Charles University

Faculty of Science

Study programme: Biology

Branch of study: Anthropology and Human genetics

Lejla Salihagić

Detection of CNV polymorphisms in regions of forensic STR loci in Czech population

Detekce CNV polymorfismů v oblastech forenzně významných STR lokusů v české populaci

Master's thesis

Supervisor: prof. RNDr. Marie Korabečná, Ph.D.

Prague, 2022

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 10.8.2022

_____

Bc. Lejla Salihagić

Poděkování:

Ráda bych poděkovala své školitelce prof. RNDr. Marie Korabečné, Ph.D. za její pomoc, rady, trpělivost a za možnost dělat diplomovou práci v oboru forenzní genetiky. Děkuji Mgr. Ivetě Zedníkové za její přátelský přistup, dostupnost a pomoc s laboratorní části práce. Chtěla bych poděkovat i Mgr. Vlastimilu Stenzlovi z Kriminalistického Ústavu v Praze, který poskytl DNA vzorky podle analýzy Národní databáze forenzních DNA profilů ČR. Obzvlašť bych chtěla poděkovat doc. Mgr. Kateřině Machové Polákové, Ph.D. za poskytování laboratoře digitální PCR na Ůstavu hematologie a krevní transfuze, Mgr. Haně Žižkové, Ph.D. a Mgr. Václavě Polívkové, Ph.D. za jejích odborné rady a sdílení bohatých zkušeností s technologií digitální PCR. V neposlední řadě také děkuji svým nejbližším za neskutečnou podporu, které se mi dostávalo jak z Bosny, tak odsud. Tato práce byla vypracovana ve spolupráci s Ústavem biologie a lékařské genetiky 1. Lékařské fakulty UK a Všeobecné fakultní nemocnice Praha a Kriminalistickým ústavem Praha.

Abstract

DNA analysis aimed at individual identification of persons is based on the examination of a panel of STRs (Short Tandem Repeats). These polymorphic loci were selected on the basis of broadly conceived international testing and the frequencies of individual alleles across different ethnicities are known.

Currently, attention is also paid to polymorphisms involving larger regions of DNA - the so-called Copy Number Variants (CNV). According to the literature (Repnikova et al., 2013), it turns out that these CNVs can also affect areas of forensically significant STR loci, resulting in, for example, the deletion of one of the alleles. The examined person, then in such an affected locus is not a homozygote, but a hemizygote. Otherwise, when duplication event of one of the alleles occurs, a tri-allelic STR arises. Such situations can then distort the result of individual identification, or even the correct evaluation of kinship relations.

In the scope of master's thesis, a methodology based on digital droplet PCR (ddPCR) will be established, which will allow verification in cases where routinely obtained DNA profiling results in the suspicion of the presence of a third allele in one of the STR loci. Samples for testing will be obtained in cooperation with the Institute of Criminology in Prague on the basis of a jointly designed research project. 200 000 DNA profiles obtained from control samples - buccal swabs - will be screened. The result of the screening will be the selection of suspicious samples for testing using the ddPCR method. We assume that we will find rare CNVs and validate their presence using the ddPCR method.

Key words: polymorphisms, Short Tandem Repeats (STR), Copy Number Variants (CNV), droplet digital PCR (ddPCR), tri-allelic, profiling

Abstrakt

DNA analýza směřující k individuální identifikaci osob je založena na vyšetření panelu polymorfismů typu STR (Short Tandem Repeats). Tyto polymorfní lokusy byly vybrány na základě široce pojatého mezinárodního testování, jsou známy frekvence jednotlivých alel napříč různými etniky.

V současné době je věnována pozornost také polymorfismům zahrnujícím větší oblasti DNA - tzv. Copy Number Variants (CNVs). Dle literatury (Repnikova et al., 2013) se ukazuje, že tyto varianty mohou postihovat také oblasti forenzně významných STR lokusů, což má za následek například deleci jedné z alel. Vyšetřovaná osoba, pak není v takto postiženém lokusu homozygotem, ale hemizygotem. V opačném případě, při duplikaci jedne z alel, vzniká trialelický STR. Takovéto situace můžou pak zkreslit výsledek individuální identifikace, eventuálně také správné vyhodnocení příbuzenských vztahů.

V rámci diplomové práce bude zavedena metodika založená na digital droplet PCR (ddPCR), která umožní ověření v případech, kdy z rutinně získaného DNA profilování vyplyne podezření na přítomnost třetí alely u jednoho z STR lokusů. Vzorky pro testování budou získány ve spolupráci s Kriminalistickým ústavem Praha na základě společně řešeného výzkumného projektu. Screeningu bude podrobeno 200 000 DNA profilů získaných z kontrolních vzorků - bukálních stěrů. Výsledkem screeningu bude vytipování podezřelých vzorků pro otestování metodou ddPCR. Předpokládáme, že nalezneme vzácně se vyskytující CNV a jejich přítomnost validujeme pomocí metody ddPCR.

Klíčová slova: polymorfismy, Short Tandem Repeats (STR), Copy Number Variants (CNV), droplet digital PCR (ddPCR), trialelický, profilování

Table of Contents

List of Abbreviations

AP3B1	Adaptor Related Protein Complex 3 Subunit Beta 1
bp	Base pair
BHQ	Black Hole Quencher™
CGH	Comparative genome hybridization
CNV	Copy number variant
CODIS	Combined DNA Index System
CSF1PO	Human c-fms proto-oncogene for CSF-1 receptor gene
CV	Coefficient of variation
ddPCR	Droplet digital polymerase chain reaction
DECIPHER	DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
ENFSI	European Network of Forensic Science Institutes
ESS	European Standard Set
FAM	Fluorescein amidite
FBI	Federal Bureau of Investigation
FGA	Human alpha fibrinogen gene
HCl	Hydrochloric Acid
HEX	Hexachloro-fluorescein
ICP	Institute of Criminalistics Prague
MMEJ	Microhomology-mediated end joining
NAHR	Non-allelic homologous recombination

NCBI	National Center for Biotechnology Information
NHEJ	Non-homologous end joining
NIST	National Institute of Standards and Technology
NTC	No Template Control
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase chain reaction
PI	Paternity Index
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
SNP	Single-nucleotide polymorphism
TH01	Tyrosine hydroxylase 1
TPOX	Human thyroid peroxidase
UCSC	University of California, Santa Cruz
UK	United Kingdom
US	United States
dUTP	Deoxyuridine Triphosphate
UV	Ultraviolet
VNTR	Variable number tandem repeat
vWA	von Willebrand Factor

1 Introduction

Forensic DNA analysis, typing, profiling or DNA fingerprinting represents the most important breakthrough in molecular biology used in criminal investigations and paternity cases solving (Butler, 2005). Through further decades it found its place in mass disaster victim identification, genetic genealogy and ancestry tests. Forensic DNA analysis from early 1990s till presence is based on examination of panel of polymorphisms known as short tandem repeats (STRs). STRs or microsatellites are polymorphic regions of DNA that consist of 2-7 bp long repeat units. STRs used in forensic DNA profiling are selected according to stringent criteria and international research. STR alleles and their frequencies in population are estimated and published in datasets (Bodner et al., 2016).

Another group of polymorphisms, with major impact on genetic diversity and phenotypic variation are copy number variants (CNVs). CNVs are strings of DNA with size of 1 kilobase (kb) to several megabases (Mb) that display copy number differences (copy number gain or copy number loss) in normal population. Until recently, researchers have been mainly interested in CNVs as potential contributors to different diseases and complex traits (Henrichsen et al., 2009; Iskow et al., 2012). According to the literature (Picanço et al., 2014; Repnikova et al., 2013; Yang et al., 2020), CNVs affecting regions of forensic STR loci can lead to deletion or gain of one of the alleles.

Forensic DNA profiling relies on examining a set of STR loci, where one locus is defined by two alleles that individual inherits in Mendelian fashion, one from mother, another from father. Forensic DNA profiles are presented by electropherograms, where one peak at certain locus indicates homozygous STR with two identical alleles, two peaks indicate heterozygous STR with two different alleles. Therefore, copy number gain or copy number loss in the region of STR may aggravate and distort data interpretation in cases of individual identification and paternity testing. Presence of a peak can then be misinterpreted as a homozygosity, although one allele has been deleted due to the copy number loss and the individual is hemizygous. In case of copy number gain, three peaks occur at affected locus, where third locus may be misinterpreted as stutter, common biology-related artifact arising due to the strand slippage during DNA synthesis.

Within the scope of this master's thesis, methodology based on droplet digital PCR (ddPCR) was set up and used to inspect and verify presence of CNVs in regions of selected STRs. Samples for testing were provided by Institute of Criminalistics Prague, where screening of more than 200 000 DNA profiles from National DNA Database of Czech Republic was performed. Screened DNA profiles were obtained from reference material – buccal swab. As a result of the screening, 42 samples with suspicion of CNV presence were selected and tested by ddPCR method – 22 samples for *FGA* locus and 20 samples for locus *D3S1358*.

2 Literature Overview

2.1 Biological background of forensic DNA analysis

An individual inherits one set of chromosomes from mother and another one from father. Chromosomes, which are identical in the size and genetic structure, make a homologous pair. Therefore, a homologous pair will contain two alternative possibilities - alleles, on the same position – loci (singular locus). Exceptions from this rule represent individuals with trisomy, somatic mutations and chimerism (Buckleton, Bright, & Taylor, 2016). In relation to the two alleles on homologous pair of chromosomes being identical or different, the state is described as homozygosity or heterozygosity, respectively. Description of alleles on the same locus represents genotype of a given locus. Combination of examined genotypes of multiple loci leads us to the DNA profile. Multiple loci examination is the key for human identity testing, since it reduces the possibility of a random match between unrelated individuals when the DNA profiles are compared (Butler, 2005).

2.2 Molecular markers - autosomal STRs

More than 99.7% of human genome is shared as identical between individuals, so uniqueness and individuality at genetic level is hidden in those remaining 0.3%. Short tandem repeats (STRs) contribute to this individuality by a unique fraction. They are dispersed throughout the human genome and according to J. M. Butler (2010), they make up 3% of human genome. Madsen et al. (2010) are mentioning even higher percentage of STR coverage, 4.3% of the genome. The ones used for purposes of forensic genetics are located in intronic, non-coding regions of DNA, implying that they do not have known impact on genetic health of an individual. Their localization is probably defined by low or non-existent selective pressure on non-coding loci, so these variants can persist in populations through generations (Buckleton, Bright, & Taylor, 2016). However, recent studies show that STR markers used in forensic casework may be playing part in gene regulation via different mechanisms, leading to the conclusion that they are potentially associated with individuals' phenotype (Wyner et al., 2020).

STRs have repeat unit of size 2-7 bp, so one of the classifications of STRs is based on size of the repeat unit. Two nucleotides in the repeat unit make dinucleotide repeats, three

trinucleotides, four tetranucleotides, five pentanucleotides and six nucleotides make hexanucleotide repeats (J. M. Butler, 2010b). This repetitive nature of STRs leads to frequent occurence of slippage events during DNA replication, which results in mutations in the number of repeats. This feature makes STRs important contributors to human genetic variation (Gymrek, 2017). STRs used in forensic genetics are mostly tetra- or pentanucleotides.

Another classification of STRs that is affecting their fitness for forensic purposes is according to their structure: simple repeat (Figure 1a), compound repeat (Figure 1b), complex repeat (Figure 1c) and complex hypervariable repeat. The latter category is not commonly used in forensic DNA profiling since their complexity makes allele nomenclature challenging and measurements are not easily comparable between laboratories (J. M. Butler, 2010b).

Figure 1: Different types of STRs according to structure. 1a) Simple repeat; 1b) Compound repeat; 1c) Complex repeat

Specificity, sensitivity, feasibility, and ability to simultaneously amplify several loci (multiplexing), automation of the process, made PCR-based STR-typing routine technique used in forensic casework (Shrivastava, Jain, & Kumawat, 2021). Suitable STR locus for purposes of forensic DNA typing is chosen based on following features:

- has discrete and distinguishable alleles;
- enables robust amplification;
- bears high power of discrimination;
- has no genetic linkage with other analyzed loci;

- rare artefact formation during amplification is observed;
- is suitable for amplification with multiple loci (multiplexing) (Goodwin, Linacre, & Hadi, 2007).

Establishment of sets of standardized STRs known as "core loci" enabled forensic laboratories all over the world to enter DNA profiles in national DNA databases and compare data internationally. In 1997 in US, Combined DNA Index System (CODIS) Core Loci working Group of Federal Bureau of Investigation (FBI) defined thirteen core autosomal STR loci and amelogenin marker on X and Y chromosome for sex determination. By 2017, seven core STR loci were added to the original set (Hares, 2015). In 1999 in Europe, European Network of Forensic Science Institutes (ENFSI) DNA Working Group established European Standard Set (ESS) of STR loci with seven core loci, and amelogenin marker. With great success of national DNA databases and international cooperation that came with signing Treaty of Prüm in 2005, seven ESS loci were not sufficient since the chance of adventitious matches was no longer negligible (Schneider, 2009). This led to addition of five loci, but different companies came with different kits for STR multiplexing, so another four loci are included in some kits.

