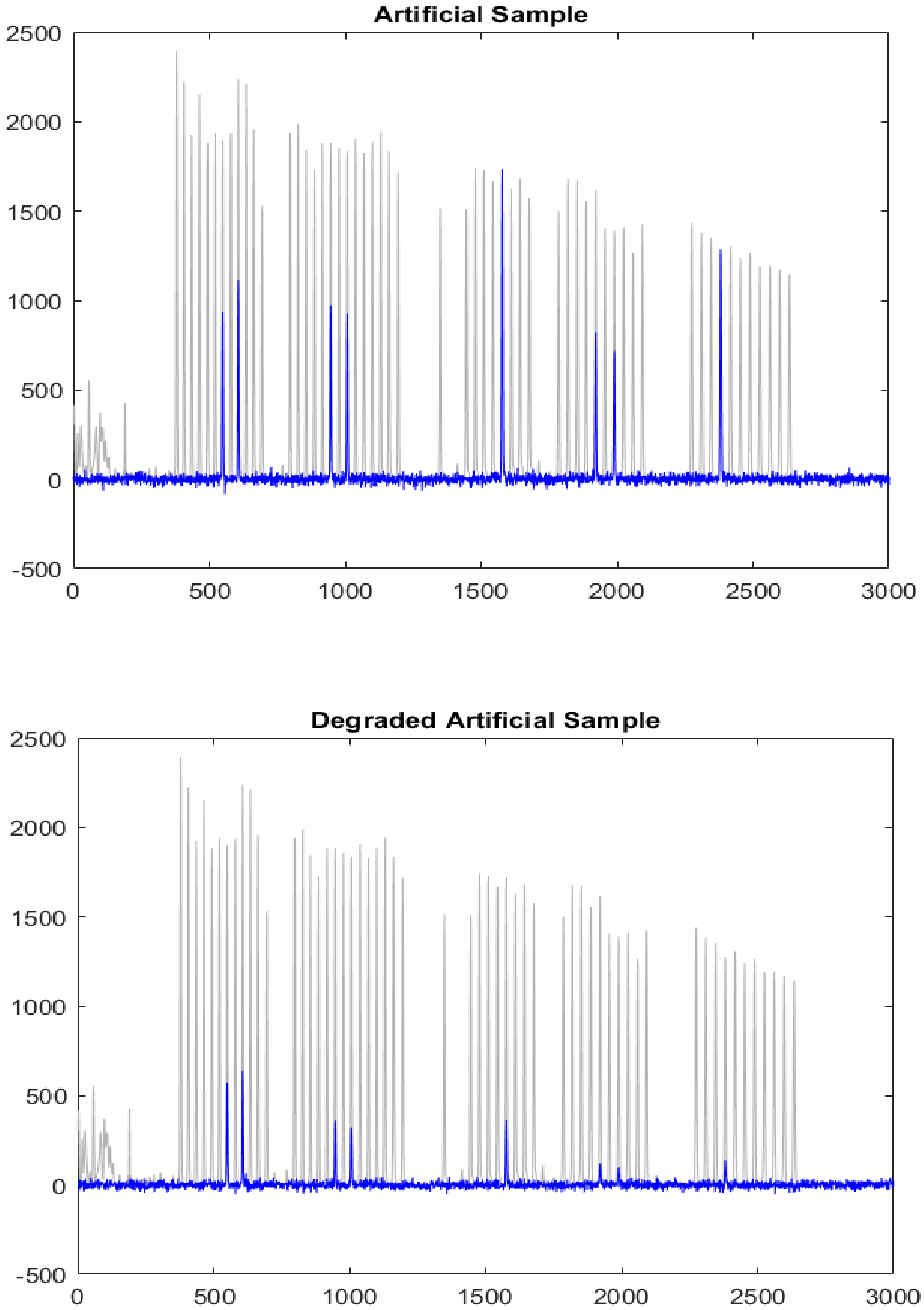
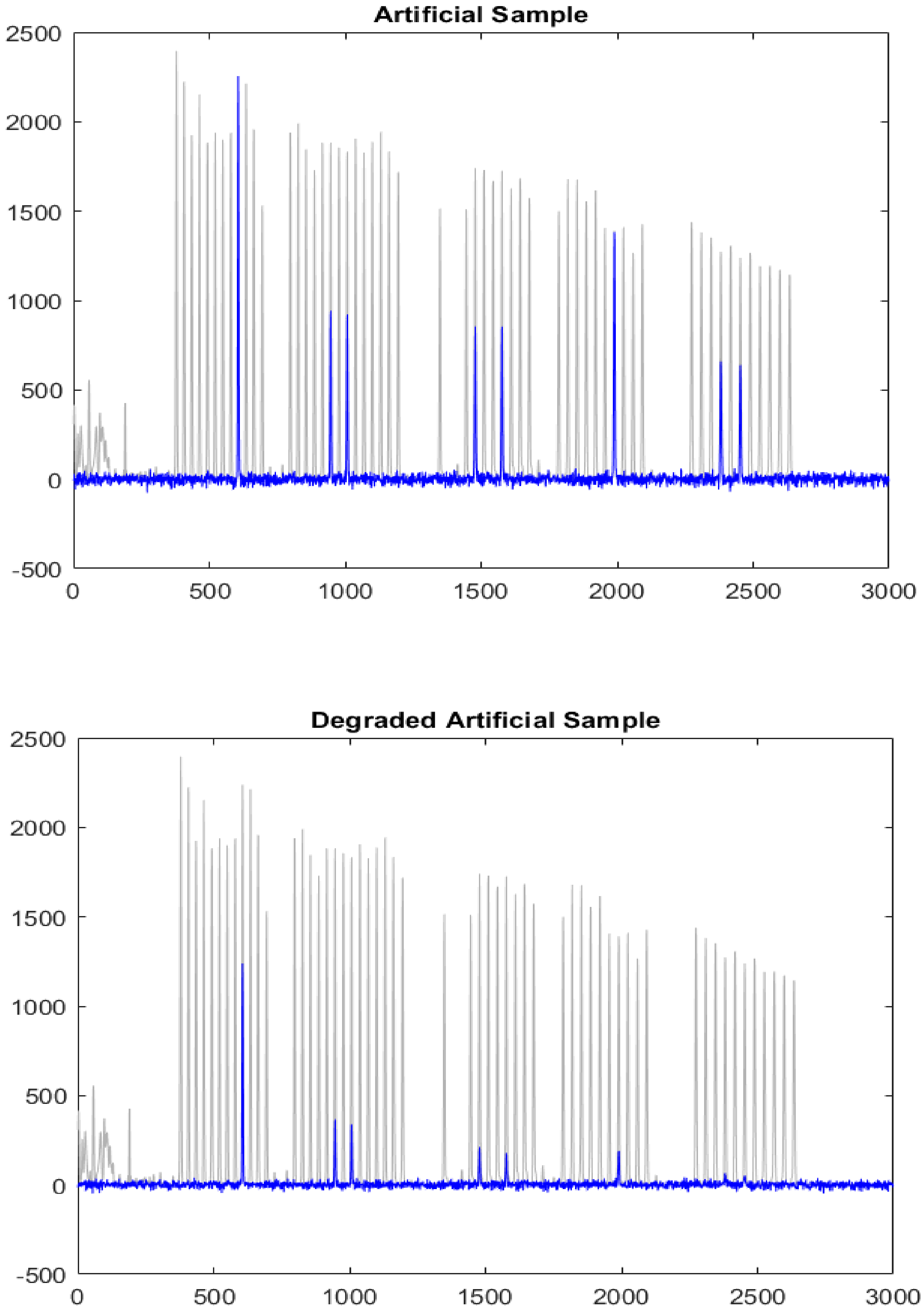


Figure 4.6: Level (7) noise added to a sample before and after the same level of degradation.


Artificial Sample

Degraded Artificial Sample

Figure 4.7: Level (10) noise added to a sample before and after the same level of degradation.

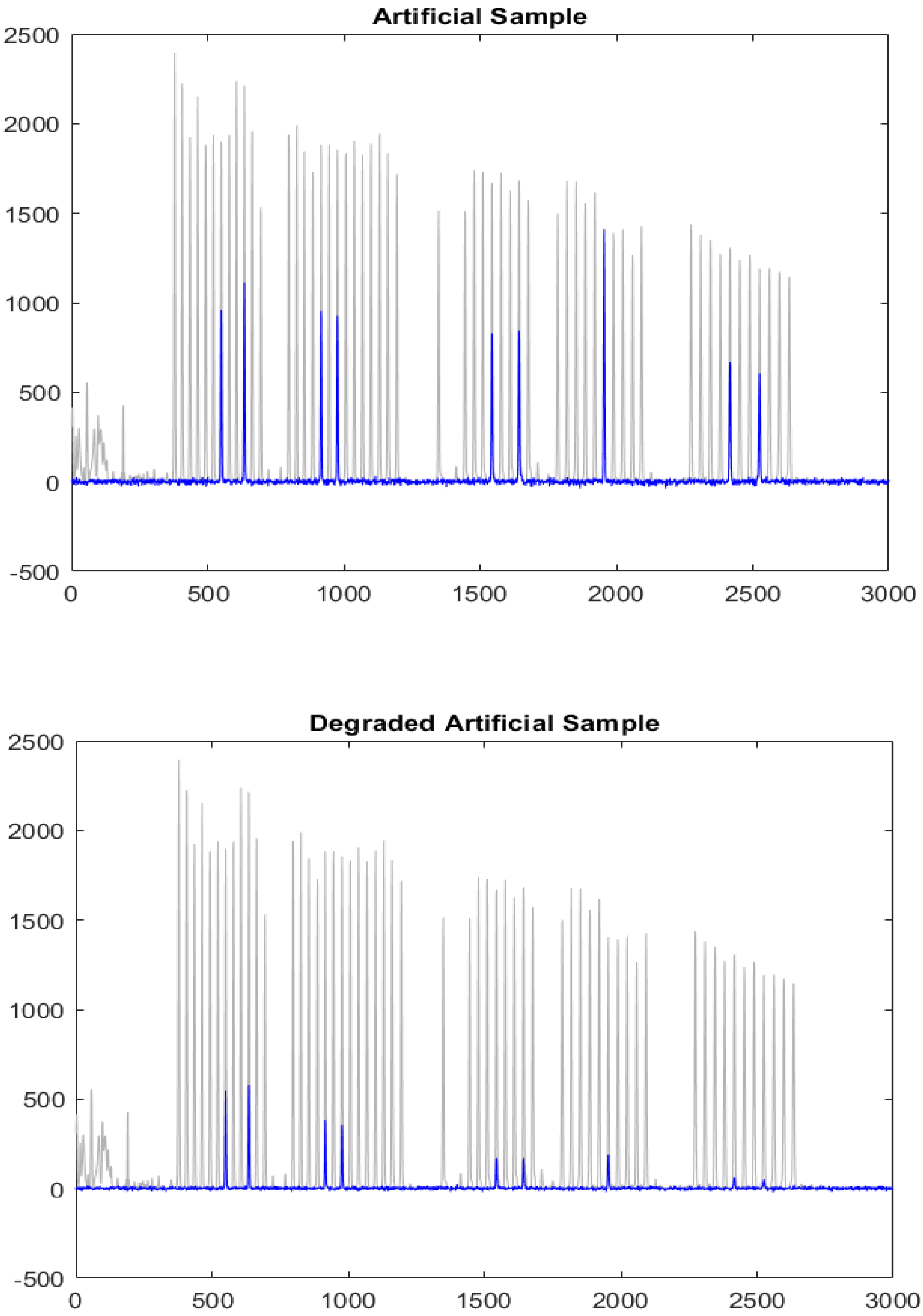


Figure 4.8: Level (15) noise added to a sample before and after the same level of degradation.

Chapter 5

Enhancement of DNA profile's peaks

5.1 Overall structure of the enhancement tool:

A GUI (see Figure 5.1) was designed to represent, view, and demonstrate the methods used: Peak finding function, Analysis function, Quality function, and Match score results. Figure 5.2 represents the relationship between the main GUI functions and other functions that are being used to analyze samples

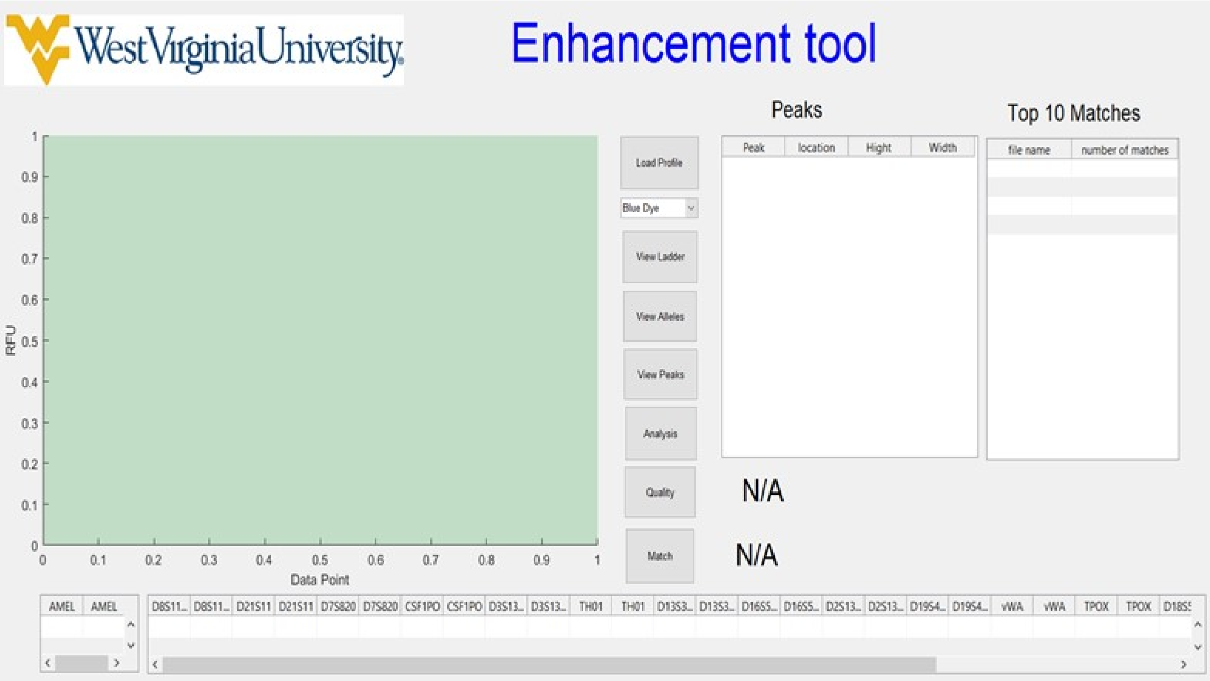


Figure 5.1: The enhancement tool GUI and available functions to analyze STR profiles

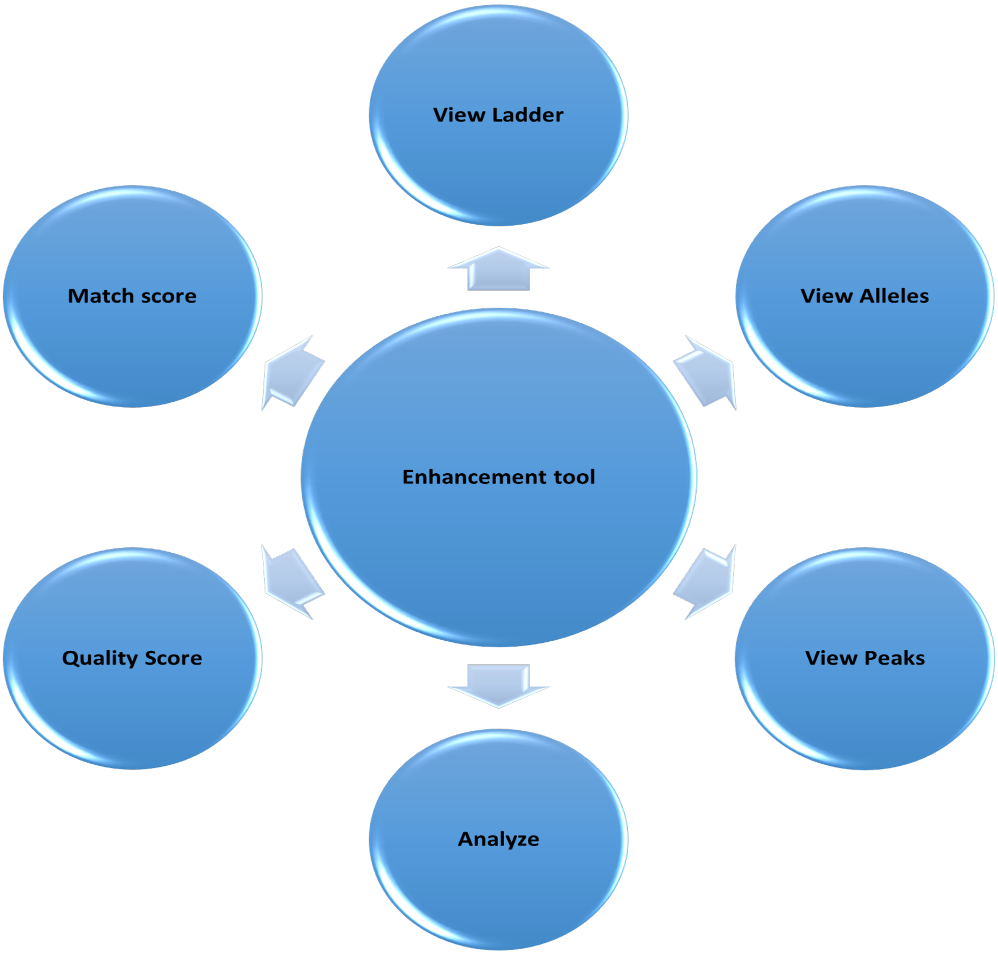


Figure 5.2: Block diagram represents the relationship between the main GUI functions and other functions that are being used to analyze samples

This GUI was developed to visualize, test, and analyze one dye at a time to allow the user to check detected peaks and perform match scores on them using available functions (Figure 5.2). Each function includes a set of instructions to perform. The results are then represented inside the GUI to allow the user to compare and analyze the outcomes.

5.2 Peak finding algorithm

The enhancement tool can be used with real data coming from a lab after converting it into a .MAT file or after running the STR generator tool and generating profiles.

- **Load Profile:**

To load an STR profile, the user clicks the “Load Profile” button to choose a sample that needs to be analyzed. A window then appears to allow the user to browse for and select the .MAT data array file. When the sample is loaded, the user can proceed to the next step—selecting the dye to start analyzing. The user can choose between four different dyes (Blue, Green, Yellow, and Red) as shown in Figure (5.3).

- **View Ladder**

The “View Ladder” button displays the ladder for the chosen dye. Clicking this button leads the GUI to run the “View Ladder” function which evaluate the chosen dye first. It then displays the correct ladder dye that corresponds to the chosen dye from the drop-down menu using the following script instruction:

```
evalin('base',['load ' 'Run_RH200-0076_2017_11_06_10_33.mat']);
popup_sel_index = get(handles.popupmenu1, 'Value');
switch popup_sel_index
case 1 % Blue dye is selected
    L1 = evalin('base','Trace_A6_Ladder1_0076_2017_11_06_10_33');
     B = L1(1,1:3500);
    plot(handles.axes1, B, 'color', [0.7 0.7 0.7]);
case 2 % Green dye is selected
    L2 = evalin('base','Trace_A6_Ladder2_0076_2017_11_06_10_33');
     G = L2(1,1:3500);
    plot(handles.axes1, G, 'color', [0.7 0.7 0.7]);
case 3 % Yellow dye is selected
    L3 = evalin('base','Trace_A6_Ladder3_0076_2017_11_06_10_33');
     Y = L3(1,1:3500);
    plot(handles.axes1, Y, 'color', [0.7 0.7 0.7]);
case 4 %red dye is selected
    L4 = evalin('base','Trace_A6_Ladder4_0076_2017_11_06_10_33');
      R = L4(1,1:3500);
    plot(handles.axes1, R, 'color', [0.7 0.7 0.7]);
end
```

This ladder dye represents all possible outcomes for all alleles that can be discovered by using the GlobalFiler kit. A light gray color is selected to allow peaks to be seen (Figure, 5.3).

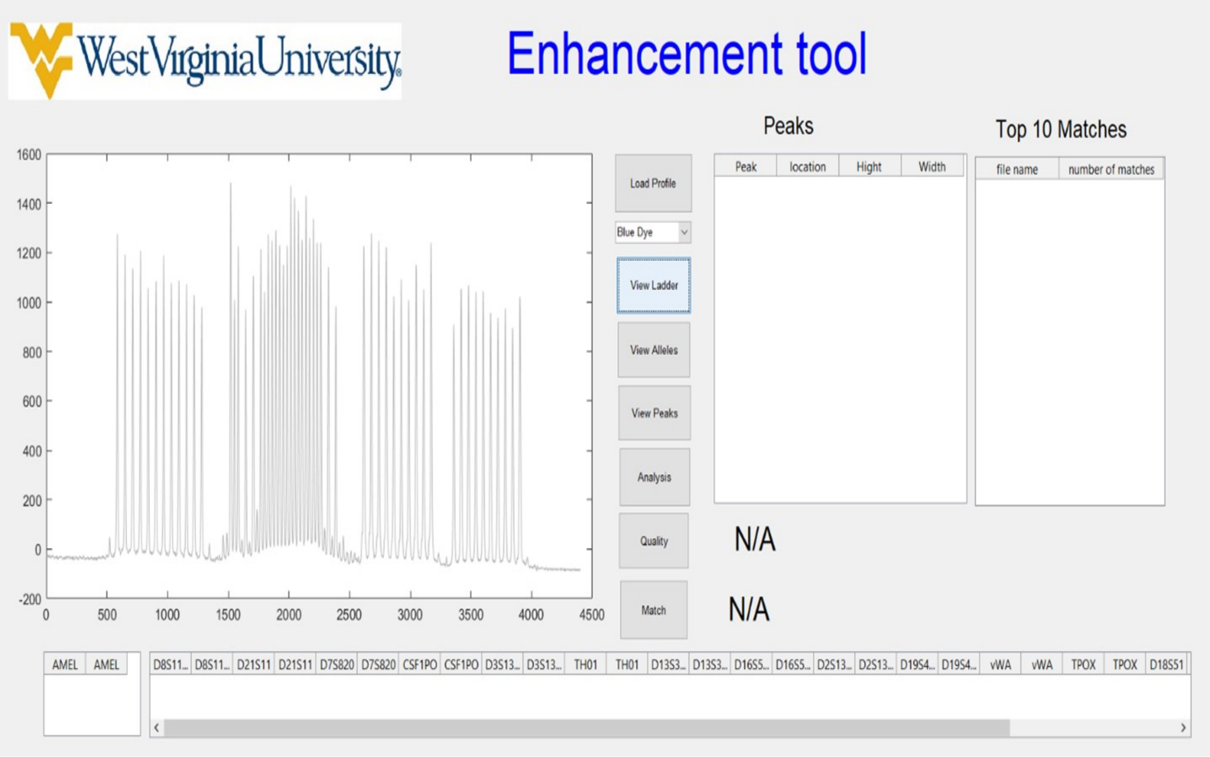


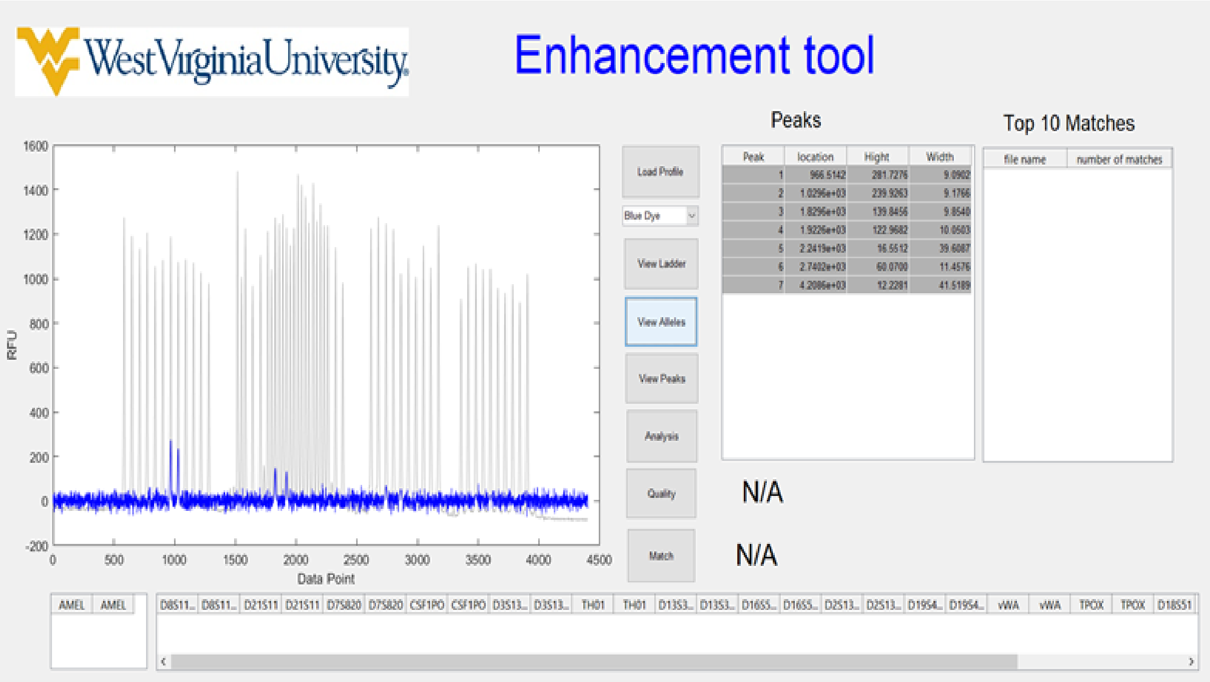
Figure 5.3: The enhancement tool after selecting the sample and choosing the blue dye to analyze the data.

- **View Alleles**

The next step is displaying the alleles for the selected samples. The user can click on the “View Alleles” button to run the script below. Using a simple “plot” instruction in MATLAB will display all the selected sample alleles with a color that fits the chosen dye (Figure 5.4): Blue for blue dye, green for green dye... etc. This script will output the peaks in the “Peaks” table on the right of the GUI. The peak finding algorithm detailed is available in Appendix C.

```
hold(handles.axes1,'on')
popup_sel_index = get(handles.popupmenu1, 'Value');
switch popup_sel_index
    case 1
```

```
        P1 = evalin('base','Final_Sample');
        plot(handles.axes1, P1(1,:), 'color', 'B');
        yf = P1(1,:);
        x = 1:length(yf);
peakgroup=5;
smoothtype=3;
WidthPoints=4; % Average number of points in half-width of peaks
SlopeThreshold=0.5*WidthPoints^-2; % Formula for estimating
value of SlopeThreshold
AmpThreshold = 11;
%AmpThreshold= mean(yf)+2*std(yf); %0.4*max(y);
smoothwidth=round(WidthPoints/2); % SmoothWidth should be
roughly equal to 1/2 the peak width (in points)
FitWidth=round(WidthPoints/2); % FitWidth should be roughly
equal to 1/2 the peak widths(in points)
P = findpeaksG(x,yf,SlopeThreshold,AmpThreshold
,smoothwidth,peakgroup,smoothtype);
```



Peak	location	Hight	Width
1	966.5142	281.7276	9.0902
2	1.0296e+03	239.9263	9.1766
3	1.8296e+03	139.8456	9.8540
4	1.9226e+03	122.9682	10.0503
5	2.2419e+03	16.5512	39.6087
6	2.7402e+03	60.0700	11.4576
7	4.2086e+03	12.2281	41.5189

file name	number of matches

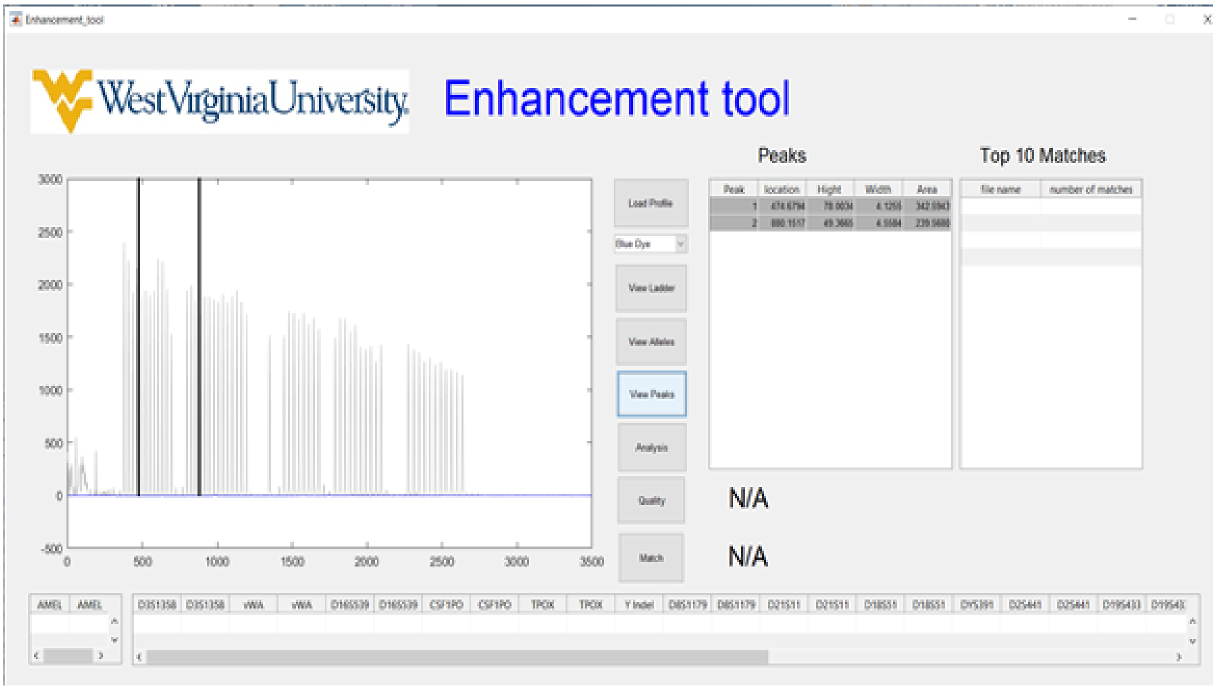
AMEL	AMEL

D8S11..	D8S11..	D21S11	D21S11	D7S820	D7S820	CSF1PO	CSF1PO	D3S13..	D3S13..	TH01	TH01	D13S3..	D13S3..	D16S5..	D16S5..	D2S13..	D2S13..	D19S4..	D19S4..	vWA	vWA	TPOX	TPOX	D18S51

Figure 5.4: The enhancement tool represents a level eight degradation sample. Once the user clicks on the View alleles button, it will show each peak's location, height, width, and size in the peaks table.