In 1997, National Institute of Standards and Technology (NIST) created Short Tandem Repeat DNA Internet DataBase or STRBase. Since then, the database gathers all the information and details about STRs useful for forensic DNA typing community. Besides general facts such as, STR and primer sequences, PCR multiplex kits and many more, STRBase also contains information about tri-allelic variants reported and verified from the laboratories worldwide (Ruitberg et al., 2001).

According to STRBase (Butler J.M. et al., 2017) *FGA* (human alpha fibrinogen) is a complex STR, nevertheless, Gettings et al., 2015 classify it as a compound tetranucleotide repeat. It is located in the third intron of the human alpha fibrinogen locus on the long arm of chromosome 4.

Repeat is defined as follows: $[TTTC]_3TTTTTTCT[CTTT]_nCTCC[TTCC]_2$ on GenBank. Mutation rate of *FGA* locus is reported to be 0,28% (Butler J.M. et al., 2017). Study of allele frequencies of *FGA* in a population sample of 1410 unrelated individuals from Czech republic indicates presence of 18 different sizes of alleles in Czech population. Out of 17

STR loci tested, it ranks as third locus with the highest power of discrimination (Šimková et al., 2009).

D3S1358 is simple tetranucleotide repeat. It is located on short arm of chromosome 3, more precisely 3p21.31. On GenBank, repeat is defined as $[AGAT]_{n.}$ Mutation rate of *D3S1358* is 0,12% (Butler J.M. et al., 2017). In previously mentioned article written by Šimková et al. 2009, *D3S1358* locus is represented by 11 differently-sized alleles in Czech population.

2.3 Multiplex PCR and fragmentation analysis

Multiplex PCR is still known to be most reliable and robust method for forensic DNA typing purposes. As any PCR, it is based on three basic steps of denaturation of DNA double strand, annealing of primers to region of interest and primers extension by DNA polymerase. Multiplex PCR enables us combination of multiple primer pairs in one reaction in order to amplify different regions of DNA. These primer pairs need to have similar annealing temperature and minimal regions of complementarity to avoid primer dimers creation, where the primers bind to one another instead to the template DNA. Commercially available multiplex STR kits are utilizing multiple fluorescent dyes that enable spectral resolution and separation of the peaks representing different DNA fragments. These fragments are then compared to internal size standard and correlated to allelic ladder, which contains allelic variants of known repeats. Application of peak detection treshold and interpretation treshold in software connected to capillary electrophoresis instrument ensure reliable data interpretation. (J. M. Butler, 2010). Forensic DNA profile obtained from STR genotyping with usage of a commercially established multiplex STR kit by Promega is presented in Figure 2, where every STR locus tested has one or two peaks, representing homozygote or heterozygote, respectively.

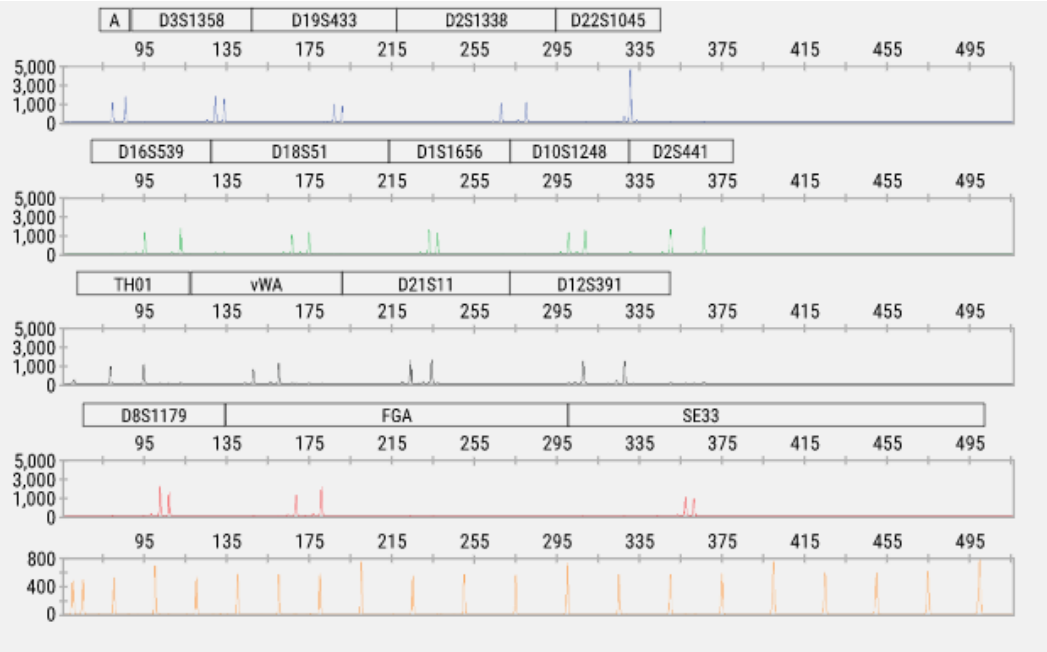


Figure 2: Forensic DNA profile - PowerPlex® ESI 17 Fast System (Adapted from: Promega, Retrieved August, 06, 2022, from https://worldwide.promega.com/products/forensic-dna-analysis-ce/str-amplification/powerplex-esx-17-and-esi-17-fast-systems/?catNum=DC1710. Copyright 2022 by Promega Corporation.

However, extra peaks occurring in DNA profiles are not so rare phenomenon. Occurrence of additional peaks may be of biology or technology-related character. Biology-related artifact peaks include most commonly observed stutters, which are produced during PCR amplification, where one product is one repeat shorter or longer than the main allele peak. Stutters occur as a result of strand slippage by DNA polymerase during elongation step. Another source of artefact peaks is incomplete 3'(+A) nucleotide addition, where Taq DNA polymerase adds an extra nucleotide, resulting in „ split peaks “ (Butler, 2010). Third, and most rare biological source of extra peaks are tri-allelic patterns which are discussed later.

2.4 Copy number variants (CNVs)

Besides previously disscused STRs, another form of genetic variation and important source of polymorphism are copy number variants (CNVs). These involve duplications, deletions, insertions and rearrangements of genomic regions 1 kilobase (kb) in length, or larger (Beckmann, Estivill & Antonarakis, 2007). According to Zarrei et al., 2015, CNVs contribute to 4,8-9,5% of human genome, and approximately 100 genes can be deleted without any effects on the phenotype. Different pathways leading to CNV formation greatly

affect their size, genomic location and potential functional impact. It has been argued that CNVs could explain the majority of human genetic variation, one of the reasons being that they account for at least five times more variable base pairs compared to single-nucleotide variants when two human genomes are compared to each other (Saitou & Gokcumen, 2020).

Determining how the CNVs affect human genome is very complex and challenging. Their size and genome coverage indicate that they most likely have pleiotropic effects, by mechanisms of increasing or decreasing gene dosage of affected and neighboring genes, transcriptome changes, chromation re-organization, changes in regulatory region and chimeric genes formation (Lauer & Gresham, 2019). Mechanism of CNV formations are diverse – recurrent CNVs, same in size and recurring in multiple individuals, are mostly formed by the mechanism of non-allelic homologous recombination (NAHR). These are often present in the regions with extensive homology, where the breakpoints are in close proximity, leading to instability. They are common in population with frequency rate higher than 1%. Usually they are smaller in size and have no impact on phenotype of the individual or might affect susceptibility to more complex diseases (Sismani et al., 2015). Non-recurrent CNVs differ in size and are more diverse in general, so are the mechanisms of their formation – non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), replication slippage, fork stalling and template switching or microhomology-mediated break-induced replication (Pös et al., 2021).

Research of copy number variants led to the formation of databases, which provide major help in the interpration of CNV data. Some of this databases collect data of clinical relevance, such as Online Mendelian Inheritance in Man (OMIM), human genome browsers (UCSC, Ensembl), DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER). In these databases, individual cases with genetic and phenotypic details are reported. Another database, containing data from healthy individuals, serving as a catalogue of CNVs as of „control“ genomes is Database of Genomic Variants (DGV) (Nowakowska, 2017).

2.5 Tri-allelic STR patterns

According to Yang et al. (2020), there are two categories of tri-allelic patterns, type 1 and 2, as shown in Figure 3. Type 1 tri-allelic pattern displays three imbalanced peaks, of which sum of peak heights of second and third peak (minor alleles) equals the peak height of the first peak (major allele). Type 2 pattern is characterized by equal allele intensities and has two possible representations: equal intensity ratio of 1:1:1 three-banded pattern with three different alleles, and pattern, with 2:1 intensity ratio, where two of three alleles are identical. If the three alleles in Type 2 pattern have the same number of repeats, one large peak is generated on electropherogram, which is almost undistinguishable from normal heterozygote. Both of these tri-allelic pattern categories have their own mechanisms of development (Clayton et al., 2004, as cited in Yang et al., 2020). Somatic cell mutation at an early developmental stage is believed to be cause of Type 1 pattern development (Rolf et al., 2002, as cited in Yang et al., 2020), while chromosomal duplication and rearrangement event in germline is considered to be origin of Type 2 pattern (Gu et al, 2008, as cited in Yang et al., 2020).

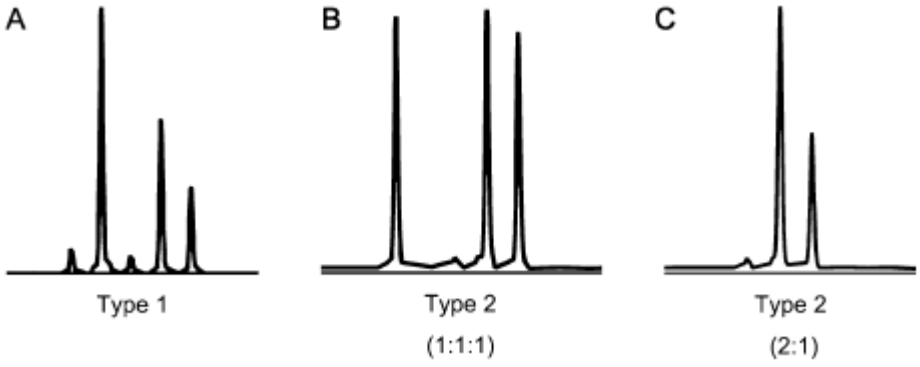


Figure 3: Two types of tri-allelic patterns. (Adapted from: Yang et al., 2020).

Yang et al. (2020) report on prevalence of tri-allelic patterns in CODIS STR loci in global populations (Table 1). They analyzed data from 1172 reported tri-allelic pattern cases. Data indicate significantly higher prevalence of Type 1 tri-allelic pattern at STR loci with higher mutation rate in germline, such as *D18S51* and *FGA*. Loci *D7S820*, *TPOX*, *TH01* and *D16S539* with lower incidence of Type 1 pattern indicate lower germline mutation rates.

Type 2 tri-allelic pattern prevalence indicates close relationship with the incidence of the Type 1 pattern, with the exception of the *TPOX* locus, where significant difference in case report numbers is observable. This is further explained by presence of identical form of *TPOX* rearrangement in ancient African population, along with more stable inheritance of Type 2 tri-allelic pattern.

Table 1: Reports of tri-allelic patterns in CODIS STR loci in global populations (No. of cases:1172). Data are divided in into two groups, according to the pattern type. (Adapted from: Yang et al., 2020).

STR Locus	**Type 1 Case No.**	**Type 2 Case No.**
D7S820	3	7
D3S1358	4	12
TPOX	4	534
TH01	5	7
D16S539	5	8
D13S317	8	15
D5S818	8	5
CSF1PO	11	10
D8S1179	25	35
vWA	54	37
D21S11	61	42
FGA	80	38
D18S51	113	41

Different cell lineages from which DNA is extracted also affect occurence of Type 1 tri-allelic patterns in forensic DNA profiles, since mutations from which they arise may be tissue dependent. This indicates that in the case of crime investigation, sample from body fluid obtained from the crime scene may be genotyped as tri-allelic at certain STR locus, while reference sample from the suspect may be bi-allelic for that same locus. Results of profiling, therefore, may be discordant if samples from crime scene and reference sample are not from same body fluid or tissue (Yang et al., 2021).

In the article by Jiao et al. (2018) case report of maternal testing was desribed, where tri-allelic patterns where observed at two STR loci, *Penta D* and *D21S11*, in child´s DNA profile. At *Penta D* locus, 1:1:1 tri-allelic pattern with equal intensity ratio of three alleles was observed, while *D21S11* locus exhibited two peaks with 1:2 intensity ratio. These results

were suggesting presence of CNVs or trisomy 21, since both of these STR loci are located on chromosome 21. Further examination revealed 1:1:1 pattern at two additional STR loci, and 2:1 pattern at three additional STR loci. Furthermore, karyotyping and whole-genome SNP array analyses confirmed that the child had trisomy 21. Information obtained from this case report further highlight the importance of further investigation of tri-allelic patterns, as well as obtaining information about clinical phenotype in cases of parentage testing.