- **View Peaks**

The "View Peaks" button runs the peak finding script which analyzes each peak that might exist in the data point file (STR profile) and lists all those findings as dark gray vertical lines (Figure 5.5). This allows the user to visualize the results of the peak-finding algorithm on the chosen STR profile before analyzing those peaks in the next steps. The method is described at the end of the Theory chapter. Please see Appendix C for a full MATLAB script.



Peaks

Peak	location	Hight	Width	Area
1	474.6794	78.0034	4.1255	342.0343
2	880.1517	49.3665	4.5584	239.5680

Top 10 Matches

file name	number of matches

N/A

N/A

AMEL	AMEL

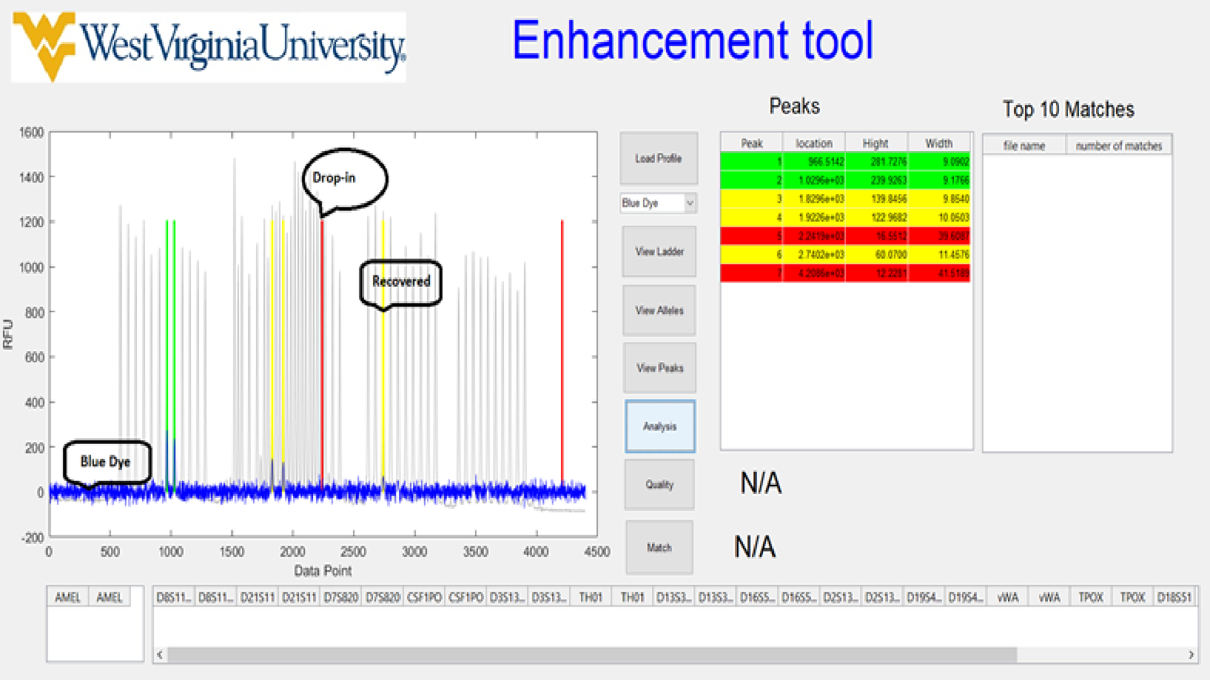
D3S1358	D3S1358	vWA	vWA	D16S539	D16S539	CSF1PO	CSF1PO	TPOX	TPOX	Y indel	D8S1179	D8S1179	D21S11	D21S11	D18S51	D18S51	DYS391	D2S441	D2S441	D19S433	D19S43

Figure 5.5: The results of all available peaks in the sample. The script will find any peak and display it in a vertical gray line to be analyzed later.

- **Analysis function**

This function gets activated through the "Analysis" button. It determines whether all discovered alleles are true or false alleles. The script for this function analyzes each peak found through the peak finding algorithm and sorts it into three main categories:

- True alleles: those alleles are above the threshold and in the acceptance, location to be in alleles in the STR sample.
- Recovered alleles were detected as true alleles using the peak finding method but under the threshold acceptance level. Those alleles' location is perfectly correlated to the allelic ladder to be considered recovered alleles. They have a yellow color vertical line. See Figure 5.6
- False allele: Each allele discovered in the used method but not in the correct location compared with the used ladder can be considered a false allele, regardless of its amplitude. See Figure 5.6 for details



Peaks

Peak	location	Hight	Width
1	966.5142	281.7276	9.0902
2	1.0296e+03	239.9263	9.1766
3	1.8296e+03	139.8456	9.8540
4	1.9226e+03	122.9682	10.0503
5	2.2419e+03	16.5512	39.6087
6	2.7402e+03	60.0700	11.4576
7	4.2086e+03	12.2281	41.5189

Top 10 Matches

file name	number of matches

Figure 5.6: The enhancement tool sorts out each allele in its category.

- **Quality function:**

The Quality function measures the quality of a signal from 0 to 1; the closer this number is to zero, the better the signal quality is. This number is calculated based on the average of four functions: 1) Average value, 2) Peaks presence, 3) standard deviation, and 4) signal-to-noise ratio. (Figure 5.7). The final score is calculated by adding all measurements and dividing the final answer by four (Equation 5.1).

$$\text{Quality Score} = \frac{\text{Average value} + \text{Peak presence} + StdDev + SNR}{4} \tag{5.1}$$

(a) **Average value:**

The average value is important to provide the mean of the analyzed signal which is an important factor in the final calculation of the signal-to-noise ratio. It is calculated using the following equation:

$$\text{Average Value} = \frac{mean - \min average}{\max average - \min average} \tag{5.2}$$

(b) **Peaks Presence:**

The Peak Presence function counts how many peaks are present in the sample which affects the quality of the signal considering the other factors. The following formula is used to calculate this factor:

$$\text{Peak presence} = \frac{\text{number of peaks} - \min}{\max - \min} \tag{5.3}$$

Where min = 4 , max = 10

(c) **Standard deviation:**

This function outputs the normalized standard deviation of the signal. First, the standard deviation of the signal is calculated, then the output signal is normalized using the following equation:

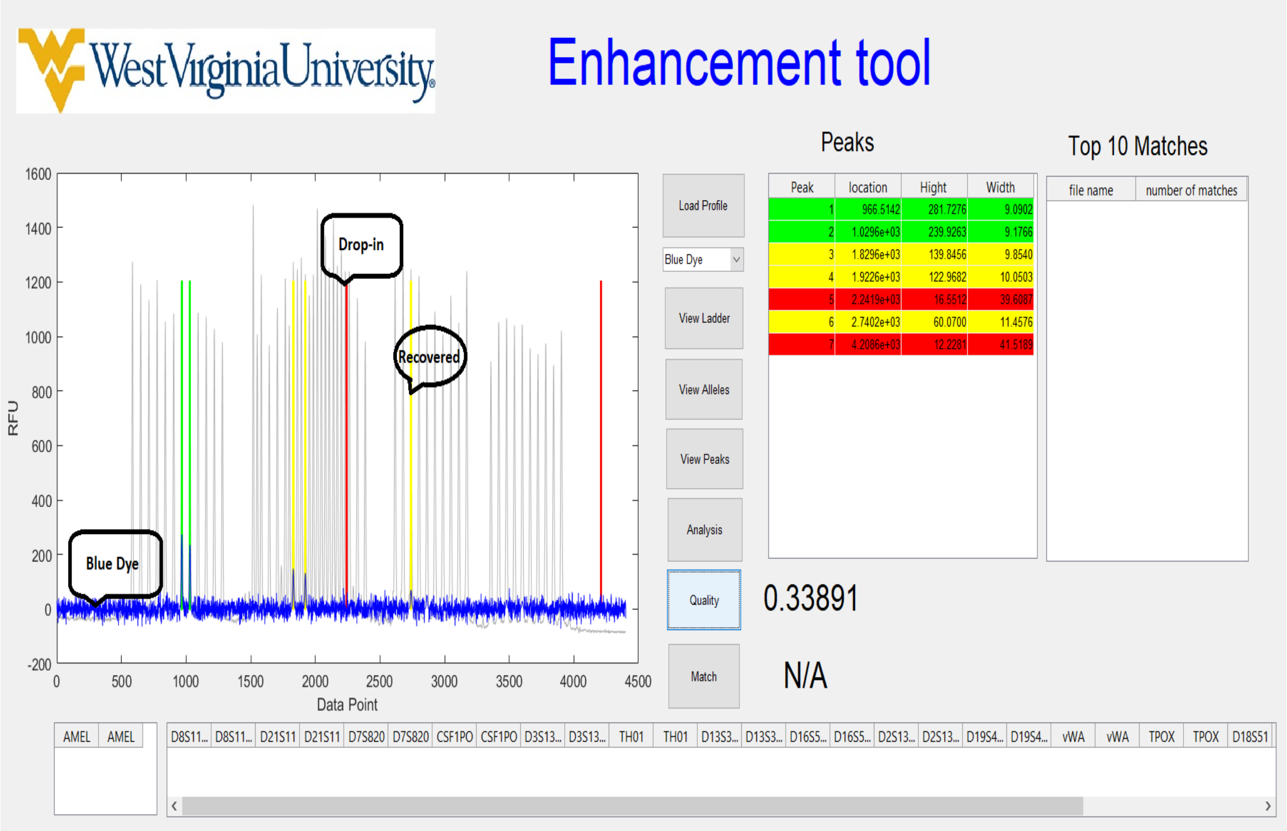
$$\text{StdDev} = \frac{StdDev - \min Std}{\max Std - \min Std} \tag{5.4}$$

Where $\min Std = 3.4167$, and $\max Std = 257.0095$

(d) **Signal to noise ratio:**

This function is a MATLAB function that can estimate and calculate the noise in the signal used, which is followed by the following formula to calculate the noise to signal ratio.

$$\text{SNR} = \frac{mean(\mathcal{F}(\varkappa))^2}{noise^2} \tag{5.5}$$



Peaks

Peak	location	Hight	Width
1	966.5142	281.7276	9.0902
2	1.0296e+03	239.9263	9.1766
3	1.8296e+03	139.8456	9.8540
4	1.9226e+03	122.9682	10.0503
5	2.2419e+03	16.5512	39.6087
6	2.7402e+03	60.0700	11.4576
7	4.2086e+03	12.2281	41.5189

Top 10 Matches

file name	number of matches

0.33891

N/A

AMEL	AMEL

D8S11...	D8S11...	D21S11	D21S11	D7S820	D7S820	CSF1PO	CSF1PO	D3S13...	D3S13...	TH01	TH01	D13S3...	D13S3...	D16S5...	D16S5...	D2S13...	D2S13...	D19S4...	D19S4...	vWA	vWA	TPOX	TPOX	D18S51

Figure 5.7: Shows the quality score for the example with blue dye.

- **Match score results:**

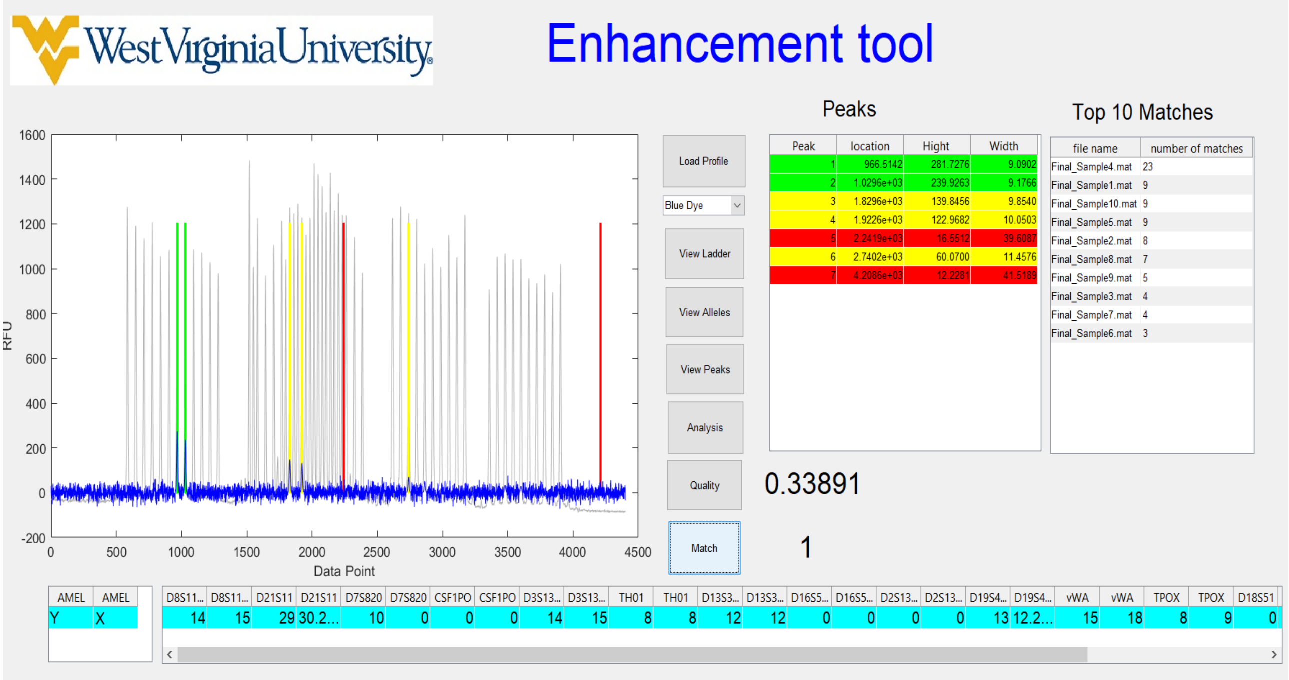
The “Match Score” algorithm compares the loaded profile alleles with each profile in the database and then represents the data in a table in ascending order (see the script below). This feature helps the user to identify which person/STR profile is the closest match to the under-examination profiles data set. It also offers the investigator a closer look at the suspect in a criminal case investigation.

```
sorted_matrix = sortrows(Top_Matches,-2);
            set(handles.uitable2,'Columnformat',({[] {'cell'}}),
            'Data', sorted_matrix)
            set(handles.text4, 'String', num2str(matches));

            if alleles (1,12) == 1
                C{1,1} = char('X');
            end
```

```
if alleles (1,12) == 2
   C{1,1} =char('Y');
end
if alleles (1,13) == 1
   C{1,2} = char('X');
end
if alleles (1,13) == 2
   C{1, 2} = char('Y');
end
alleles1 = alleles (1, 1:11);
alleles2 = alleles (1, 14:38);
alleles = [alleles1 alleles2];
columnformat = {'char'};
set(handles.uitable3 , 'Data', alleles ,
'BackgroundColor', [0 1 1]);
set(handles.uitable4 ,'Columnformat',({[]
{'char'}}), 'Data', C, 'BackgroundColor', [0 1 1])
```

The backend script compares each allele from the loaded profile with the same allele in the first profile in the database. Each matched allele adds one to the number of matches until it tests 38 alleles. It then moves on to the following profile and repeats the same testing pattern while keeping the number of matches associated with each profile. Once the entire script is completed, it orders all the profiles that exist in the target folder according to the number of matches in ascending order. It then puts the highest ten matches into the table next to the Peaks table, as shown in Figure (5.8).



Peaks

Peak	location	Hight	Width
1	966.5142	281.7276	9.0902
2	1.0296e+03	239.9263	9.1766
3	1.8296e+03	139.8456	9.8540
4	1.9226e+03	122.9682	10.0503
5	2.2419e+03	16.5512	39.6087
6	2.7402e+03	60.0700	11.4576
7	4.2086e+03	12.2281	41.5189

Top 10 Matches

file name	number of matches
Final_Sample4.mat	23
Final_Sample1.mat	9
Final_Sample10.mat	9
Final_Sample5.mat	9
Final_Sample2.mat	8
Final_Sample8.mat	7
Final_Sample9.mat	5
Final_Sample3.mat	4
Final_Sample7.mat	4
Final_Sample6.mat	3

0.33891

1

AMEL	AMEL
Y	X

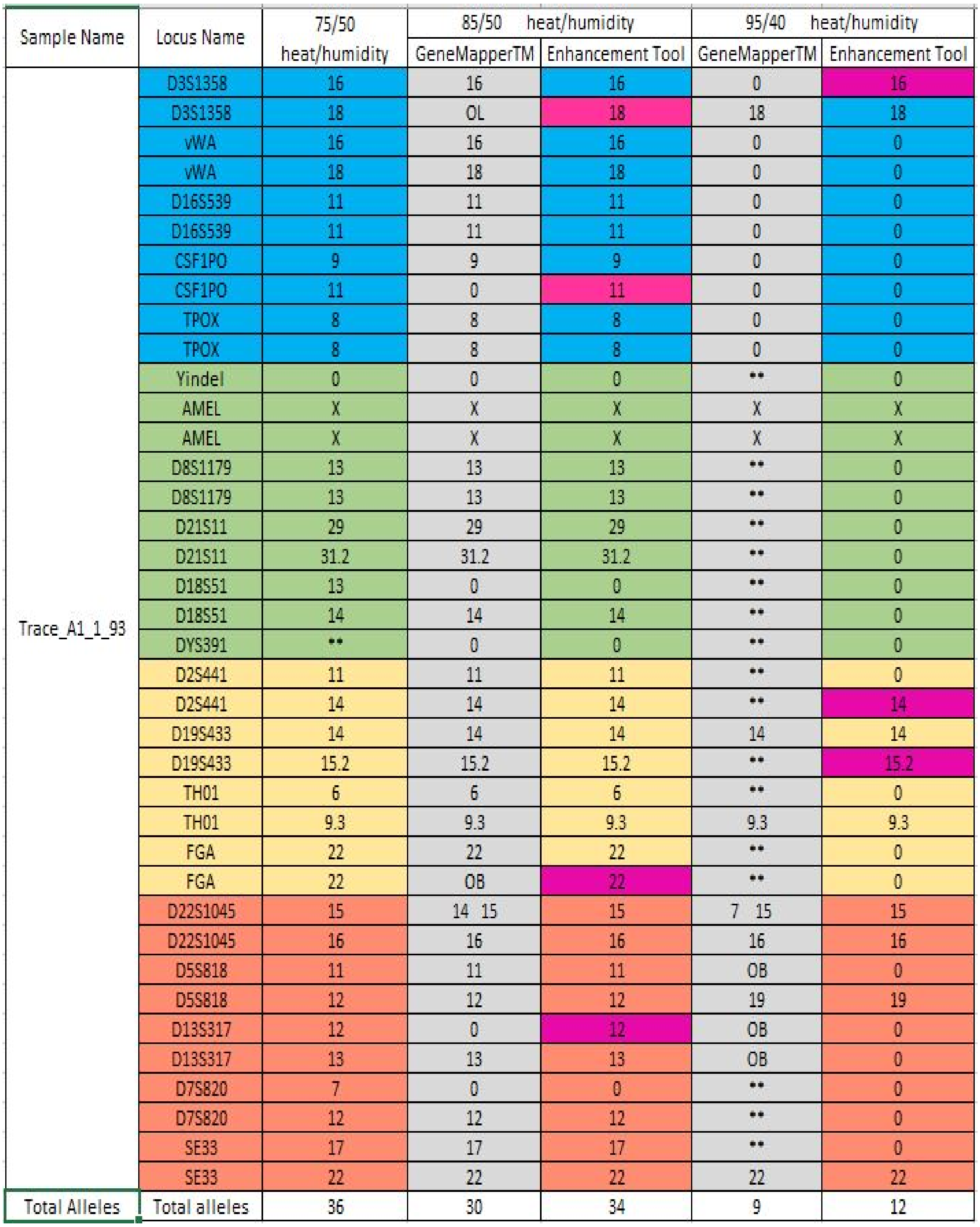
D8S11...	D8S11...	D21S11	D21S11	D7S820	D7S820	CSF1PO	CSF1PO	D3S13...	D3S13...	TH01	TH01	D13S3...	D13S3...	D16S5...	D16S5...	D2S13...	D2S13...	D19S4...	D19S4...	vWA	vWA	TPOX	TPOX	D18S51
14	15	29	30.2...	10	0	0	0	14	15	8	8	12	12	0	0	0	0	13	12.2...	15	18	8	9	0

Figure 5.8: Shows the Top 10 matches profiles comparing to the under-study profile

5.3 Analyzing real samples

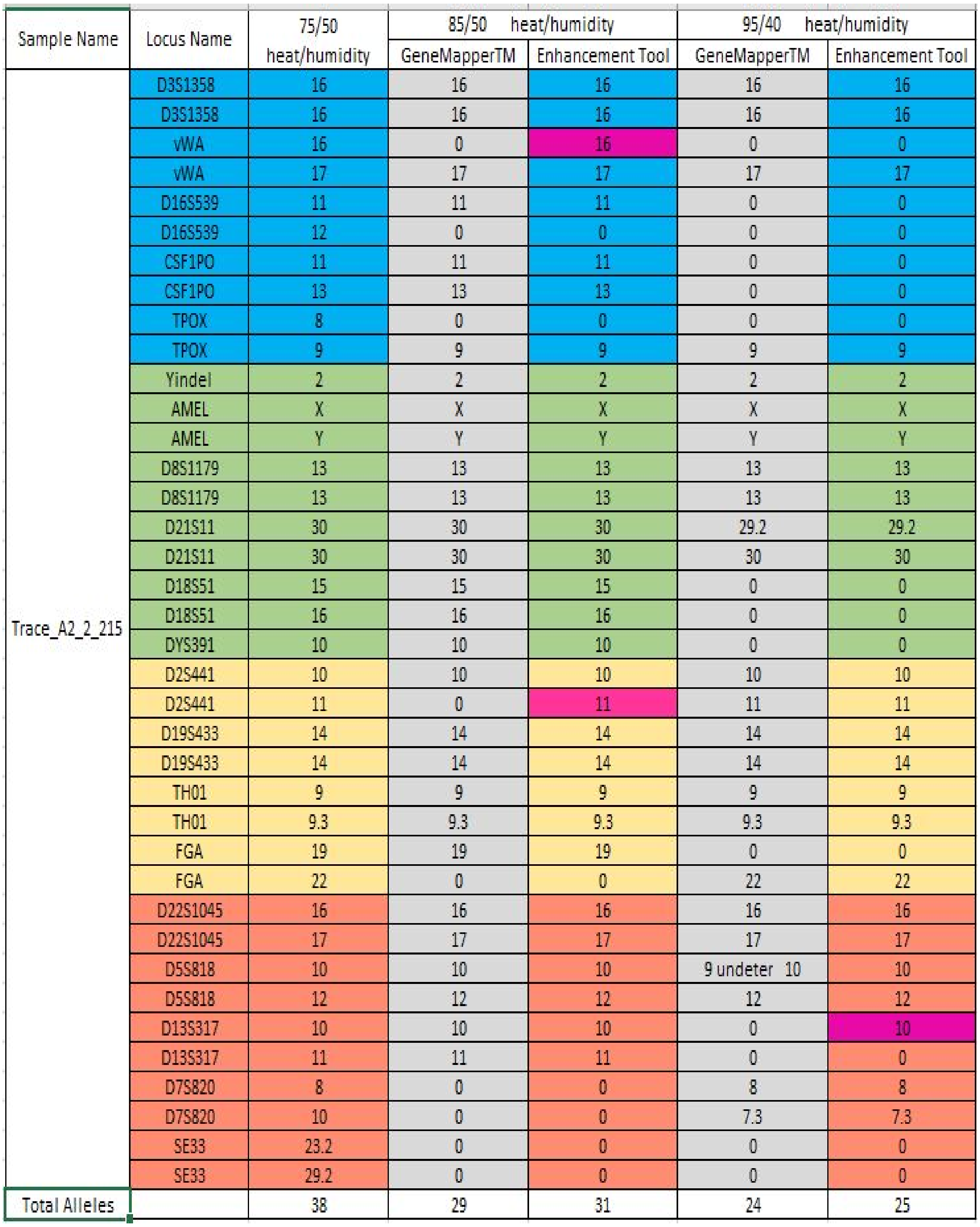
The enhancement tool was used to enhance and examine 42 profiles from seven participants in this study. The samples were enriched individually, and the results were recorded to demonstrate the differences between them. All the 75/50 heat/humidity degradation profiles that had all the correct alleles were called without any dropouts. In the case of samples with the 85/50 heat/humidity combination, a different dropout/recovery ratio was noticed between the *GeneMapper*^{TM} ID-X Software, *RapidHIT*^{TM} 200, and the enhancement tool developed. When using the *GeneMapper*^{TM} ID-X Software, 95/40 heat/humidity combinations had a higher dropout rate. *GeneMapper*^{TM} produced better results that can be immediately accessed compared with *GeneMapper*^{TM}. Results from the instrument lab were taken from a study [79] made to evaluate the *RapidHIT*^{TM} 200 on degraded biological samples. The study [79] used *GeneMapper*^{TM} as a reference to evaluate the results of *RapidHIT*^{TM} [79]. In the current study, the FSA samples were converted into .MAT files that were run through the enhancement tool. The findings were recorded into an Excel sheet to test the credibility of the method on real data samples. Figures 5.9 to 5.15 summarize the outcome of the enhancement tool using the Gaussian alleles finding method. This would improve the alleles calling ratio by more than 20 % when dealing with degraded samples.

The strength of signal processing in finding and recovering dropped-out alleles when using a mathematical method empowers and enhances the signal to pick up a degraded allele that would not be called using other software/methods.