Vidal & Cassar (2008) reported a case study of type 2 tri-allelic pattern inheritance at locus *D3S1358* from paternal grandmother to granddaughter. To further test posibility of partial duplication of chromosome 3 where *D3S1358* locus is placed, 11 additional STRs across chromosome 3 were tested and tri-allelic patterns were not detected at any of them, which indicates that duplication event was closely linked only to the region of *D3S1358*.

However, occurence of tri-allelic patterns may positively influence discriminatory power such as in case desribed by Wang et al. (2015), where missing girl was discovered raped and murdered in Qishan county, China. Semen retrieved from the crime scene, pointed to one of her neighbours, who however, had identical twin brother living with him. Presence of tri-allelic pattern at the *vWA* locus on STR profile of the brother ruled him out as the perpetrator of the crime. This mutation was found by genotyping in blood, semen and buccal samples of the individual, but hair root sheath cells were genotyped as biallelic. This implies that one-step mutation, in this case insertion of one repeat, occured after formation of zygote. At early stage of development, after division of blastocyte, one of the blastocytes was unaffected while the other had tri-allelic pattern. This further led to mosaicism, where cells from blood, semen and buccal smear had three alleles at *vWA* locus and hair root sheath cells only two alleles.

Analysis of results of 32 850 samples from clinical array comparative genomic hybridization (CGH) in the study by Repnikova et al. (2013) revealed the presence of CNVs at 9 out of 13 CODIS autosomal STR loci regions in 32 individuals, which indicates overall low frequency of CNVs in these regions. Region containing *TPOX* STR was affected most frequently by CNV.

As of March 2017, STRBase counts 401 tri-allelic variant reports in regions of STR loci, out of which 40 are reported at *FGA* locus and 11 at *D3S1358* locus (Butler J.M., Vallone P.M., Gettings K.B., Borsuk L.A., Ruitberg C.M., 2017).

2.6 Droplet digital PCR (ddPCR)

Recent development of digital PCR technologies, when compared to real-time quantitative PCR (qPCR), enable more specific, reproducible and precise, absolute quantification of nucleic acids, without the need for a standard curve (C. M. Hindson et al., 2013). The workflow of droplet-based dPCR is based on random distribution of PCR mix with target nucleic acid across same-sized partitions, or droplets. This leads to number of droplets containing one or more nucleic acid templates, while others contain none of it. After PCR takes place, these two groups of droplets are differentiated by the presence/absence of the fluorescence signal. Termostable droplets are passed through droplet reader, where the detection of presence of fluorescence released from probes indicate a droplet with 0, 1 or more target sequences (Härmälä et al., 2017). Poisson statistics model is then applied to calculate copy number of target nucleic acid template per droplet, and it is derived from the number of positive droplets and the total number of droplets present in reaction (Karlin-Neumann & Bizouarn, 2018).

Key for successful quantification by ddPCR is limiting dilution, where the sample is diluted to the degree where some droplets will contain sequence of interest, but not all of them. This enables ddPCR system to image fluorescence intensity for each droplet separately, normalize it and assign it as either positive or negative (Karlin-Neumann & Bizouarn, 2018). In theory, Poisson statistics is then used and in ddPCR is based on considering probability of any droplet containing none, one or more target sequences, and it estimates average number of molecules per droplet. When average droplet volume is known, Poisson statistics can be used to estimate concentration of the nucleic acid in reaction. It does so by using following equation (Karlin-Neumann & Bizouarn, 2018):

$$[Target\ DNA] = -ln\left(1 - \frac{N_P}{N_T}\right)\frac{1}{V_P} copy\ number\ per\ unit\ volume,$$

where [Target DNA] is the number of target DNA molecules per unit volume of PCR mix, N_p is the number of positive droplets, N_T is the total number of partitions, V_P is average droplet volume (Karlin-Neumann & Bizouarn, 2018).

However, due to the stochastic nature of biological processes and unpredictability of pipetting certain amount of target sequence in total DNA amount put into reaction, theoretical calculation may not be enough. Modern ddPCR systems, such as QX200 AutoDG

system used in this research, are able to overcome limiting dilution as precondition for correct measurement, and robustly quantify target sequences even in cases when multiple sequences occupy the same droplet. Here again, Poisson distribution is applied, considering each droplet as a subsample of a larger sample to calculate average target occupancy per partition, λ. Certainly, it is known that negative droplet contains 0 target sequences, but it is not certain how many target molecules are present in positive droplets. By calculating numbers of negative and positive droplets, λ is then calculated, which further enables estimation of total number of target molecules at the beginning of PCR and target concentration after the reaction Subsampling, or amount of target DNA drawn and put into the reaction, and partitioning, or occupancy of one droplet by different number of target molecules, are main sources of variability in ddPCR data analysis. Uncertainty arising from these two variables is defined by total Poisson statistical error (Karlin-Neumann & Bizouarn, 2018).

3 Research Objectives

Master's thesis focuses on the establishment of a methodology for detection of CNVs in regions of forensic STR loci in DNA samples of individuals from Czech population using droplet digital PCR (ddPCR) method. For accurate DNA profile interpretation, the presence of a tri-allelic pattern should be confirmed by a technology which is independent on fragmentation analysis performed using capillary electrophoresis.

Our goal was to optimize ddPCR protocol that is suitable for CNV detection in the regions of STR loci. To this date, application of ddPCR in these purposes is not published in any literature available on Internet. We assume that two sets of samples provided by ICP have third allele present at loci *FGA* and *D3S1358* and that ddPCR could be right choice for confirmation of tri-allelic pattern indicated by capillary electrophoresis.

In the article by Repnikova et al., (2013) mentioned in 2.5 Tri-allelic patterns subchapter, it was discussed that CNVs involving STR markers could be present at higher rate than reported in their article, but at the smaller size than defined by their laboratories, when CGH array technology is used. We assume that ddPCR technology is capable of detection of smaller-sized CNVs by proper primer and probe design for selected loci.

4 Materials and Methods

4.1 Samples

DNA samples for the purposes of ddPCR analysis were provided by Institute of Criminalistics Prague, after screening of more than 200 000 DNA profiles from National DNA Database. Following criteria were considered during selection of samples:

1. **Electropherogram indicates potential presence of tri-allelic pattern at one of the STR loci.**
2. **DNA profile is the result of reference sample DNA analysis.** DNA analysis of buccal swab reference sample excludes possibility of tri-allelic pattern presence due to DNA sample mixture. Furthermore, in most cases reference sample contains DNA of higher quality and concentration than the one found at the crime scene.
3. **DNA profile belongs to an individual from the Czech population.**
4. **DNA from the sample is trackable.** DNA profiles from 2009 to 2019 were subjected to screening since DNA isolated and stored earlier is hardly tracked. Further, DNA obtained from recently collected samples will presumably have lower degradation rate.
5. **Volume of DNA is > 5 µl.** Restriction digestion protocol used in the experiment included adding of 5 µl of DNA sample since the lower volume could be insufficient for DNA quantification and CNV detection.

4.2 Primers and Probes design

Primers and probes were designed with the usage of Primer3 program (Untergasser et al., 2012). FASTA formats of STR sequences defined in the latest, fifth version of Forensic STR Sequence Structure Guide (Phillips et al., 2018) were used for primers and probes localization. Since STRs are repetitive regions, primers and probes were localized adjacent to the actual target STR locus, as shown in Figures 4 and 5 for FGA and D3S1358 STR, respectively. Primers and probes sequences are presented in Table 2. TaqMan probes for detection of amplified sequences were labeled by FAM (fluorescein amidite) reporter dye and BHQ (Black Hole Quencher™) quencher dye.

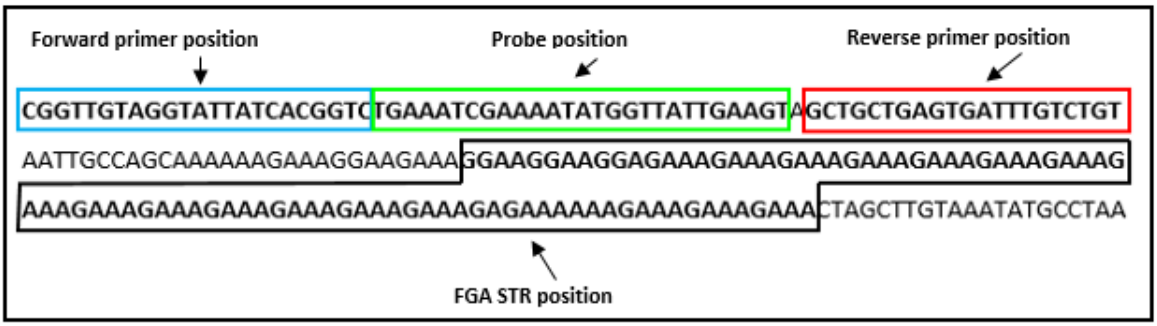


Figure 4: Position of primers, probe in relation to FGA STR position.

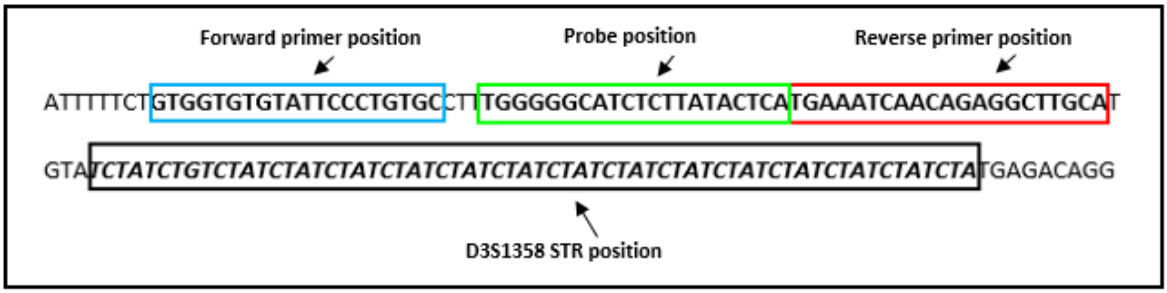


Figure 5: Position of primers, probe in relation to D3S1358 STR position.

Table 2: Primers and probes sequences for FGA and D3S1358 STR loci.

FGA	
Forward primer	5'-CGG TTG TAG GTA TTA TCA CGG TC-3'
Reverse primer	5'-ACA GAC AAA TCA CTC AGC AGC-3'
FAM probe	5'-TGA AAT CGA AAA TAT GGT TAT TGA AGT-3'
D3S1358	
Forward primer	5'-GTG GTG TGT ATT CCC TGT GC-3'
Reverse primer	5'-TGC AAG CCT CTG TTG ATT TCA-3'
FAM probe	5'-TGG GGG CAT CTC TTA TAC TCA-3'

AP3B1 assay (ID: dHsaCP2500315, manufactured by Bio-Rad) with HEX (hexachlorofluorescein) reporter-dye and Iowa Black® quencher-dye labeled probe designed and recommended by Bio-Rad was used as reference target for copy number detection. This assay has been designed and wet-lab validated specifically for ddPCR

technology in Bio-Rad laboratories. *AP3B1*, or adaptor related protein complex 3 subunit 1 is a gene that encodes protein responsible for organelle biogenesis associated with melanosomes, platelet dense granules and lysosomes. The protein encoded by *AP3B1* gene is part of heterotetrameric AP-3 protein complex ("AP3B1 adaptor related protein complex 3 subunit beta 1 (AP3B1, Gene - NCBI, 2022). The gene is represented by one allele on both homologous chromosomes in human cells in healthy individual, therefore, the copy number is always 2.

4.3 DNA Isolation

Since the amount and quality of every DNA sample from ICP was limited and relatively low, DNA used for ddPCR optimization was isolated from buccal swab of healthy female individual (named BUK001).

Buccal swab collection procedure

1. An individual whose buccal swab is collected should not eat or drink for an hour before the procedure.
2. An individual should rinse her/his mouth with water before buccal swab collection.
3. Gloves are obligatory and special caution should be taken when manipulating with the swab. Swab should not touch any external surface before or after the collection procedure.
4. Buccal cells are collected using a Isohelix™ SK-2SDNA/RNA buccal swab, by rubbing firmly inner side of cheeks, lower and upper lip for approximately 1 minute. Swab is then placed into a 2 ml collection tube and closed.

DNA isolation was performed on MagCore® Compact Automated Nucleic Acid Extractor using MagCore® Genomic DNA Tissue Kit manufactured by RBC Bioscience.

DNA Isolation procedure

1. Stick is separated from the swab head and removed. 500 µl GT buffer and 20 µl Proteinase (10mg/ml) are added to the collection tube with the swab head.
2. Sample lysate is incubated at 55°C for 30 min.

3. Lysate is then pipetted onto Filter Column (part of MagCore® Genomic DNA Tissue Kit) and centrifuged 10,000 x g for 5 min in Eppendorf MiniSpin centrifuge.
4. Filtrated lysate is transferred to the collection tube.
5. Tube is put into the correct well of T-rack of MagCore® Compact Automated Nucleic Acid Extractor. Cartridge, pipette tip and elution tube from the MagCore® Genomic DNA Tissue Kit are put into correct wells of T-rack.
6. DNA extraction is performed with program 401, with the elution volume of 60 µl and Tris-HCl elution solution.

4.4 DNA Concentration Measurements

Concentration of the obtained DNA sample was measured by applying 1 µl of the samples on NanoDrop™ One Microvolume UV-Vis Spectrophotometer. Final concentration of the DNA sample BUK001 was 80 ng/µl with 260/280 purity ratio 1.88. In general, DNA with with 260/280 purity ratio of ~ 1.8 is considered "pure" (Lucena-Aguilar et al., 2016). Concentration and purity of selected DNA samples from ICP are shown in Tables 3 and 4. Concentration measurements were further used for calculating amount of DNA entering the PCR.

Table 3: Concentration and A260/A280 purity ratio of selected samples for FGA STR testing.

Sample	Concentration (ng/ µl)	A260/A280 purity ratio
FGA-1	432,1	1.01
FGA-2	286,1	0.97
FGA-3	432,4	1.06
FGA-4	266,9	1.01
FGA-5	293,6	1.05
FGA-6	138,4	0.95
FGA-7	226,1	0.81
FGA-8	116,6	0.98
FGA-9	234,8	1.12
FGA-10	110,5	0.94
FGA-11	234,7	1.03
FGA-12	396,5	1.23
FGA-13	97,7	0.97
FGA-14	85,7	1.07
FGA-15	273,2	1.02
FGA-16	349,4	1.17
FGA-17	225,5	1.07
FGA-18	347,5	1.22
FGA-19	97,8	1.01
FGA-20	140,8	1.14
FGA-21	160,2	1.01
FGA-22	99,5	1.03

Table 4: Concentration and A260/A280 purity ratio of selected samples for D3S1358 STR testing.

Sample	Concentration (ng/ µl)	A260/A280 purity ratio
D3S1358-1	253	1.01
D3S1358-2	117,8	0.96
D3S1358-3	100,9	0.95
D3S1358-4	342,8	1.30
D3S1358-5	275,7	1.05
D3S1358-6	153,7	1.10
D3S1358-7	472,4	1.13
D3S1358-8	172,6	1.14
D3S1358-9	347,4	1.28
D3S1358-10	184,6	1.07
D3S1358-11	222,9	1.16
D3S1358-12	304,9	1.06
D3S1358-13	66,3	0.95
D3S1358-14	287,9	1.06
D3S1358-15	254,9	1.01
D3S1358-16	83,7	1.12
D3S1358-17	72,8	0.98
D3S1358-18	161,9	1.00
D3S1358-19	178,6	0.99
D3S1358-20	585,4	1.27

4.5 DNA Restriction

Enzymatic digestion of DNA samples prior to ddPCR was performed to separate CNVs that may be closely linked. Without enzymatic digestion, tandemly arranged CNVs could stick together throughout ddPCR, resulting in inaccurate quantification. Restriction digestion also improves accessibility of the template region and decreases sample viscosity (Bio-Rad manufacturer guidelines; B. J. Hindson et al., 2011). Enzyme used in this experiment is HaeIII, recommended by Bio-Rad according to reference gene assay choice.

Its source is bacteria Haemophilus aegyptius. Recognition site at which HaeIII cleaves DNA is shown in Figure 6.

Figure 6: Restriction site of HaeIII enzyme indicated by red arrows.

Enzymatic digestion reaction with the final volume of 20 µl was prepared according to Table 5. Reaction tubes with prepared mix were centrifuged for a few seconds at maximum speed in a microcentrifuge and incubated at 37°C for 1 hour. Enzyme was thermally inactivated by incubation at 65°C for 15 minutes.

Table 5: Components and volumes of restriction digestion reaction by HaeIII.

Component	**Volume (per reaction)**
Nuclease-free H_2O	11,8 ul
Buffer C 10X Buffer, Promega	2 ul
Bovine Serum Albumin, Acetylated	0,2 ul
DNA	5 ul
Mix components by pipetting, then add:	
Restriction enzyme Hae III, 10u/µl (Promega)	1 ul

4.6 Gradient PCR

For both STR assays, gradient PCR was performed to determine optimal annealing temperature. The predicted annealing temperature was calculated based on the melting temperature (T_m) of the primers and template. Feature of C1000 thermal cycler, T_a calculator was used to determine T_m of combination of primers and probe (Bio-Rad Laboratories, 2008). Gradient range was set up from 62 to 48°C (62, 60.9, 59.1, 56.5, 53.4, 50.7, 49 and 48°C). For both *FGA* and *D3S1358*, annealing temperature profiles were similar and positive droplets separation from negative droplets was clear and unambiguous (Figure 7).

Protocol for gradient PCR was set as follows:

- 95°C – 10 min
- 39 cycles of 94°C for 30 sec followed by 48-62°C for 1 min
- 98°C – 10 min
- 8°C – infinite hold

Ramp rate for all steps was set to 2°C/cycle.

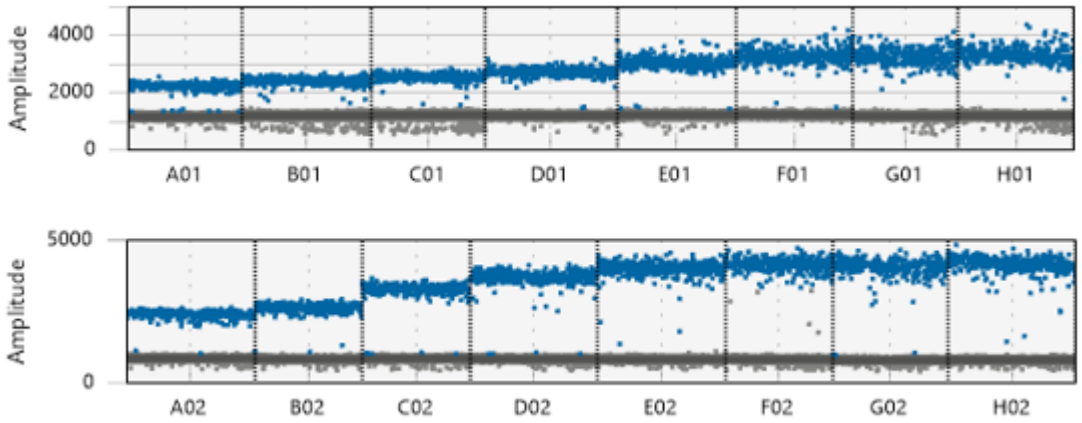


Figure 7: Gradient PCR for FGA primers and probe (upper plot) and D3S1358 primers and probe (lower plot). Blue droplets indicate FAM positive droplets, gray droplets are negative. Wells A01-H01 (FGA) and A02-H02 (D3S1358) – 62, 60.9, 59.1, 56.5, 53.4, 50.7, 49 and 48°C annealing temperatures.

Due to the annealing temperature recommendation by Bio-Rad for *AP3B1* reference gene assay, annealing temperature for further ddPCR experiments was set to be 60°C.

4.7 ddPCR Optimization

ddPCR optimization was performed due to the presence of droplet "rain" in final ddPCR reports of tested samples, as shown in Figure 8. Figure 8a depicts fluorescence amplitude and separation of positive and negative droplets in two channels. In Channel 1, FAM positive droplets contain *FGA* STR amplicon. In Channel 2, HEX positive droplets contain *AP3B1* reference gene amplicon. In both channels, presence of rain negatively influences determination of threshold and accuracy of results. The "rain" is term describing droplets that fall into the range between the droplets recognized as positive or negative. Figure 8b shows graph of Channel 1 vs. Channel 2 fluorescence amplitude, where droplets are scattered and fail to form defined clusters. Origin of rain is not strictly defined since different parameters of reaction are influencing optimal differentiation of positive and negative

droplets. Some of these parameters and factors are amount of DNA input, annealing temperature, assay design, number of PCR cycles, pipetting technique, which may result in droplet damage (Karlin-Neumann & Bizouarn, 2018).

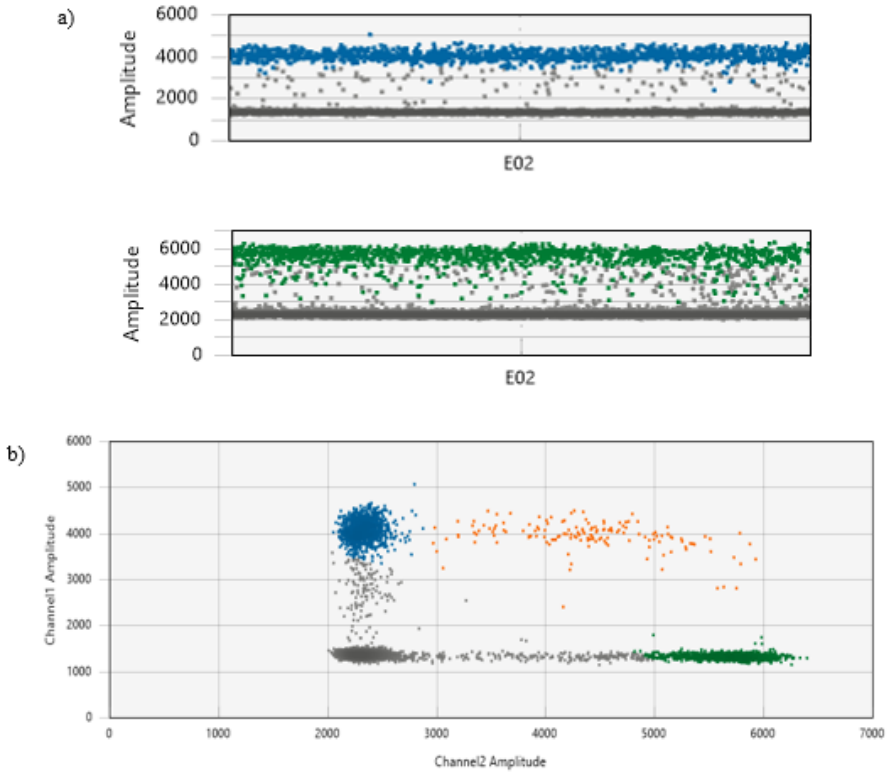


Figure 8: Poor droplet separation. a) Blue dots indicate FAM positive droplets (Channel 1). Green dots indicate HEX positive droplets (Channel 2). Gray droplets are negative in both channels. Determination of threshold that separates positive from negative droplets is challenging due to the rain. b) Blue cluster – FAM positive droplets; green cluster – HEX positive droplets; orange cluster – FAM+HEX positive droplets; gray cluster – negative droplets. Clusters are poorly defined due to the rain.

The first step of the optimization was to determine optimal amount of input DNA. Recommended amount of DNA for ddPCR analyses is 10-66 ng (Karlin-Neumann & Bizouarn, 2018). Reaction was tested with 10, 40 and 160 ng of input BUK001 DNA in duplicates. DNA was digested with HaeIII enzyme prior to PCR according to digestion protocol described in Chapter 4.5. After digestion reaction, ddPCR was set-up.

ddPCR protocol

1. Master mix for PCR reaction was prepared according to Table 6.
2. Every well contained 15 µl of master mix and 5 µl of restricted DNA, except for the last well with NTC (No Template Control) containing water instead the DNA sample. Every DNA sample was pipetted in triplicate.
3. PCR plate was covered with PCR plate heat seal foil.
4. PX1 PCR Plate Sealer (Bio-Rad) was adjusted to heat at 180°C. Plate was inserted and foil was sealed.
5. Plate was taken out and centrifuged at 180 x g, 1 min, 4°C.
6. Plate was transferred to multi plate shaker, 1600 rpm, 1 min.
7. Plate was centrifuged at 180 x g, 1 min, 4°C.
8. QX200 Automated droplet generator (Bio-Rad) was prepared according to Bio-Rad manual and plate was placed inside.
9. After droplets were ready, the PCR plate with droplets was sealed with the foil as described in step 3 and inserted into C1000 Touch Thermal Cycler (Bio-Rad).
10. PCR program was defined as follows:
    - 95°C – 10 min
    - 40 cycles of 94°C for 30 sec followed by 60°C for 1 min
    - 98°C – 10 min
    - 8°C – infinite hold

    Ramp rate for all steps was set to 2.5°C/cycle.
11. Plate was transferred to QX200 Droplet Reader (Bio-Rad). Reader was set-up via QuantaSoft software according to Bio-Rad manual.

Table 6: Components and volumes of mastermix for ddPCR.