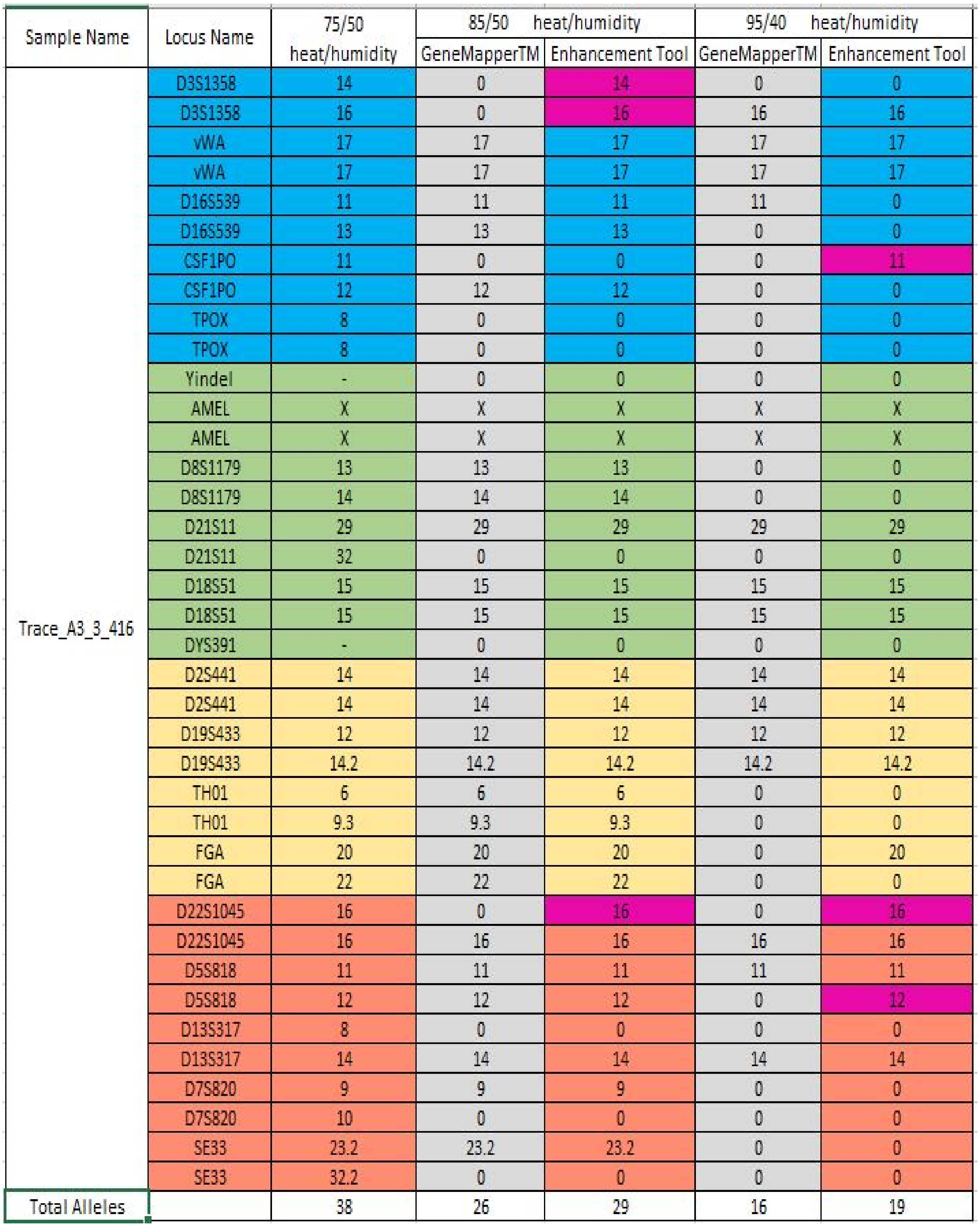
Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A1_1_93	D3S1358	16	16	16	0	16
	D3S1358	18	OL	18	18	18
	vWA	16	16	16	0	0
	vWA	18	18	18	0	0
	D16S539	11	11	11	0	0
	D16S539	11	11	11	0	0
	CSF1PO	9	9	9	0	0
	CSF1PO	11	0	11	0	0
	TPOX	8	8	8	0	0
	TPOX	8	8	8	0	0
	Yindel	0	0	0	**	0
	AMEL	X	X	X	X	X
	AMEL	X	X	X	X	X
	D8S1179	13	13	13	**	0
	D8S1179	13	13	13	**	0
	D21S11	29	29	29	**	0
	D21S11	31.2	31.2	31.2	**	0
	D18S51	13	0	0	**	0
	D18S51	14	14	14	**	0
	DYS391	**	0	0	**	0
	D2S441	11	11	11	**	0
	D2S441	14	14	14	**	14
	D19S433	14	14	14	14	14
	D19S433	15.2	15.2	15.2	**	15.2
	TH01	6	6	6	**	0
	TH01	9.3	9.3	9.3	9.3	9.3
	FGA	22	22	22	**	0
	FGA	22	OB	22	**	0
	D22S1045	15	14 15	15	7 15	15
	D22S1045	16	16	16	16	16
	D5S818	11	11	11	OB	0
	D5S818	12	12	12	19	19
	D13S317	12	0	12	OB	0
	D13S317	13	13	13	OB	0
	D7S820	7	0	0	**	0
	D7S820	12	12	12	**	0
	SE33	17	17	17	**	0
	SE33	22	22	22	22	22
Total Alleles	Total alleles	36	30	34	9	12

Figure 5.9: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



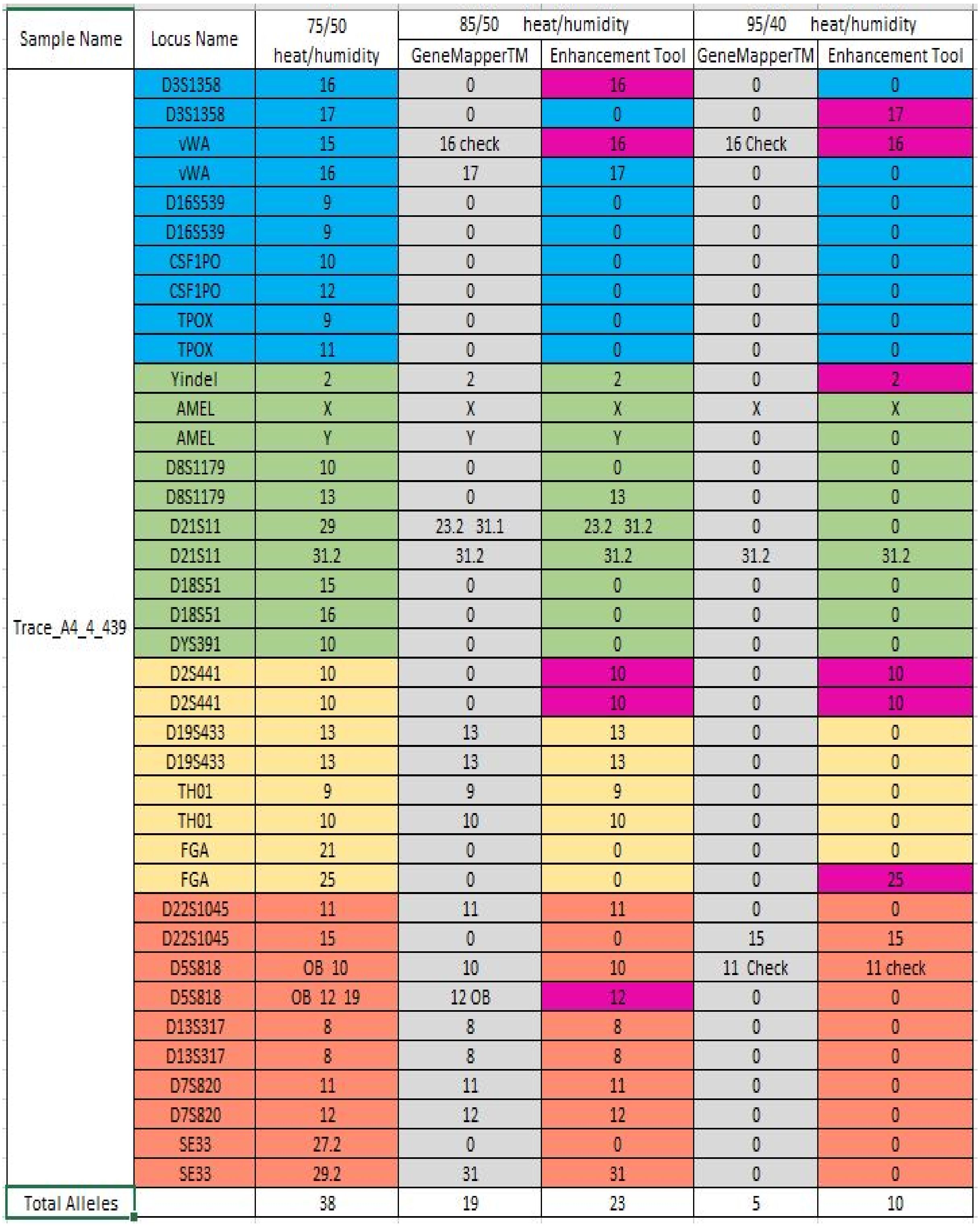
Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A2_2_215	D3S1358	16	16	16	16	16
	D3S1358	16	16	16	16	16
	vWA	16	0	16	0	0
	vWA	17	17	17	17	17
	D16S539	11	11	11	0	0
	D16S539	12	0	0	0	0
	CSF1PO	11	11	11	0	0
	CSF1PO	13	13	13	0	0
	TPOX	8	0	0	0	0
	TPOX	9	9	9	9	9
	Yindel	2	2	2	2	2
	AMEL	X	X	X	X	X
	AMEL	Y	Y	Y	Y	Y
	D8S1179	13	13	13	13	13
	D8S1179	13	13	13	13	13
	D21S11	30	30	30	29.2	29.2
	D21S11	30	30	30	30	30
	D18S51	15	15	15	0	0
	D18S51	16	16	16	0	0
	DYS391	10	10	10	0	0
	D2S441	10	10	10	10	10
	D2S441	11	0	11	11	11
	D19S433	14	14	14	14	14
	D19S433	14	14	14	14	14
	TH01	9	9	9	9	9
	TH01	9.3	9.3	9.3	9.3	9.3
	FGA	19	19	19	0	0
	FGA	22	0	0	22	22
	D22S1045	16	16	16	16	16
	D22S1045	17	17	17	17	17
	D5S818	10	10	10	9 undeter 10	10
	D5S818	12	12	12	12	12
	D13S317	10	10	10	0	10
	D13S317	11	11	11	0	0
	D7S820	8	0	0	8	8
	D7S820	10	0	0	7.3	7.3
	SE33	23.2	0	0	0	0
	SE33	29.2	0	0	0	0
Total Alleles		38	29	31	24	25

Figure 5.10: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



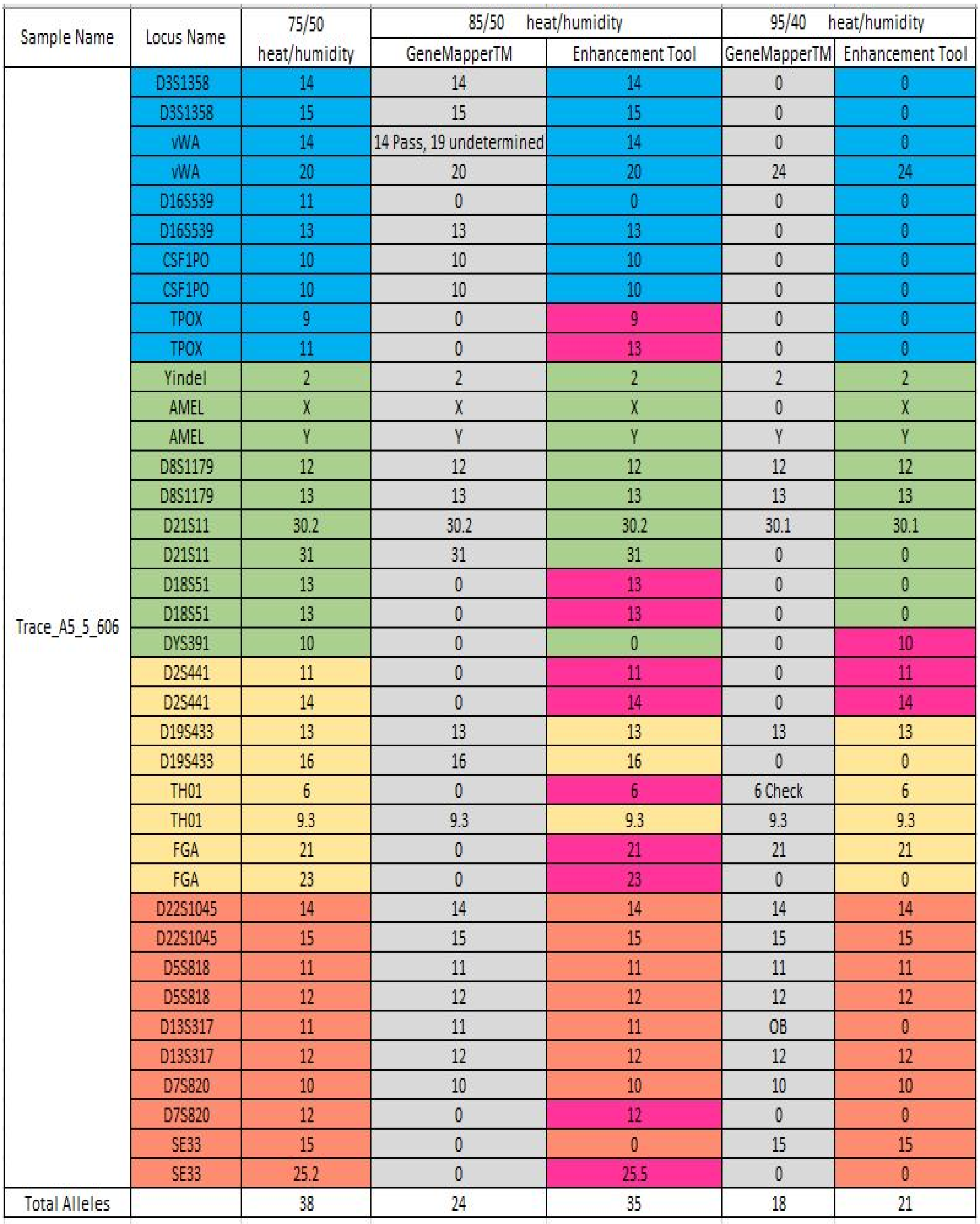
Sample Name	Locus Name	75/50	85/50 heat/humidity		95/40 heat/humidity	
		heat/humidity	GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A3_3_416	D3S1358	14	0	14	0	0
	D3S1358	16	0	16	16	16
	vWA	17	17	17	17	17
	vWA	17	17	17	17	17
	D16S539	11	11	11	11	0
	D16S539	13	13	13	0	0
	CSF1PO	11	0	0	0	11
	CSF1PO	12	12	12	0	0
	TPOX	8	0	0	0	0
	TPOX	8	0	0	0	0
	Yindel	-	0	0	0	0
	AMEL	X	X	X	X	X
	AMEL	X	X	X	X	X
	D8S1179	13	13	13	0	0
	D8S1179	14	14	14	0	0
	D21S11	29	29	29	29	29
	D21S11	32	0	0	0	0
	D18S51	15	15	15	15	15
	D18S51	15	15	15	15	15
	DYS391	-	0	0	0	0
	D2S441	14	14	14	14	14
	D2S441	14	14	14	14	14
	D19S433	12	12	12	12	12
	D19S433	14.2	14.2	14.2	14.2	14.2
	TH01	6	6	6	0	0
	TH01	9.3	9.3	9.3	0	0
	FGA	20	20	20	0	20
	FGA	22	22	22	0	0
	D22S1045	16	0	16	0	16
	D22S1045	16	16	16	16	16
	D5S818	11	11	11	11	11
	D5S818	12	12	12	0	12
	D13S317	8	0	0	0	0
	D13S317	14	14	14	14	14
	D7S820	9	9	9	0	0
	D7S820	10	0	0	0	0
	SE33	23.2	23.2	23.2	0	0
	SE33	32.2	0	0	0	0
Total Alleles		38	26	29	16	19