Component	**Volume (per reaction)**
SuperMix for Probes (No dUTP) (Bio-Rad)	10 µl
F + R primer (18 µM each)	1 µl
Probe (5 µM)	1 µl
AP3B1 Assay (Bio-Rad)	1 µl
Nuclease-free H_2O	2 µl

Second step of optimization was to determine if a) higher number of PCR cycles; b) lower volume of reference gene *AP3B1* assay in reaction and, c) dilution of input DNA to 40 ng after DNA restriction may positively influence droplet differentiation in case of AP3B1.

PCR reaction was set up as follows:

1. 95°C – 10 min
2. 50 cycles of 94°C for 30 sec followed by 60°C for 1 min
3. 98°C – 10 min
4. 8°C – infinite hold

Ramp rate for all steps was set to 2.5°C/cycle.

Amount of input DNA and volume of reference gene *AP3B1* assay were modified according to Table 7. Other components of the reaction and its amount/volume were identical as in previously described ddPCR protocol. Every reaction mix was run in duplicates.

Table 7: Modification of amount of input DNA and volume of reference gene AP3B1 assay in reaction.

Reaction Mix	Amount of input DNA	Volume of AP3B1 assay
1	10 ng	1 µl
2	40 ng	1 µl
3	40 ng after DNA restriction	1 µl
4	10 ng	0,75 µl (+ 0,25 µl H_2O)
5	40 ng	0,75 µl (+ 0,25 µl H_2O)
6	40 ng after DNA restriction	0,75 µl (+ 0,25 µl H_2O)

4.8 ddPCR testing of STR loci

For *FGA* STR locus, 22 DNA samples of interest alongside with 20 DNA control samples were tested for possible presence of third allele. For *D3S1358* locus, 20 samples of interest and 19 control samples were tested. In both cases, samples provided by ICP in which electropherograms indicated clear presence of two alleles at given locus of interest were used as control. All samples were tested in triplicates. To achieve better precision (higher number of accepted droplets) the results from three technical replicates were then merged for evaluation of each sample. Every well with PCR mix and sample was manually inspected, if the total number of droplets formed was 10 000 or higher. Separation of clusters on 2-D

plot was then checked to determine if double positive, double negative, FAM-positive and HEX-positive droplets were correctly assigned to the clusters. Wells containing the same sample were then merged and analyzed.

4.9 Statistical tests

In CNV data interpretation, concept of confidence interval is applied and calculated by Bio-Rad software, by applying Poission statistics. 95% confidence interval is defned by two numbers, minimum and maximum of the variable, which define the range in between which estimated value will lie 95% of the time. It is represented by error bars shown in Graphs with copy number and ratio target gene : reference gene calculations. During data analysis, it was analysed if confidence interval for copy number calculation is >1, which implicates unambigous measurement of copy number.

Statistical calculations of Z-score and Mann-Whitney U test were performed in order to confirm or reject null hypothesis (H_0). Null and alternative hypotheses were defined as follows:

H_0 = There is no significant difference in copy numbers between datasets of samples of interest and control samples.

H_1 = Copy numbers in dataset of samples of interest are higher than copy numbers in control samples dataset.

Z-score indicates number of standard deviations of a given sample from the mean. It was calculated for every sample of interest separately. After calculation mean and standard deviation of both samples of interest and control samples datasets, Z-score for every sample was calculated as follows:

$Z = (\text{Sample value} - \mu_{(\text{Control samples})})/\sigma_{(\text{Control samples})}$, where:

Z = Z-score
Sample value = Copy number measurement for given sample
$\mu_{(\text{Control samples})}$ = Mean of control samples
$\sigma_{(\text{Control samples})}$ = Standard deviation of control samples

The samples with Z-score higher than +3 were considered as samples with copy number of target sequence higher than two, at least in a fraction of cells according to literature dealing with copy number detection using ddPCR also in mosaic samples (Tan et al., 2019).

During the evaluation and selection of samples with potential duplication of target sequences, we took in account also the extent of 95% confidence intervals calculated by Bio-Rad software. Copy number, as well as ratio target : reference gene were calculated by Bio-Rad software using Poisson statistics.

Mann-Whitney U test was used to compare differences between datasets of samples of interest and control samples and to analyze probability of copy number from samples of interest dataset being greater than copy number from control samples dataset. It was calculated using Social Statistics Calculator (Social Science Statistics, 2022), with settings of statistical significance at 0.05 and two-tailed hypothesis. Firstly, sample sizes of two separate datasets are separately aligned from lowest to higher value. Then, ranks are assigned to sample size, and sum and mean of ranks are calculated. Calculator is then used to calculate U-value, Z-score of datasets and p-value.

5 Results

5.1 ddPCR optimization results

Difference in droplet separation after first ddPCR optimization with 10, 40 and 160 ng are presented in Figure 9. These results indicate 10 and 40 ng as the optimal amount of input DNA for CNV analysis of STR loci by ddPCR. Reaction with 160 ng of input DNA creates droplet rain and makes setting up threshold extremely difficult, especially in case of Channel 2 with HEX positive droplets. Nevertheless, reaction seems to be optimized in case of *FGA* assay (Channel 1), whereas in case of reference gene *AP3B1* assay (Channel 2) it is still quite challenging to determine threshold, even with input DNA of 10 and 40 ng.

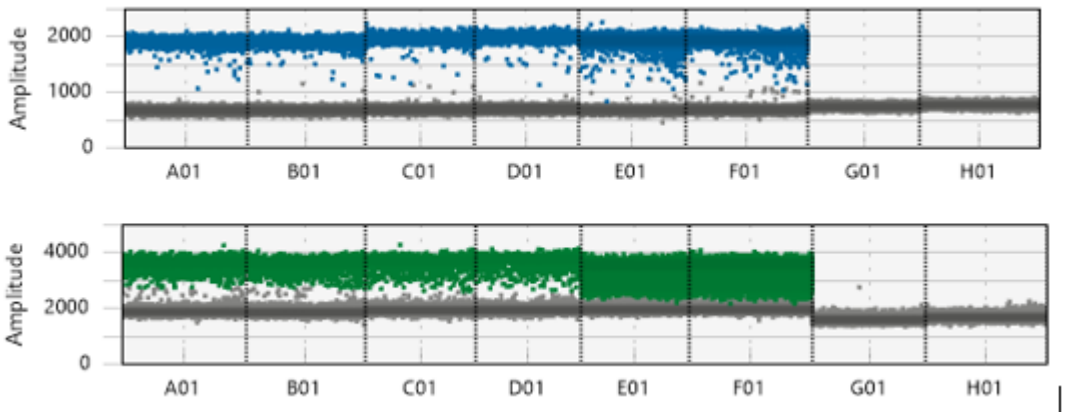


Figure 9: Droplet separation after first ddPCR optimalization. Blue droplets indicate FAM positive droplets (Channel 1). Green droplets indicate HEX positive droplets (Channel 2). Gray droplets are negative in both channels. Wells A01-B01 – 10 ng of input DNA; w wells C01-D01 – 40 ng of input DNA; wells E01-F01 – 160 ng of input DNA; wells G01-H01 – No template control (NTC).

Difference in droplet separation after second ddPCR optimization are presented in Figure 10. Separation of droplets in wells E04 and F04 indicate that reaction is optimized when number of PCR cycles is increased to 50 and 40 ng of DNA input is used after DNA digestion, which corresponds to Reaction Mix 3 from the Table 7. Lower volume of reference gene *AP3B1* assay in reaction affects droplet separation rather negatively, as represented by wells A05-F05.

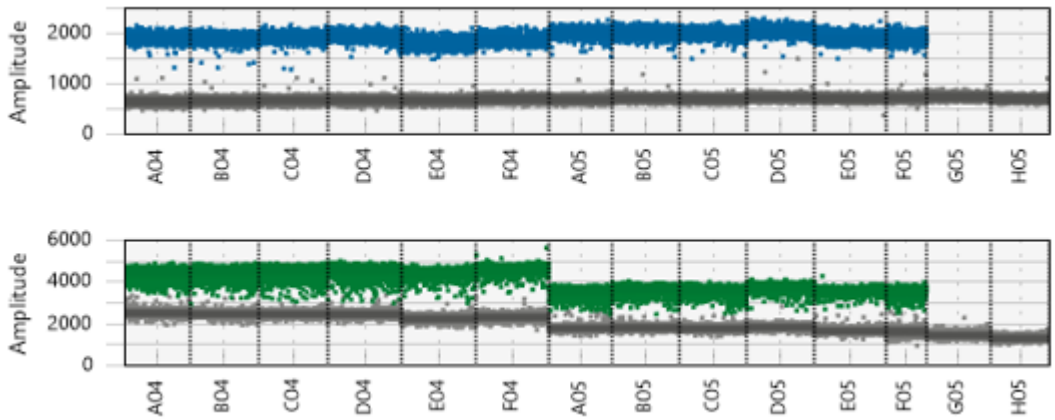


Figure 10: Droplet separation after second ddPCR optimalization. Blue dots indicate FAM positive droplets (Channel 1). Green dots indicate HEX positive droplets (Channel 2). Gray droplets are negative in both channels. Wells A04-B04 – Reaction Mix 1; wells C04-D04 – Reaction Mix 2; wells E04-F04 – Reaction Mix 3; wells A05-B05 – Reaction Mix 4; wells C05-D05 – Reaction Mix 5; wells E05-F05 – Reaction Mix 6; wells G05-H05 – No template control (NTC).

However, when every well is manually inspected and treshold between negative and positive droplets is determined, results shown in Graph 1 indicate optimized reactions in every reaction mix, since results of copy number calculations clearly indicate copy number ~ 2 for and ratio ~ 1 for *FGA : AP3B1* with low error bars in both measurements. For further experiments, number of PCR cycles was incresed to 50, volume of reference gene assay was 1 µl/reaction and input DNA was diluted to 40 ng prior to ddPCR.

In second optimization experiment, average number of accepted droplets per well was 15 769, and average number of positive droplets per well was 2137.

Graph 1: Upper - Copy number calculation (Confidence interval 95%); Lower - Ratio FGA : AP3B1. Wells M01-M06 indicate BUK001 sample tested with 6 different reaction mixes. Reaction M01-M06 – sample with Reaction Mix 1-6, respectively.

5.2 STR loci testing results

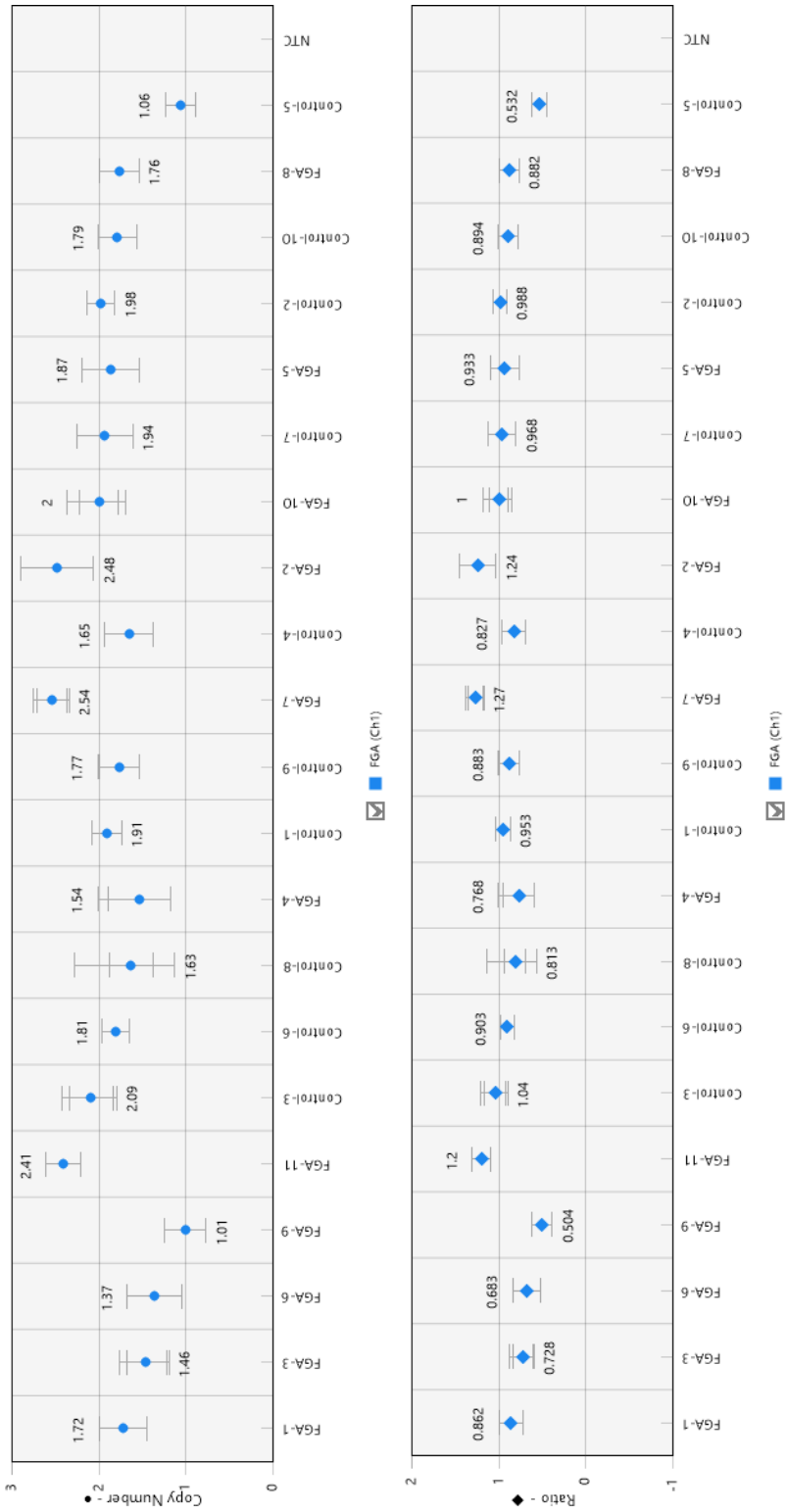
5.2.1 *FGA* STR locus testing results

Tested samples were divided onto two PCR plates, each of them containing 11 samples of interest, 10 control samples and NTC. Graph 2 and Graph 3 show copy number and ratio *FGA : AP3B1* calculations for both PCR plates, respectively. Two control samples had measurements of 95% confidence interval >1, Control-7 and Control-14 (Graphs 2 and 3).