Figure 5.11: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



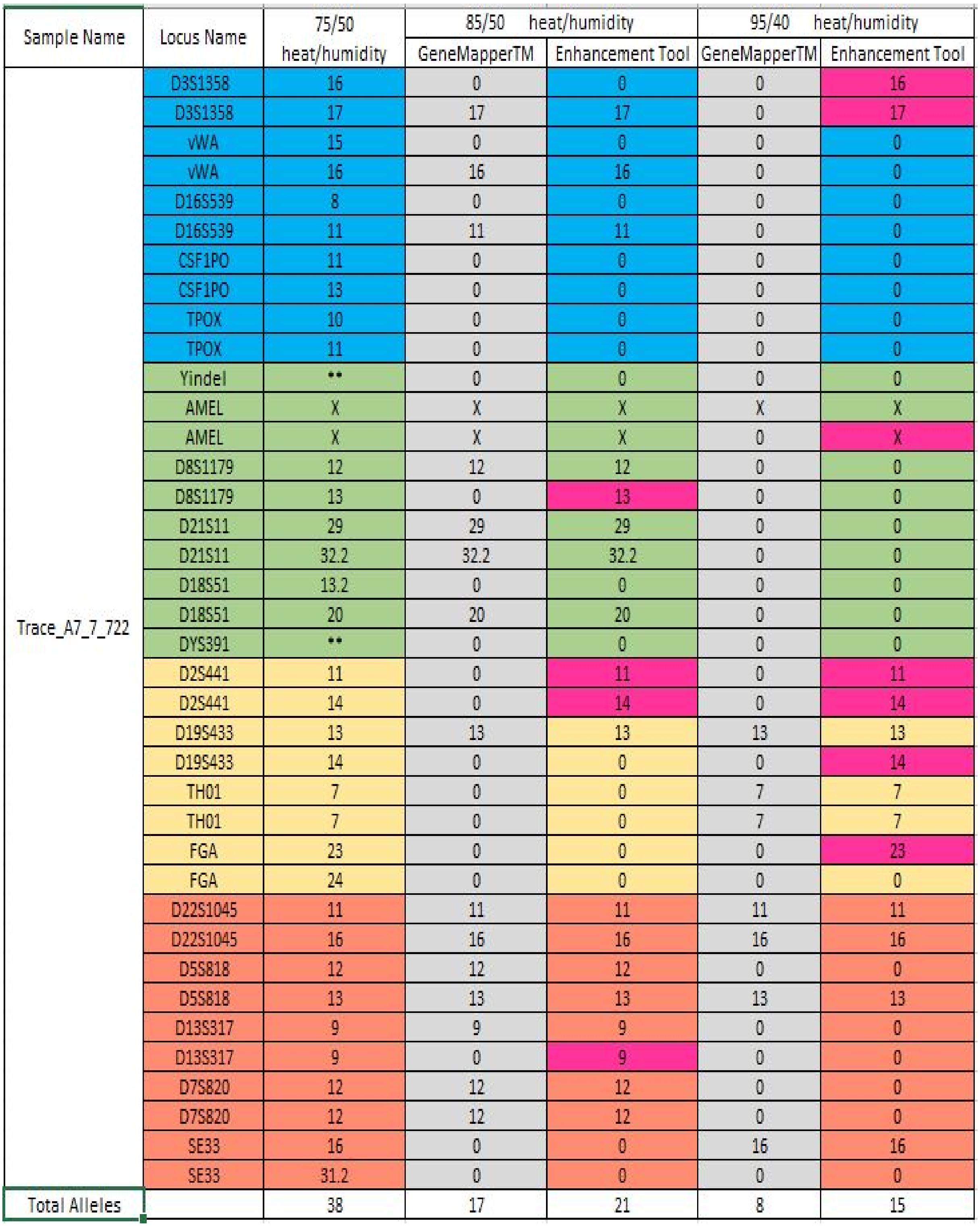
Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A4_4_439	D3S1358	16	0	16	0	0
	D3S1358	17	0	0	0	17
	vWA	15	16 check	16	16 Check	16
	vWA	16	17	17	0	0
	D16S539	9	0	0	0	0
	D16S539	9	0	0	0	0
	CSF1PO	10	0	0	0	0
	CSF1PO	12	0	0	0	0
	TPOX	9	0	0	0	0
	TPOX	11	0	0	0	0
	Yindel	2	2	2	0	2
	AMEL	X	X	X	X	X
	AMEL	Y	Y	Y	0	0
	D8S1179	10	0	0	0	0
	D8S1179	13	0	13	0	0
	D21S11	29	23.2 31.1	23.2 31.2	0	0
	D21S11	31.2	31.2	31.2	31.2	31.2
	D18S51	15	0	0	0	0
	D18S51	16	0	0	0	0
	DYS391	10	0	0	0	0
	D2S441	10	0	10	0	10
	D2S441	10	0	10	0	10
	D19S433	13	13	13	0	0
	D19S433	13	13	13	0	0
	TH01	9	9	9	0	0
	TH01	10	10	10	0	0
	FGA	21	0	0	0	0
	FGA	25	0	0	0	25
	D22S1045	11	11	11	0	0
	D22S1045	15	0	0	15	15
	D5S818	OB 10	10	10	11 Check	11 check
	D5S818	OB 12 19	12 OB	12	0	0
	D13S317	8	8	8	0	0
	D13S317	8	8	8	0	0
	D7S820	11	11	11	0	0
	D7S820	12	12	12	0	0
	SE33	27.2	0	0	0	0
	SE33	29.2	31	31	0	0
Total Alleles		38	19	23	5	10

Figure 5.12: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



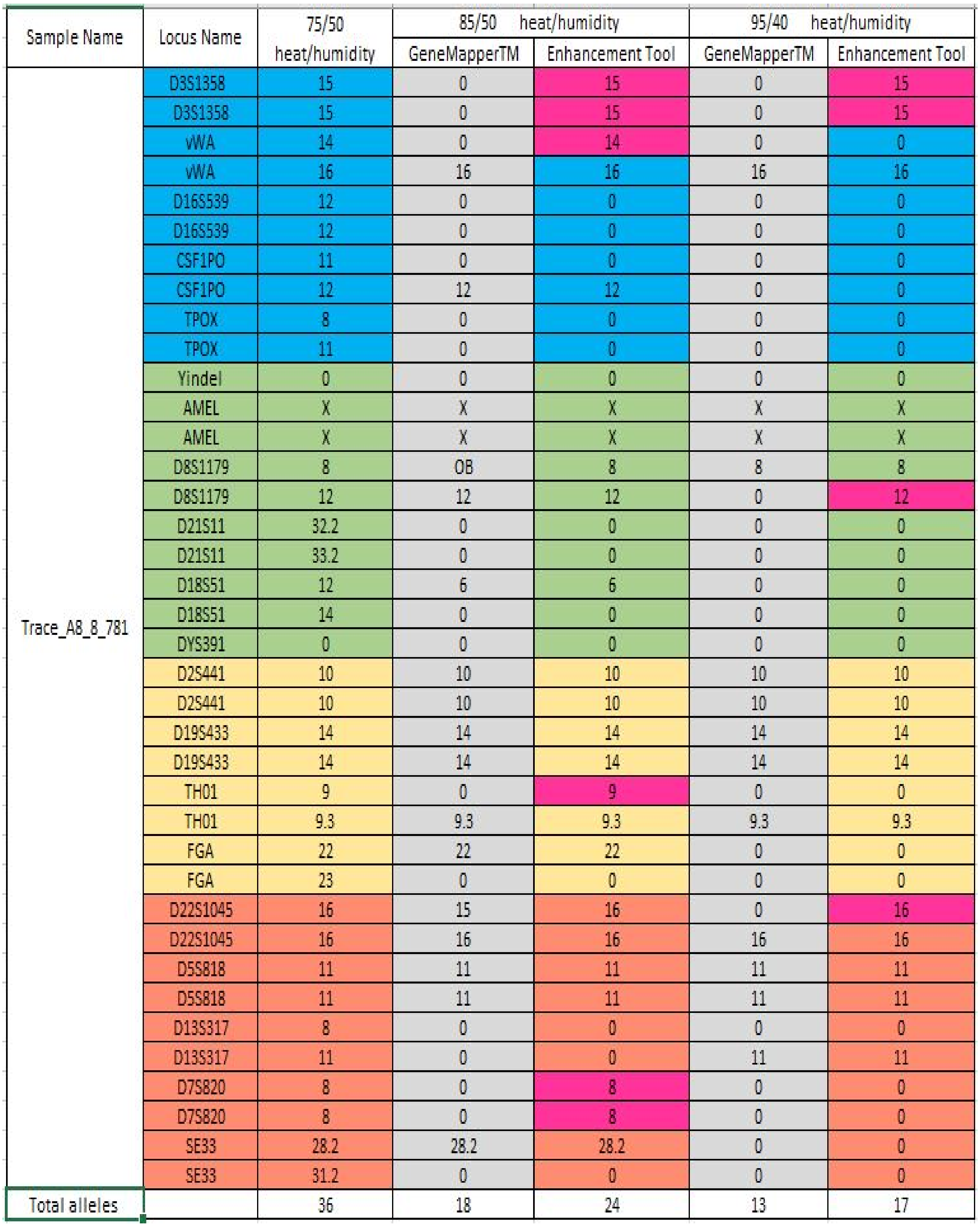
Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A5_5_606	D3S1358	14	14	14	0	0
	D3S1358	15	15	15	0	0
	vWA	14	14 Pass, 19 undetermined	14	0	0
	vWA	20	20	20	24	24
	D16S539	11	0	0	0	0
	D16S539	13	13	13	0	0
	CSF1PO	10	10	10	0	0
	CSF1PO	10	10	10	0	0
	TPOX	9	0	9	0	0
	TPOX	11	0	13	0	0
	Yindel	2	2	2	2	2
	AMEL	X	X	X	0	X
	AMEL	Y	Y	Y	Y	Y
	D8S1179	12	12	12	12	12
	D8S1179	13	13	13	13	13
	D21S11	30.2	30.2	30.2	30.1	30.1
	D21S11	31	31	31	0	0
	D18S51	13	0	13	0	0
	D18S51	13	0	13	0	0
	DYS391	10	0	0	0	10
	D2S441	11	0	11	0	11
	D2S441	14	0	14	0	14
	D19S433	13	13	13	13	13
	D19S433	16	16	16	0	0
	TH01	6	0	6	6 Check	6
	TH01	9.3	9.3	9.3	9.3	9.3
	FGA	21	0	21	21	21
	FGA	23	0	23	0	0
	D22S1045	14	14	14	14	14
	D22S1045	15	15	15	15	15
	D5S818	11	11	11	11	11
	D5S818	12	12	12	12	12
	D13S317	11	11	11	OB	0
	D13S317	12	12	12	12	12
	D7S820	10	10	10	10	10
	D7S820	12	0	12	0	0
	SE33	15	0	0	15	15
	SE33	25.2	0	25.5	0	0
Total Alleles		38	24	35	18	21

Figure 5.13: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A7_7_722	D3S1358	16	0	0	0	16
	D3S1358	17	17	17	0	17
	vWA	15	0	0	0	0
	vWA	16	16	16	0	0
	D16S539	8	0	0	0	0
	D16S539	11	11	11	0	0
	CSF1PO	11	0	0	0	0
	CSF1PO	13	0	0	0	0
	TPOX	10	0	0	0	0
	TPOX	11	0	0	0	0
	Yindel	**	0	0	0	0
	AMEL	X	X	X	X	X
	AMEL	X	X	X	0	X
	D8S1179	12	12	12	0	0
	D8S1179	13	0	13	0	0
	D21S11	29	29	29	0	0
	D21S11	32.2	32.2	32.2	0	0
	D18S51	13.2	0	0	0	0
	D18S51	20	20	20	0	0
	DYS391	**	0	0	0	0
	D2S441	11	0	11	0	11
	D2S441	14	0	14	0	14
	D19S433	13	13	13	13	13
	D19S433	14	0	0	0	14
	TH01	7	0	0	7	7
	TH01	7	0	0	7	7
	FGA	23	0	0	0	23
	FGA	24	0	0	0	0
	D22S1045	11	11	11	11	11
	D22S1045	16	16	16	16	16
	D5S818	12	12	12	0	0
	D5S818	13	13	13	13	13
	D13S317	9	9	9	0	0
	D13S317	9	0	9	0	0
	D7S820	12	12	12	0	0
	D7S820	12	12	12	0	0
	SE33	16	0	0	16	16
	SE33	31.2	0	0	0	0
Total Alleles		38	17	21	8	15

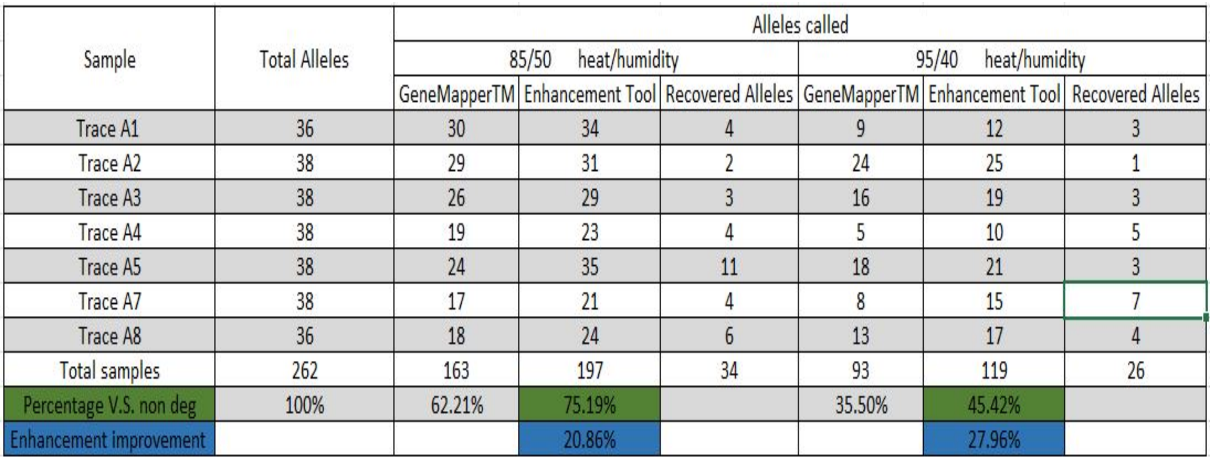
Figure 5.14: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels



Sample Name	Locus Name	75/50 heat/humidity	85/50 heat/humidity		95/40 heat/humidity	
			GeneMapperTM	Enhancement Tool	GeneMapperTM	Enhancement Tool
Trace_A8_8_781	D3S1358	15	0	15	0	15
	D3S1358	15	0	15	0	15
	vWA	14	0	14	0	0
	vWA	16	16	16	16	16
	D16S539	12	0	0	0	0
	D16S539	12	0	0	0	0
	CSF1PO	11	0	0	0	0
	CSF1PO	12	12	12	0	0
	TPOX	8	0	0	0	0
	TPOX	11	0	0	0	0
	Yindel	0	0	0	0	0
	AMEL	X	X	X	X	X
	AMEL	X	X	X	X	X
	D8S1179	8	OB	8	8	8
	D8S1179	12	12	12	0	12
	D21S11	32.2	0	0	0	0
	D21S11	33.2	0	0	0	0
	D18S51	12	6	6	0	0
	D18S51	14	0	0	0	0
	DYS391	0	0	0	0	0
	D2S441	10	10	10	10	10
	D2S441	10	10	10	10	10
	D19S433	14	14	14	14	14
	D19S433	14	14	14	14	14
	TH01	9	0	9	0	0
	TH01	9.3	9.3	9.3	9.3	9.3
	FGA	22	22	22	0	0
	FGA	23	0	0	0	0
	D22S1045	16	15	16	0	16
	D22S1045	16	16	16	16	16
	D5S818	11	11	11	11	11
	D5S818	11	11	11	11	11
	D13S317	8	0	0	0	0
	D13S317	11	0	0	11	11
	D7S820	8	0	8	0	0
	D7S820	8	0	8	0	0
	SE33	28.2	28.2	28.2	0	0
	SE33	31.2	0	0	0	0
Total alleles		36	18	24	13	17

Figure 5.15: The enhancement tool's actual peaks detection capability versus GeneMapperTM for one sample with different degradation levels

The number of called alleles was improved substantially (see Figure 5.16), leading to the recall of 34 more alleles using the enhancement tool in 84/50 heats/humidity samples and 28 alleles in 95/40 heats/humidity samples. Better results by 20.86% were achieved when working with 85/50 heat humidity degradation and 27.96% for 95/40 heat humidity degradation. Every sample was tested and enhanced separately to ensure and record those results. Although the sample was degraded, this study demonstrates that applying signal processing and mathematical techniques can improve the outcomes of a sample by more than 20%. Using Gaussian or Lorentzian functions can increase the number of alleles by applying mathematical calculations to detect the peaks.



Sample	Total Alleles	Alleles called					
		85/50 heat/humidity			95/40 heat/humidity		
		GeneMapperTM	Enhancement Tool	Recovered Alleles	GeneMapperTM	Enhancement Tool	Recovered Alleles
Trace A1	36	30	34	4	9	12	3
Trace A2	38	29	31	2	24	25	1
Trace A3	38	26	29	3	16	19	3
Trace A4	38	19	23	4	5	10	5
Trace A5	38	24	35	11	18	21	3
Trace A7	38	17	21	4	8	15	7
Trace A8	36	18	24	6	13	17	4
Total samples	262	163	197	34	93	119	26
Percentage V.S. non deg	100%	62.21%	75.19%		35.50%	45.42%	
Enhancement improvement			20.86%			27.96%	

Figure 5.16: shows Final result after using the enhancement tools on real data samples

Chapter 6

Conclusion and Future work

This chapter summarizes the value of this research, establishes the foundation for all future research, and suggests some directions for continuous engineering improvements in signal processing in DNA typing analysis.

6.1 Conclusion

A new method of generating STR profiles has been developed to provide scientists and forensic lab experts with unlimited STR profiles based on genetic population. A MATLAB code and GUI are implemented to provide unlimited STR profiles based on user recommendations. Those profiles are based on mixed population genetics where a higher variation is noticed and probabilities in all races in FBI population genetic release of 2017.

A degradation feature is added to allow the user to generate a degradation sample matching the non-degraded ones. The user also can choose the level of degradation between (1) and (20), which depends on what he would like to do his research on. The degradation is based on an exponential function being multiply to the original STR profile where the small base pair have a smaller degradation comparing the higher base pair loci.

Dr.Dawson research engineering group developed a peak finding method based on mathematical solution of first and second derivative of the DNA STR profile signal. This signal processing method will allow detection for a high number of alleles buried in the noise. By using his method, any small shaped signal can be detected which can be classified as an allele. Then matching this peak with allelic ladder of GlobalFiler to see if it has the location to be considered a true allele.

To maximize the benefits coming from this approach, and to provide a better way to represent the data. A GUI is developed which can load the STR profile for any degraded/non degraded artificial profile and run the main three comparison methods to better visualize and study the data. (1) Analysis, (2) Quality score, (3) match score.

Analysis feature allows the user to provide analysis on each discovered peak based on peak finding algorithm then match those peaks with the allelic ladder and label them with green, yellow and red color depends on the peak. Quality score is a combination of four calculations to determine the quality of the analysis after calculating the amount of noise built into it during amplification and detection steps in the lab. Match score is way to determine the highest STR profile match to the one is getting analyzed. Using this method, the code can run the database and order the entire data set from the highest to the lowest match. Which's considered one of the quickest ways to scan the sample.

6.2 Future work

This research is considered the spark of signal processing efforts in DNA typing analysis. It can be continued and improved in varies ways depends on the end goal we're trying to achieve. Firstly, adding a feature to the STR generator can allow the user to choose between different races to generate. Secondly, it can be added to a computer/mobile application where it can take a sample and rabidly test the individual to see if he/she is wanted for justice before entering the borders. Thirdly, you can add the ability to modify the dataset to include more accurate probabilities when new population genetics gets released to generate more accurate data and up to date to each race. Fourthly, you can add the ability to generate STR profiles to a different chemical kit with a different allelic ladder. Finally, you can provide a customized the enhancement tool to detect mixture profiles and do analysis on given conditions.

References

[1] Buccal Swab sample technique. Available at: $< https : //dnacenter.com/ >$.

[2] K. R. Hearn, Laboratory Exercise DNA Extraction Techniques. 2010.

[3] UV absorbance DNA quantitation. Available at: $< https : //www.bmglabtech.com/uv - absorbance - dna - quantitation/ >$.

[4] Various multiplex PCR systems. Available at: $< https : //www.biorad.com/featured/en/multiplex - pcr.html >$.

[5] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[6] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[7] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[8] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[9] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[10] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[11] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[12] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[13] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[14] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[15] A complete profile. Available at: $< https : //www.strbase.nist.gov >$.

[16] Differences between homozygous and heterozygous loci. Available at: $< https : //www.genome.gov/geneticsglossary/ >$.

[17] shows a green dye for individual missing information at two of its loci. Available at: < *https* : *//www.genome.gov/geneticsglossary/* >.

[18] J. Butler, Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers. 2005.

[19] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[20] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[21] RapidHit 200 DNA profiling system. Available at: < *https* : *//www.fbi.gov* >.

[22] ANDE DNA profiling system. Available at: < *https* : *//www.ande.com* >.

[23] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[24] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology. 2011.

[25] Gaussian fit method. Available at: < *https* : *//en* : *wikipedia* : *org/wiki/Gaussianfunction* >.

[26] J. M. Butler, Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press, 2005.

[27] J. M. Butler, Advance topics in forensic DNA typing methodology. 2011.

[28] NANOPORE Technology real-time devices for DNA and RNA sequencing Available at: < *https* : *//nanoporetech.com/products/minion* >.

[29] J. M. Butler, Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press, 2005.

[30] P. de Knijff, Son, give up your gun: presenting Y-STR results in court, vol. 6. 2003.

[31] A. Caliebe, A. Jochens, S. Willuweit, L. Roewer, and M. Krawczak, No shortcut solution to the problem of Y-STR match probability calculation, vol. 15. Elsevier, 2015.

[32] W. P. . K. M., Less is more - Length reduction of STR amplicons using redesigned primers. international Journal of Legal Medicine, 2001.

[33] J. M. Butler, Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press, 2005.

[34] e. a. . Whitaker, J. P., Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in high degraded samples. 1995.

[35] e. a. Takahashi, M., Evaluation of five polymorphic microsatellite markers for typing DNA from decomposed human tissues. 1997.

[36] e. a. . Schneider, P.M., STR analysis of artificially degraded DNA - results of a collabrorative European exercise. 2004.

[37] J. M. Butler, Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press, 2005.

[38] C. J. T. C. B. J. and W. B.S., Characterization of new miniSTR loci to aid analysis of degraded DNA. Journal of Forensic Sciences, 44, 1999.

[39] P. Gill. Forensic Science International, 91, 1997.

[40] G. P. S. R. P. R. C. T. W. J. and B. J. Forensic Science International, 91, (1998a).

[41] G. P. S. B. C. T. W. J. U. A. and B. J.S., Proceedings of the Ninth International Symposium on Human Identification,. Madison, Wisconsin: Promega Corporation., (1998b).

[42] T. Y. F. I. P. V. L. soto M. Farfan M.J. Carracedo A. and S. P. Forensic Science International, 2003.

[43] e. a. Findlay, I., DNA fingerprinting from single cells. Nature, 389, 1997.

[44] B. J. M. . H. C. R., Scientific issues with analysis of low amounts of DNA Available at: *http : //www.promega.com/profiles/*1301/1301$_0$2*.html*. Profiles in DNA, 13(1), 2010.

[45] *http : //www.forensicsciencesimplified.org/dna/how.html*.

[46] s. a. Hopwood, A.J., Integrated microfluidic system for rapid forensic DNA analysis : Sample collection to DNA profile. 2010.

[47] s. a. Hurth, C., An automated instrument for human STR identification: Design, characterization, and experimental valiadtion. 2010.

[48] e. a. Tsukada, K., Multiplex short tandem repeat typing in degraded samples using newly designed primers for the TH01, TPOX, CSF1PO and vWA Loci. Legal Medicine, 4, 2002.

[49] e. a. Butler, J.M., The development of reduced size STR Amplicons as tools for analysis of degraded DNA. Journal of Forensic Science, 48(5), 2003.

[50] e. a. Mulero, J.J., Development and validation of the AmpFlSTR MiniFiler PCR Amplification Kit: a min- iSTR multiplex for the analysis of degraded and/ or PCR inhibited DNA. Journal of Forensic Science, 53, 2008.

[51] e. a. Hill, C.CR., Concordance between the AmpflSTR MiniFiler and AmpFlSTR Identifiler PCS amplification kits in the Kuwaiti population. Journal of Forensic Science, 52(4), 2007.

[52] e. a. Drabek, J., A study of the effects of degradation and template concentration on the efficiency of the STR miniplex primer sets. Journal of Forensic Science, 49(4), 2004.

[53] e. a. Chung, D.T., A study on the effects of degradation and template concentration on the efficiency of the STR miniplex primer sets. Journal of Forensic Sciences, 49, 2004.

[54] e. a. Schumm, J.W., Robust STR multiplexes for challenging casework samples. Progress in Forensic Genetics, 10, ICS 1261, 2004.

[55] e. a. Hellmann, A., STR typing of human telogen hairs - a new approach. International Journal of Legal Medicine, 2001.

[56] a. a. Muller, K., Improved STR typing of telogen hair root and hair shaft DNA. Electrophoresis, 2007.

[57] e. a. Opel, K. L., The applicaton of miniplex primer sets in the analysis of degraded DNA from human skeletal remains. Journal of Forensic Science, 2006.

[58] e. a. Dixon, L.A., Analysis of artificially degraded DNA usin STRs and SNPs - results of collaborative Eropean (EDNAP) exercise. Nature Biotechnology,25, 2006.

[59] e. a. Gill, P., The evolution of DNA database - recommendations for new European STR loci. Forensic Science International, (2006).

[60] e. a. Ohtaki, H., A powerful, novel, multiplex typing system for six short tandem repeat loci and the allele frequency distributions in two Japanese regional populations. Electropjoresis, 23, 2002.

[61] C. M. D. . B. J.M., Characterization of new miniSTR loci to aid analysis of degraded DNA. Journal of Forensic Sciences, 50(1), 2005.

[62] e. a. Hill, C. R., Characterization of 26 new miniSTR loci for improved analysis of degraded DNA SAMPLES. Journal of Forensic Sciences, 53(1), 2008.

[63] e. a. Coble, M. D., Characterization and performance of new miniSTR loci for typing degraded samples. Elsevier Science: Amsterdam, The Netherlands, International Congress series. Progress in Forensic Genetics 11, 2006.

[64] e. a. Hill, C.R., A 26lex autosomal STR assay to aid human identity testing. Journal of Forensic Sciences, 54, 2009.

[65] J.-A. B. Luke Volgin, Duncan Taylor and M.-H. Lin, Validation of a neural network approach for STR typing to replace human reading. Forensic Science International, 2021.

[66] J. M. Davis and J. C. Giddings, Statistical theory of component overlap in multicomponent chromatograms. American Chemical Society, 55 ed., 1983.

[67] M. Woldegebriel and G. Vivo-Truyols, Probabilistic Model for Untargeted Peak Detection in LC/MS Using Bayesian Statistics. Institute for Molecular Sciences, University of Amsterdam, 2015.

[68] M. A. R. J. D. T. Moroose and T. Ambrose, Detecting STR peaks in degraded DNA samples. in Proc. 4th International Conference on Bioinformatics and Computational Biology (BICoB), 2012.

[69] NIST 1036 Revised U.S. Population Dataset (July 2017) . Available at: $< https : //strbase.nist.gov/NISTpop.htm >$.

[70] IntegenX, Inc. - RapidHIT 200 Rapid DNA Profiling System. Available at: $< https : //www.wmddetectorselector.army.mil/detectorPages/253.aspx >$.

[71] FBI Rapid dna official website. Available at: $< https : //www.fbi.gov/services/laboratory/biometric - analysis/codis/rapid - dna >$.

[72] ANDE DNA IDs^{TM} Rapid DNA system. Available at: $< https : //www.ande.com/ >$.

[73] S. Hussain, ICE's Rapid DNA Testing on Migrants at the Border Is Yet Another Iteration of Family Separation. Electronic Frontier Foundation, 2019.

[74] Special Operators Are Using Rapid DNA Readers. Defense One, 2020.

[75] equation to calculate all three parameters of the peak (height, maximum position and width) available at: `https://www.wolframalpha.com/input/?i=solve+a\%3Dlog\%28h\%29-\%284*log\%282\%29p\%5E2\%29\%2Fw\%5E2\%2Cb\%3D\%288*log\%282\%29*p\%29\%2Fw\%5E2\%2C+c\%3D-\%284*log\%282\%29\%29\%2Fw\%5E2+for+h\%2Cp\%2Cw`.