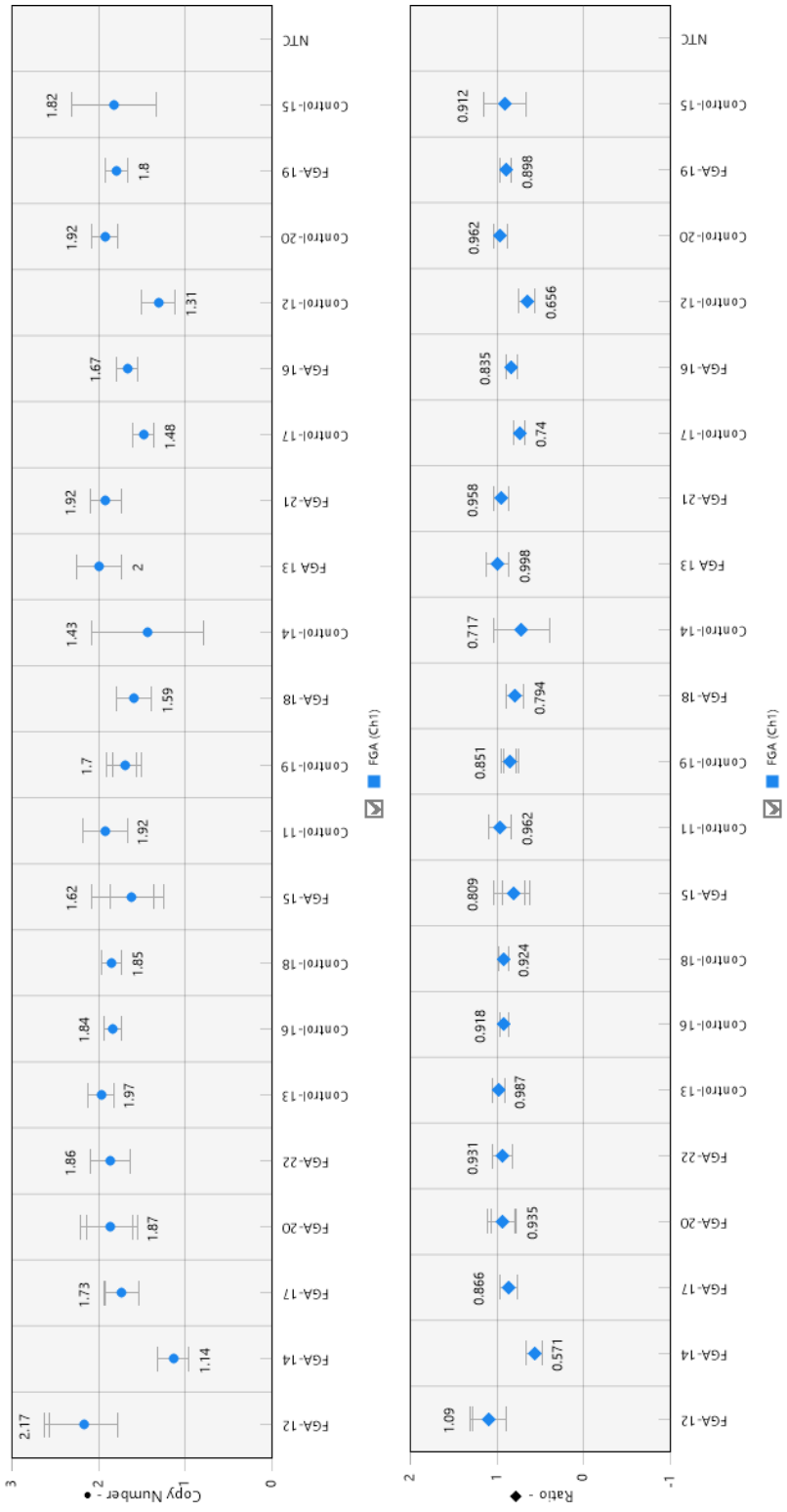
Average number of accepted droplets per well in both PCR plates was 17 327, while the average number of positive droplets per well was 233.

Table 8 shows calculated copy numbers for all the samples tested. Population mean and standard deviation were then calculated for datasets of tested and control samples. Graph 4 plots distribution of copy numbers of *FGA* in tested samples (FGA-1 – FGA 22) and control samples (Control-1 – Control 20). Z-score values for samples FGA-1 to FGA-22 are calculated and presented in Table 9. Mann-Whitney U test calculations are shown in Table 10.

According to copy number calculations from Table 8 and z-score calculations from Table 9, three samples, FGA-2, FGA-7 and FGA-11 with potential tri-allelic FGA STR pattern are chosen and represented by electropherograms provided by Institute of Criminalistics in modified Figures 11, 12 and 13, respectively.



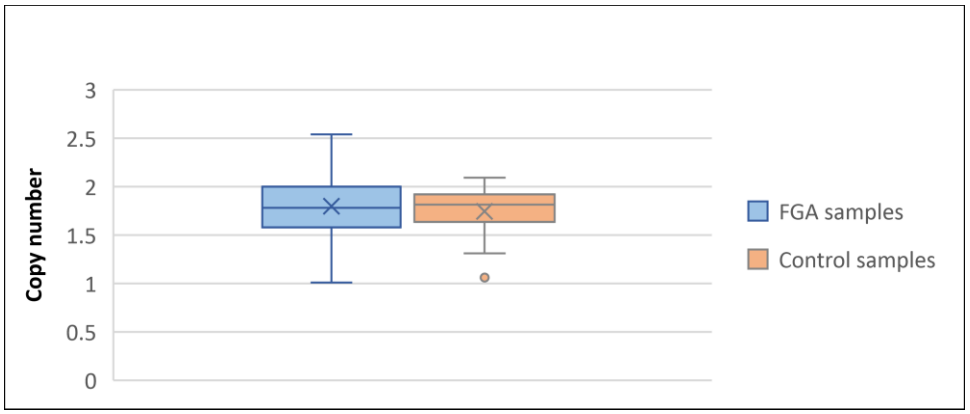
Graph 2: PCR Plate 1: Samples FGA-1 – FGA-11; Control-1 – Control-10; NTC Upper: Copy number calculation (Confidence interval 95%); Lower: Ratio FGA : AP3B1.



Graph 3: PCR Plate 2: Samples FGA-11 – FGA-22; Control-11 – Control-20; NTC Upper: Copy number calculation (Confidence interval 95%); Lower: Ratio FGA : AP3B1.

Table 8: Copy number values of FGA samples and control samples (Samples marked with red color indicate CNV > 2).

Sample	CNV	Control Sample	CNV
FGA-1	1.72	Control-1	1.91
FGA-2	**2.48**	Control-2	1.98
FGA-3	1.46	Control-3	2.09
FGA-4	1.54	Control-4	1.65
FGA-5	1.87	Control-5	1.06
FGA-6	1.37	Control-6	1.81
FGA-7	**2.54**	Control-7	1.94
FGA-8	1.76	Control-8	1.63
FGA-9	1.01	Control-9	1.77
FGA-10	2	Control-10	1.79
FGA-11	**2.41**	Control-11	1.92
FGA-12	2.17	Control-12	1.31
FGA-13	2	Control-13	1.97
FGA-14	1.14	Control-14	1.43
FGA-15	1.62	Control-15	1.82
FGA-16	1.67	Control-16	1.84
FGA-17	1.73	Control-17	1.48
FGA-18	1.59	Control-18	1.85
FGA-19	1.8	Control-19	1.7
FGA-20	1.87	Control-20	1.92
FGA-21	1.92		
FGA-22	1.86		
MEAN	**1.797**		**1.744**
STANDARD DEVIATION	**0.388**		**0.255**



Graph 4: Distribution of copy number values of FGA samples of interest and control samples.

Table 9: Z-score values of samples of interest for FGA STR testing (Samples marked with red color: Z-score > 3).

Sample	Z-score
FGA-1	-0.09
FGA-2	2.89
FGA-3	-1.11
FGA-4	-0.80
FGA-5	0.49
FGA-6	-1.47
FGA-7	**3.12**
FGA-8	0.06
FGA-9	-2.88
FGA-10	1.00
FGA-11	2.61
FGA-12	1.67
FGA-13	1.00
FGA-14	-2.37
FGA-15	-0.49
FGA-16	-0.29
FGA-17	-0.05
FGA-18	-0.60
FGA-19	0.22
FGA-20	0.49
FGA-21	0.69
FGA-22	0.45

Table 10: Mann-Whitney U test calculation for FGA samples.

Dataset	n	Mean of ranks	Sum of ranks	U-value, Z-score, p
FGA samples	22	21.77	479	U = 214, Z = 0.13, p = 0.88
Control samples	20	21.2	424	

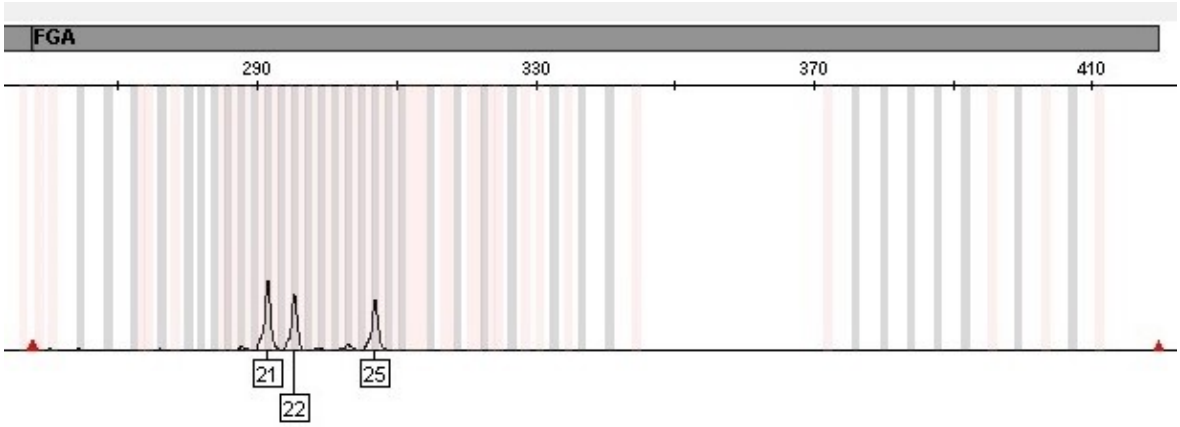


Figure 11: FGA-2 sample electropherogram – FGA STR locus with three alleles, 21,22 and 25.

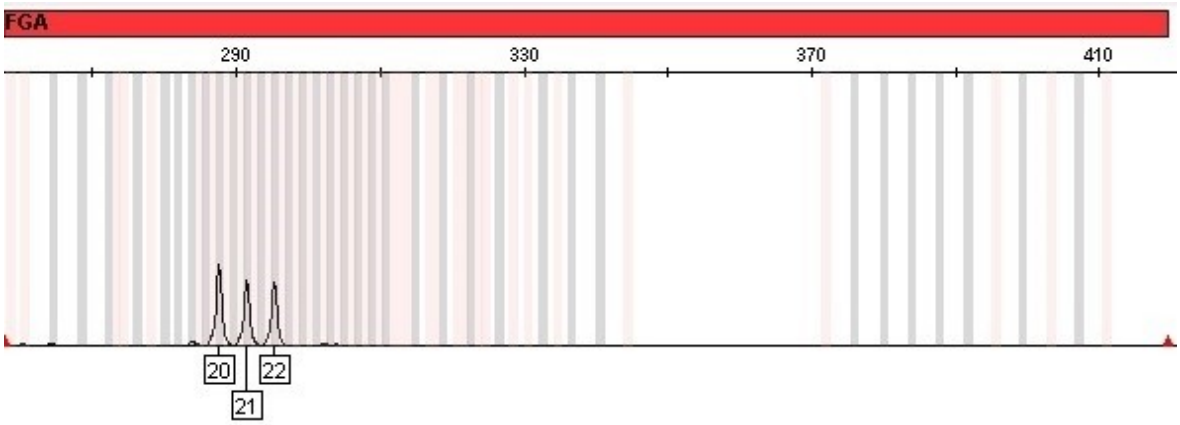


Figure 12: FGA-7 sample electropherogram – FGA STR locus with three alleles, 20, 21 and 22.

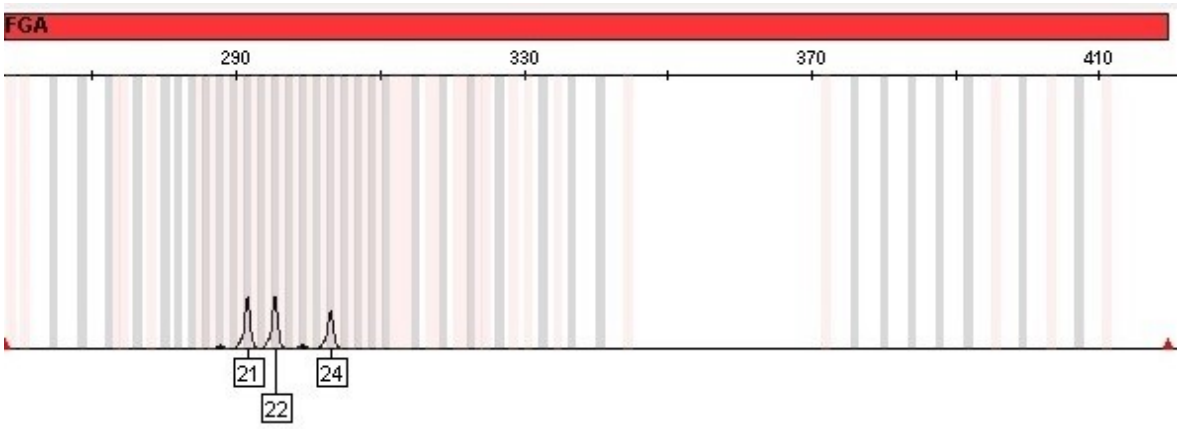


Figure 13: FGA-11 sample electropherogram – FGA STR locus with three alleles, 21, 22 and 24.

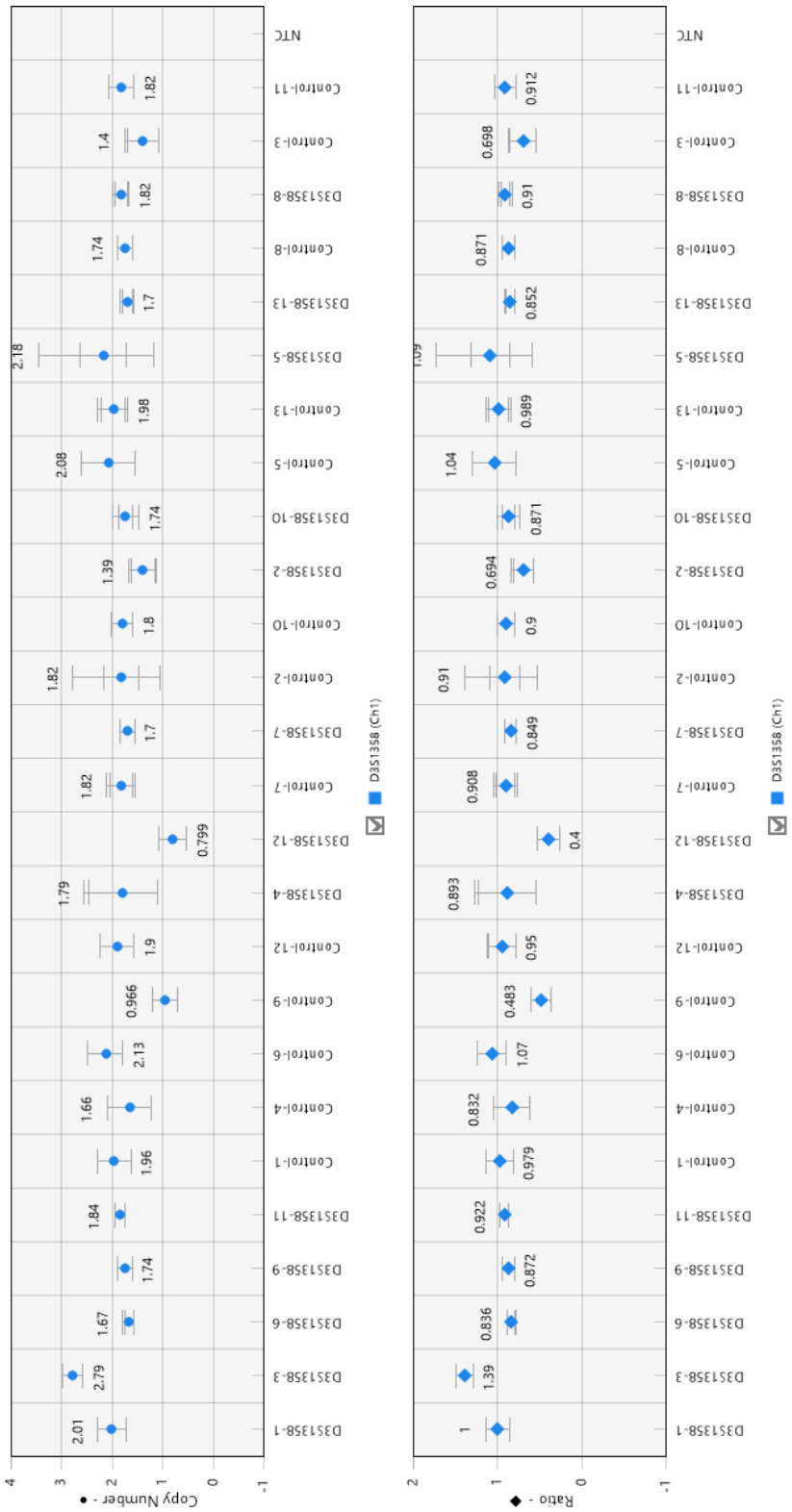
5.2.2 *D3S1358* STR locus testing results

Tested samples were divided onto two PCR plates, first plate containing 13 samples of interest, 13 control samples and NTC. Second plate contains 7 samples of interest, 6 control samples and NTC. Calculations of copy number and ratio *D3S1358 : AP3B1* are plotted on Graph 5 and Graph 6 for both PCR plates, respectively. One sample had measurement of 95% confidence interval >1, namely D3S1358-4 (Graph 5).

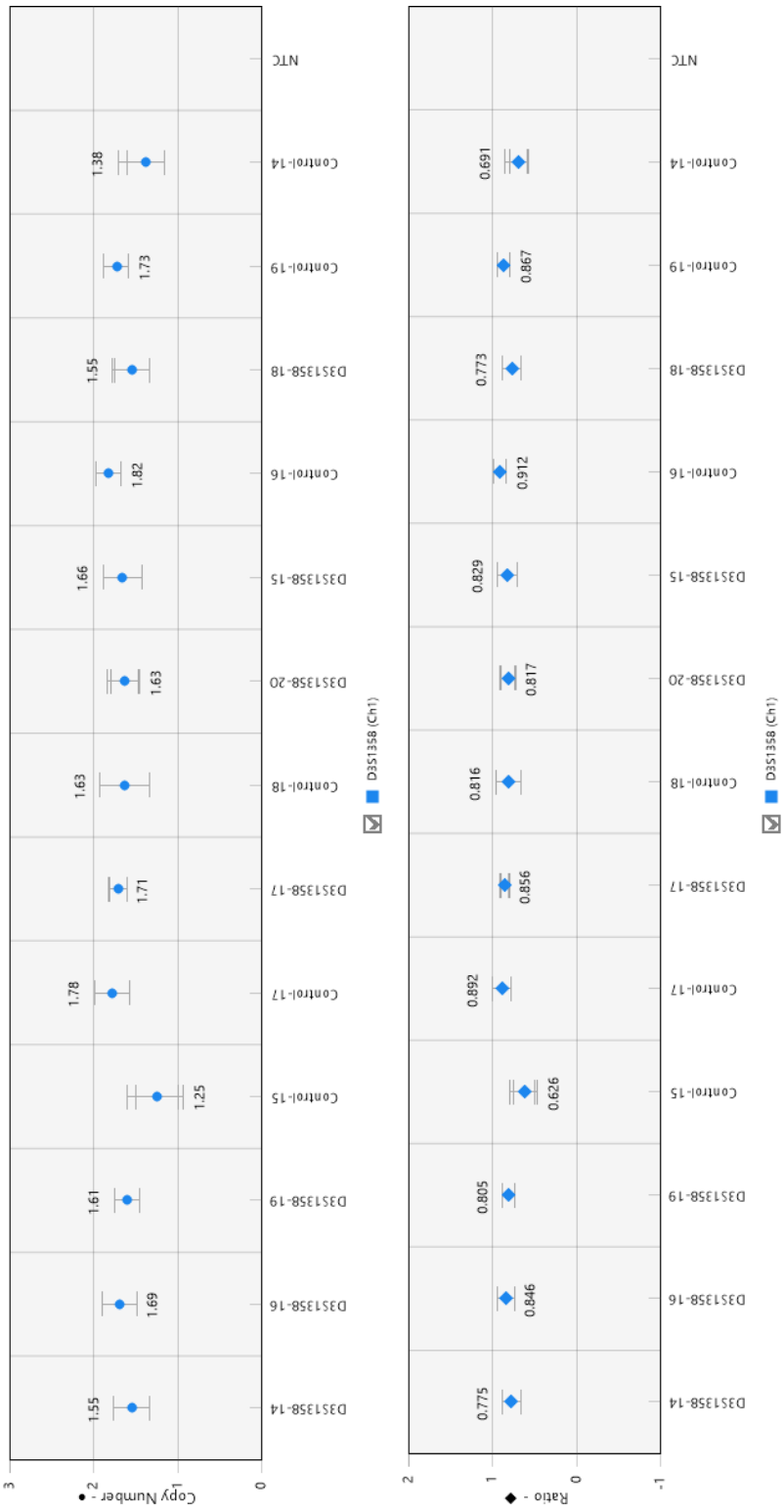
Average number of accepted droplets per well in both PCR plates was 15 381, while the average number of positive droplets per well was 247.

Table 11 shows calculated copy numbers for all the samples tested. Due to the copy number <1, samples D3S1358-12 and Control-9 were excluded from the datasets when population mean and standard deviation were calculated. Graph 7 plots distribution of copy numbers, with previously mentioned samples excluded from both datasets. Standard Z-score values for selected D3S1358 samples are calculated and presented in Table 12. Mann-Whitney U test calculations are shown in Table 13.

According to copy number calculations from Table xx and Z-score calculations, one sample, D3S1358-3 with potential tri-allelic *D3S1358* STR pattern is chosen and represented by electropherogram provided by Institute of Criminalistics in modified Figure 14.



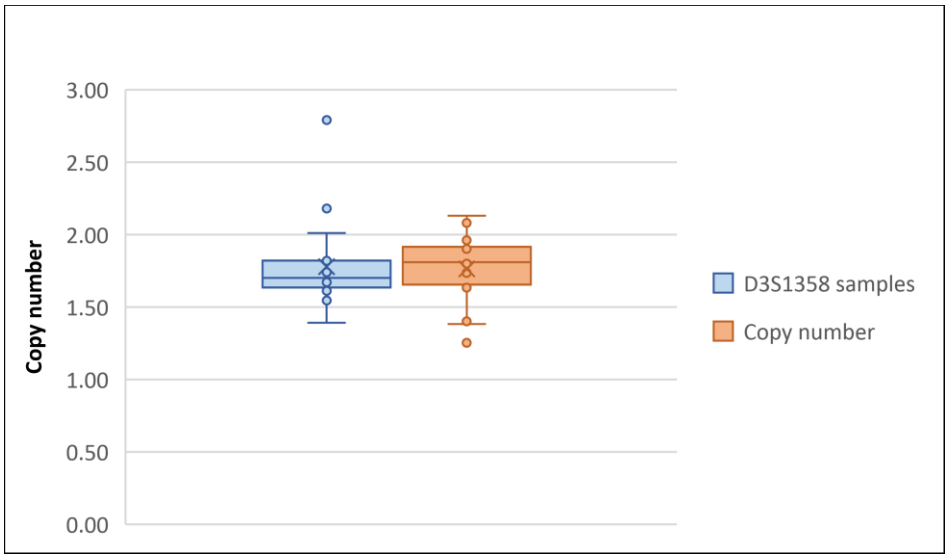
Graph 5: PCR Plate 1: Samples D3S1358-1 – D3S1358-13; Control-1 – Control-13; NTC Upper: Copy number calculation (Confidence interval 95%); Lower: Ratio D3S1358 : AP3B1.