[76] Gaussian Equation and functions. Available at: $< https : //en.wikipedia.org/wiki/Gaussian_{f}unction >$.

[77] e. Richard G. Lyons, Streamlining Digital Signal Processing: A "Tricks of the Trade" Guidebook.

[78] S. P. . M. M. . G. G. . S. Andrew, Peak picking and the assessment of separation performance in two-dimensional high performance liquid chromatography. 2010.

[79] A. Kim, The Evaluation of the RapidHITTM 200 on Degraded Biological Samples. 2019.

Chapter 7

Appendices

7.1 Appendix A

Allele	**D3S1358**	**vWA**	**D16S539**	**CSF1PO**	**TPOX**
1					
2					
3.2					
5			0.0048		0.0005
6					0.0319
6.4					
7				0.0232	0.0072
8			0.0212	0.0212	0.4662
9			0.1626	0.0294	0.1377
9.3					
10			0.1081	0.2321	0.0599
10.1					
11	0.0005	0.0014	0.2915	0.2736	0.2444
11.2					
11.3					
12	0.0014	0.0010	0.2568	0.3446	0.0512
12.2					
12.3					
13	0.0029	0.0034	0.1371	0.0656	0.0010
13.2					
13.3					
14	0.0874	0.0956	0.0217	0.0092	
14.2					
15	0.3045	0.1347	0.0005	0.0010	
15.2	0.0005				
15.3					
16	0.2828	0.2302			

Allele	**D3S1358**	**vWA**	**D16S539**	**CSF1PO**	**TPOX**
16.1					
16.2					
16.3					
17	0.2042	0.2621			
17.1					
17.2					
17.3					
18	0.1057	0.1800			
18.2					
18.3					
19	0.0092	0.0787			
19.1	0.0010	0.0116			
20					
20.1					
20.2					
21		0.0014			
21.2					
22					
22.2					
22.3					
23					
23.2					
24					
24.2					
25					
25.2					
26					
Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX

Table 7.1: blue dye alleles probabilities (CODIS 2017)

7.2 Appendix B

Allele "1"	**Allele "2"**	**Probabilities**
8	8	0.0093
8	9	0.0086
8	10	0.0057
8	11	0.0280
8	12	0.0294
8	13	0.0112
8	14	0.0041
8	15	0.0001351
9	8	0.0086
9	9	0.008
9	10	0.0053
9	11	0.0259
9	12	0.0272
9	13	0.0104
9	14	0.0038
9	15	0.000125
10	8	0.0057
10	9	0.0053
10	10	0.0035
10	11	0.0171
10	12	0.018
10	13	0.0069
10	14	0.0025
10	15	8.246e-05
11	8	0.028
11	9	0.0259
11	10	0.0171

Allele "1"	**Allele "2"**	**Probabilities**
11	11	0.0844
11	12	0.0886
11	13	0.0338
11	14	0.0122
11	15	0.00041
12	8	0.0294
12	9	0.0272
12	10	0.0180
12	11	0.0886
12	12	0.093
12	13	0.0355
12	14	0.0128
12	15	0.00043
13	8	0.0112
13	9	0.0104
13	10	0.0069
13	11	0.0338
13	12	0.0355
13	13	0.0135
13	14	0.0049
13	15	0.000163
14	8	0.0041
14	9	0.0038
14	10	0.0025
14	11	0.0122
14	12	0.0128
14	13	0.0049

Allele "1"	**Allele "2"**	**Probabilities**
14	14	0.0018
14	15	5.8e-05
15	8	0.00014
15	9	0.00013
15	10	8.246e-05
15	11	0.00041
15	12	0.00043
15	13	0.000163
15	14	5.88e-05
15	15	1.96e-06

Table 7.2: A combined probability for alleles one and two for D13S317 Locus

7.3 Appendix C

- **Peak finding Script**

```
function P=findpeaksG(x,y,SlopeThreshold,AmpThreshold,
smoothwidth,peakgroup,smoothtype)
% function P=findpeaksG(x,y,SlopeThreshold,AmpThreshold,
smoothwidth,peakgroup,smoothtype)
% Function to locate the positive peaks in a noisy
x-y time series data
% set.  Detects peaks by looking for downward
zero-crossings in the first
% derivative that exceed SlopeThreshold. Returns
list (P) containing peak
% number and position, height, width,
and area of each peak, determined
% by least-squares fitting of a Gaussian to
”peakgroup” data points across
% the top of each peak. Arguments ”slopeThreshold”
, ”ampThreshold” and
% ”smoothwidth” control peak sensitivity.
Higher values will neglect
% smaller features. ”Smoothwidth” is the width
of the smooth applied before
% peak detection; larger values ignore narrow peaks.
If smoothwidth=0, no
% smoothing is performed.
% ”Peakgroup” is the number points around the
top part of the peak that are
% taken for measurement. If Peakgroup=0 the
local maximum is takes as the
% peak height and position.
% ”smoothtype” determines the smooth algorithm:
```

```
%   If smoothtype=1, rectangular (sliding-average or boxcar)
%   If smoothtype=2, triangular (2 passes of sliding-average)
%   If smoothtype=3, pseudo-Gaussian (3 passes of sliding-average)
% Examples:
% findpeaksG(0:.01:2,humps(0:.01:2),0,-1,5,5)
% x=[0:.01:50];y=(1+cos(x)).^2;P=findpeaksG(x,y,0,-1,5,5)
% x=[0:.01:5]';y=x.*sin(x.^2).^2;P=findpeaksG(x,y,0,-1,5,5)
% x=[-10:.1:10];y=exp(-(x).^2);P=findpeaksG(x,y,0.005,0.3,3,5,3)
%
% Find, measure, and plot noisy peaks with unknown positions
%   x=-50:.2:50;
%   y=exp(-(x).^2)+exp(-(x+50*rand()).^2)+.02.*randn(size(x));
%   plot(x,y,'m.')
%   P=findpeaksG(x,y,0.001,0.2,5,5,3);
%   text(P(:,2),P(:,3),num2str(P(:,1)))
%   disp('          peak #    Position    Height')
%   disp(P)
if nargin~=7;smoothtype=1;end  % smoothtype=1
if not specified in argument
if smoothtype>3;smoothtype=3;end
if smoothtype<1;smoothtype=1;end
if smoothwidth<1;smoothwidth=1;end
smoothwidth=round(smoothwidth);
peakgroup=round(peakgroup);
if smoothwidth>1
    d=fastsmooth(deriv(y),smoothwidth,smoothtype);
else
    d=deriv(y);
end
n=round(peakgroup/2+1);
P=[0 0 0 0 0];
vectorlength=length(y);
```

```
peak=1;
for j=2*round(smoothwidth/2)-1:length(y)-smoothwidth-1
if sign(d(j)) > sign (d(j+1)) % Detects zero-crossing
if d(j)-d(j+1) > SlopeThreshold % if slope of
derivative is larger than SlopeThreshold
if y(j) > AmpThreshold  % if height of peak is
larger than AmpThreshold
xx=zeros(size(peakgroup));yy=zeros(size(peakgroup));
for k=1:peakgroup % Create sub-group of points near peak
groupindex=j+k-n+2;
if groupindex<1, groupindex=1;end
if groupindex>vectorlength, groupindex=vectorlength;end
xx(k)=x(groupindex);
yy(k)=y(groupindex);
end % for k=1:peakgroup,...
if peakgroup>2
[Height,Position,Width]=gaussfit(xx,yy);
PeakX=real(Position);
% Compute peak position and height of fitted parabola
PeakY=real(Height);
MeasuredWidth=real(Width);
% if the peak is too narrow for least-squares
technique to work
% well, just use the max value of y
in the sub-group of points near peak.
else
PeakY=max(yy);
pindex=val2ind(yy,PeakY);
PeakX=xx(pindex(1));
MeasuredWidth=0;
end % if peakgroup>2,...
% Construct matrix P. One row for each peak detected,
% containing the peak number, peak position (x-value) and
% peak height (y-value). If peak measurement fails and
% results in NaN, or if the measured peak height is less
% than AmpThreshold, skip this peak
if isnan(PeakX) || isnan(PeakY) || PeakY<AmpThreshold
% Skip this peak
else % Otherwise count this as a valid peak
P(peak,:) = [round(peak) PeakX PeakY MeasuredWidth
1.0646.*PeakY*MeasuredWidth];
peak=peak+1; % Move on to next peak
end % if isnan...
end % if y(j) > AmpThreshold,...
        end %  if d(j)-d(j+1) >...
    end %  if sign(d(j)) >...
end % for j=2*round(smoothwidth/2)-1:...
% ------------------------------------------------------------
function [index,closestval]=val2ind(x,val)
% Returns the index and the value of the element of
vector x that is closest to val
% If more than one element is equally close,
returns vectors of indicies and values
% Tom O'Haver (toh@umd.edu) October 2006
% Examples: If x=[1 2 4 3 5 9 6 4 5 3 1], then
val2ind(x,6)=7 and val2ind(x,5.1)=[5 9]
% [indices values]=val2ind(x,3.3) returns
indices = [4 10] and values = [3 3]
dif=abs(x-val);
index=find((dif-min(dif))==0);
closestval=x(index);

function d=deriv(a)
% First derivative of vector using
```

```
2-point central difference.
% T. C. O'Haver, 1988.
n=length(a);
d=zeros(size(a));
d(1)=a(2)-a(1);
d(n)=a(n)-a(n-1);
for j = 2:n-1
    d(j)=(a(j+1)-a(j-1)) ./ 2;
end

function SmoothY=fastsmooth(Y,w,type,ends)
% fastbsmooth(Y,w,type,ends) smooths vector Y with smooth
%  of width w. Version 2.0, May 2008.
% The argument "type" determines the smooth type:
% If type=1, rectangular (sliding-average or boxcar)
% If type=2, triangular (2 passes of sliding-average)
% If type=3, pseudo-Gaussian (3 passes of sliding-average)
% The argument "ends" controls how the "ends" of the signal
% (the first w/2 points and the last w/2 points) are handled.
% If ends=0, the ends are zero.  (In this mode the elapsed
% time is independent of the smooth width). The fastest.
% If ends=1, the ends are smoothed with progressively
% smaller smooths the closer to the end. (In this mode the
% elapsed time increases with increasing smooth widths).
% fastsmooth(Y,w,type) smooths with ends=0.
% fastsmooth(Y,w) smooths with type=1 and ends=0.
% Example:
% fastsmooth([1 1 1 10 10 10 1 1 1 1],3)= [0 1 4 7 10 7 4 1 1 0]
% fastsmooth([1 1 1 10 10 10 1 1 1 1],3,1,1)= [1 1 4 7 10 7 4 1 1 1]
% T. C. O'Haver, May, 2008.
if nargin==2, ends=0; type=1; end
if nargin==3, ends=0; end
```

```
switch type
case 1
SmoothY=sa(Y,w,ends);
case 2
SmoothY=sa(sa(Y,w,ends),w,ends);
case 3
SmoothY=sa(sa(sa(Y,w,ends),w,ends),w,ends);
end

function SmoothY=sa(Y,smoothwidth,ends)
w=round(smoothwidth);
SumPoints=sum(Y(1:w));
s=zeros(size(Y));
halfw=round(w/2);
L=length(Y);
for k=1:L-w
    s(k+halfw-1)=SumPoints;
    SumPoints=SumPoints-Y(k);
    SumPoints=SumPoints+Y(k+w);
end
s(k+halfw)=sum(Y(L-w+1:L));
SmoothY=s./w;
% Taper the ends of the signal if ends=1.
if ends==1
    startpoint=(smoothwidth + 1)/2;
    SmoothY(1)=(Y(1)+Y(2))./2;
    for k=2:startpoint
        SmoothY(k)=mean(Y(1:(2*k-1)));
        SmoothY(L-k+1)=mean(Y(L-2*k+2:L));
    end
    SmoothY(L)=(Y(L)+Y(L-1))./2;
end
```

```
% ------------------------------------------------------------------------
function a=rmnan(a)
% Removes NaNs and Infs from vectors,
replacing with nearest real numbers.
% Example:
%  >> v=[1 2 3 4 Inf 6 7 Inf  9];
%  >> rmnan(v)
%   ans =
%   1 2  3  4 4  6  7 7  9
la=length(a);
if isnan(a(1)) || isinf(a(1)),a(1)=0;end
for point=1:la
    if isnan(a(point)) || isinf(a(point))
        a(point)=a(point-1);
    end
end

function [Height, Position, Width]=gaussfit(x,y)
% Converts y-axis to a log scale, fits a parabola
% (quadratic) to the (x,ln(y)) data, then calculates
% the position, width, and height of the
% Gaussian from the three coefficients of the
% quadratic fit.  This is accurate only if the data have
% no baseline offset (that is, trends to zero far off the
% peak) and if there are no zeros or negative values in y.
%
% Example 1: Simplest Gaussian data set
% [Height, Position, Width]=gaussfit([1 2 3],[1 2 1])
%     returns Height = 2, Position = 2, Width = 2
%
% Example 2: best fit to synthetic noisy Gaussian
% x=50:150;y=100.*gaussian(x,100,100)+10.*randn(size(x));
% [Height,Position,Width]=gaussfit(x,y)
%    returns [Height,Position,Width] clustered
around 100,100,100.
%
% Example 3: plots data set as points and best-fit
Gaussian as line
% x=[1 2 3 4 5];y=[1 2 2.5 2 1];
% [Height,Position,Width]=gaussfit(x,y)
% plot(x,y,'o',linspace(0,8),Height.*gaussian(linspace(0,8),
Position,Width))
% Copyright (c) 2012, Thomas C. O'Haver

% To prevent problems from taking the log
of zero or negative values,
% make the lowest value of y equal to 1%
of the maximum value.
maxy=max(y);
for p=1:length(y)
    if y(p)<(maxy/100),y(p)=maxy/100;end
end % end of for p=1:length(y),
logabsy=log(abs(y));
sizex=size(x);
sizey=size(logabsy);
if (sizex(1)==sizey(1))
    [coef,~,MU]=polyfit(x',logabsy',2);
else
    [coef,~,MU]=polyfit(x,logabsy',2);
end
c1=coef(3);c2=coef(2);c3=coef(1);
% Compute peak position and height or fitted parabola
Position=-((MU(2).*c2/(2*c3))-MU(1));
Height=exp(c1-c3*(c2/(2*c3))^2);
```

```
Width=norm(MU(2).*2.35703/(sqrt(2)*sqrt(-1*c3)));
```