Graph 6: PCR Plate 2: Samples D3S1358-14 – D3S1358-20; Control-14 – Control-19; NTC Upper: Copy number calculation (Confidence interval 95%); Lower: Ratio D3S1358 : AP3B1.

Table 11: Copy number values of samples of interest and control samples (Samples marked with red color indicate CNV > 2; samples marked with blue color indicate CNV < 1).

Sample	CNV	Control Sample	CNV
D3S1358-1	2.01	Control-1	1.96
D3S1358-2	1.39	Control-2	1.82
D3S1358-3	**2.79**	Control-3	1.4
D3S1358-4	1.79	Control-4	1.66
D3S1358-5	2.18	Control-5	2.08
D3S1358-6	1.67	Control-6	2.13
D3S1358-7	1.7	Control-7	1.82
D3S1358-8	1.82	Control-8	1.74
D3S1358-9	1.74	**Control-9**	**0.966**
D3S1358-10	1.74	Control-10	1.8
D3S1358-11	1.84	Control-11	1.82
D3S1358-12	**0.799**	Control-12	1.9
D3S1358-13	1.7	Control-13	1.98
D3S1358-14	1.55	Control-14	1.38
D3S1358-15	1.66	Control-15	1.25
D3S1358-16	1.69	Control-16	1.82
D3S1358-17	1.71	Control-17	1.78
D3S1358-18	1.55	Control-18	1.63
D3S1358-19	1.61	Control-19	1.73
D3S1358-20	1.63		
MEAN	**1.777**		**1.762**
STANDARD DEVIATION	**0.298**		**0.233**



Graph 7: Distribution of copy number values of D3S1358 samples of interest and control samples.

Table 12: Z-score values of samples of interest for D3S1358 STR testing (Samples marked with red color: Z-score > 3).

Sample	Z-score
D3S1358-1	1.07
D3S1358-2	-1.59
D3S1358-3	**4.42**
D3S1358-4	0.13
D3S1358-5	1.80
D3S1358-6	-0.39
D3S1358-7	-0.26
D3S1358-8	0.26
D3S1358-9	-0.09
D3S1358-10	-0.09
D3S1358-11	0.34
D3S1358-13	-0.26
D3S1358-14	-0.90
D3S1358-15	-0.44
D3S1358-16	-0.30
D3S1358-17	-0.21
D3S1358-18	-0.92
D3S1358-19	-0.64
D3S1358-20	-0.54

Table 13: Mann-Whitney U test calculation for D3S1358 samples.

Dataset	n	Mean of ranks	Sum of ranks	U-value, Z-score, p
D3S1358 samples	19	17.42	331	U = 141, Z = -0.89, p = 0.36
Control samples	18	20.67	372	

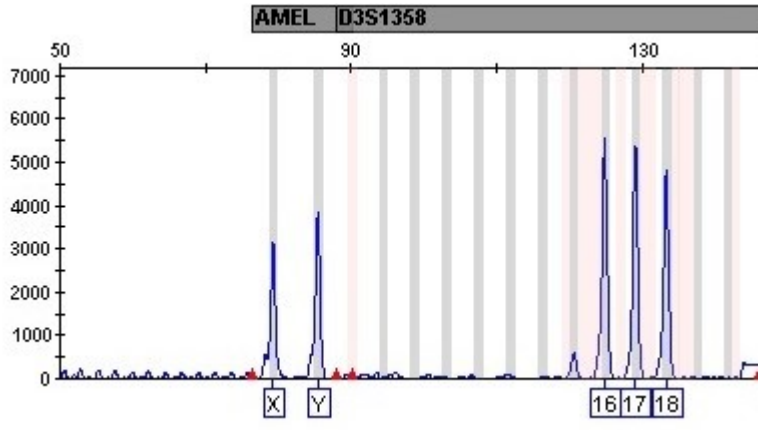


Figure 14: D3S1358-3 sample electropherogram – D3S1358 STR locus with three alleles, 16, 17 and 18.

6 Discussion

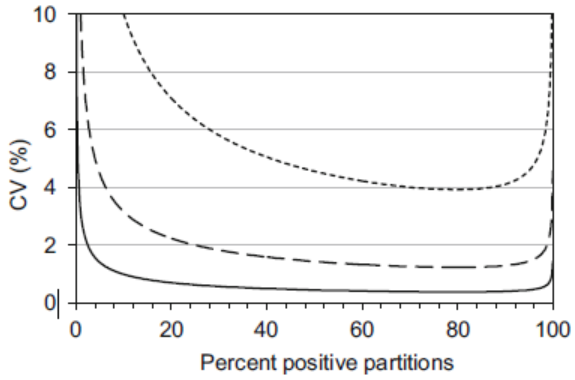
DdPCR optimization process involved two steps to obtain readable and understandable data suitable for copy number determination. Two main factors that influenced reaction were amount of input DNA and number of PCR cycles. It was confirmed that amount of 40 ng of input DNA is optimal and increase in number of cycles to 50 facilitated threshold determination between negative and positive droplets, especially for reference gene *AP3B1*-positive droplets.

DNA BUK001 sample which was used for optimization was isolated right before conducting the experiment, so the quality of DNA was presumably high and purity ratio A260/A280 was 1.88, which indicates pure DNA. Results in Graph 1 indicate optimized protocol for DNA extracted from reference sample, in our case buccal swab. Even in case of different reaction mixes that we have used in second optimization step, copy number was approaching integer number 2 in every reaction, and *FGA : AP3B1* ratio was close to integer 1.

Tables 3 and 4 with measured DNA concentrations of samples from Institute of Criminalistics Prague indicates extremely low A260/A280 purity ratio in tested samples. DNA in ICP laboratory was isolated by robotic extraction with DNA IQ™ System on Biomek® Laboratory Automatic Workstation, which is system using silica-coated magnetic beads to extract DNA, designed specifically for samples for forensic DNA profiling and paternity testing. They haven't reported any challenges with using this DNA in multiplex PCR and following capillary electrophoresis.

Since ddPCR is a technology known to be resistant to different PCR inhibitors that would normally cause problems in other PCR technologies, such as qPCR, it was assumed that absolute quantification of DNA will be unambiguous and copy numbers will be close or equal to integer values of 2, or 3 in case of copy number gain, which was the case with DNA BUK001, where the measured copy number was ~2 for every PCR setup. Nevertheless, values measured and represented in Results section for FGA and D3S1358 samples of interest and control samples indicate copy number values that are differently distributed, sometimes even approaching copy number 1.

Average number of accepted droplets in experiment with DNA BUK001 is comparable to average of accepted droplets in experiments with samples from ICP, however average number of positive droplets is almost tenfold higher in the case of BUK001 DNA quantification experiment. Effect of positive droplet partition on coefficient of variation (CV) and therefore precision of measurement, is shown in Graph 8. In case of *FGA* and *D3S1358* STR experiments, percentage of positive droplets in reaction were ~1.3% and 1.6%, respectively. This indicates that coefficient of variation for tested STR copy numbers may even exceed 10%.



Graph 8: Coefficient of variation (CV) based on percent of positive droplets per reaction in reactions containing 1000 (dotted line), 10000 (dashed line) and 100000 (solid line) accepted droplets. (Adapted from: Karlin-Neumann & Bizouarn, 2018).

Since these lower values of copy numbers and lower average number of positive droplets occured in both samples of interest and control samples datasets, we think that the reason may be variable DNA quality and degradation rate of samples, since they are obtained in period of 2009-2019.

Since copy number values for individual samples was not clear and unambiguous, statistical methods were applied to further analyze data. Z-score calculation for individual samples provided information about number of standard deviations away from the mean of copy number of control samples dataset. For *FGA* STR copy number quantification, Z-score values 2.89, 3.12 and 2.61 for samples FGA-2, FGA-7 and FGA-11, respectively, indicated possible copy number gain. For *D3S1358* STR copy number quantification, Z-score value 2.79 for sample D3S1358-3 indicated copy number gain. The value of the Z-score was not the only considered selection criterion for the seach for samples with potential target sequence duplications.

We evaluated also the extent of 95% confidence intervals keeping in mind the limited precision of our measurements performed on archive samples providing low numbers of positive droplets in analysis. Electropherograms of selected samples for both explored STR loci provided by ICP are represented in Results section and indicate presence of Type 2 1:1:1 tri-allelic pattern, as defined by Yang et al., (2020), for samples FGA-2, FGA-7, FGA-11 and D3S1358-3.

Mann-Whitney U test was used to measure possibility of sample with suspicion of tri-allelic pattern presence to have greater copy number than sample from control samples dataset. For FGA samples, mean of ranks in datasets of samples of interest and control samples were 21.77 and 21.2, respectively. The distributions in the two datasets did not differ significantly (Mann-Whitney U = 214, $n_1 = 22$, $n_2 = 20$, $Z = 0.13$, p = 0.88 two-tailed), therefore, we couldn't reject null hypothesis that there is no significant difference in copy numbers between datasets of samples of interest and control samples. For D3S1358 samples, mean of ranks in datasets of samples of interest and control samples were 17.42 and 20.67, respectively. The distributions in the two datasets did not differ significantly (Mann-Whitney U = 141, $n_1 = 19$, $n_2 = 18$, Z = -0.89, p = 0.36 two-tailed), for these datasets, we also could not reject null hypothesis.

Research limitations

First limitation that was encountered was the selection of samples from ICP. Since tri-allelic patterns are rare phenomenon, and out of ~ 200 000 DNA profiles from National DNA Database of Czech Republic, we have found 22 samples for *FGA* and 20 samples for *D3S1358* STR locus testing which is number too low to further obtain any population-relevant data. Obtained samples were of very low purity, volume, and possibly severely degraded and any further manipulation, such as purification of samples, could result in massive losses.

Second limitation was primer and probe design for STR loci. It is a challenging task since DNA polymerase is error-prone when it comes to amplification of repetitive sequences. For that reason, our primers and probes were designed to attach right next to the STR locus of interest. This means that possible mutation in the region selected for primer and probe annealing, or duplication event in which this region is not duplicated, not duplicated as a

whole, but only partially, or duplicated, translocated and then inversed, leads to underestimated copy number quantification of STR locus of interest.

7 Conclusion

We assumed that ddPCR technology is suitable choice for confirmation of results indicating tri-allelic pattern on STR marker on capillary electrophoresis. The DNA sample that we used for reaction optimization indicated value that closely approached integer copy number 2 in all the reactions, while results of samples of interest and control samples from ICP resulted in data with great variability, leading to more complex interpretation and analysis. Nonetheless, after statistical calculations, we were able to select 3 samples out of 22 in which there is indication of presence of third allele for *FGA* STR locus, and 1 sample out of 20 for *D3S1358* locus.

We assumed that ddPCR technology could overcome limitations of CGH array technology described by Repnikova et al., (2013) and quantify CNVs involving STR markers with greater precision. According to data obtained from our experiments, ddPCR technology could be suitable for copy number quantification in case of pure and non-degraded DNA samples. Lower purity ratio and DNA degradation negatively affect ddPCR amplification which results in low number of positive droplets, which in turn limits the precision of copy number measurements.

Goal of this Master's thesis was to optimize ddPCR protocol for detection of copy number variants in the regions of STR loci. In spite of limitations regarding DNA quality and purity of tested samples, ddPCR protocol suitable for detection of mentioned variants from DNA isolated from buccal swab was established. Further analysis of selected samples by next-generation sequencing could reveal more reliable information about copy number gains in STR regions.

8 References

AP3B1. (2022). https://www.ncbi.nlm.nih.gov/gene/8546

Beckmann, J. S., Estivill, X., & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, *8*(8), 639–646. https://doi.org/10.1038/nrg2149

Bodner, M., Bastisch, I., Butler, J. M., Fimmers, R., Gill, P., Gusmão, L., Morling, N., Phillips, C., Prinz, M., Schneider, P. M., & Parson, W. (2016). Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). *Forensic Science International: Genetics*, *24*, 97–102. https://doi.org/10.1016/j.fsigen.2016.06.008

Buckleton, J. S., Bright, J. A., & Taylor, D. (2016). *Forensic DNA Evidence Interpretation*. Taylor \& Francis. https://books.google.cz/books?id=dH90DwAAQBAJ

Butler, J. M. (2010a). *Chapter 10 - STR Genotyping and Data Interpretation* (J. M. B. T.-F. of F. D. N. A. T. Butler (Ed.); pp. 205–227). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374999-4.00010-2

Butler, J. M. (2010b). *Chapter 8 - Short Tandem Repeat Markers* (J. M. B. T.-F. of F. D. N. A. T. Butler (Ed.); pp. 147–173). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374999-4.00008-4

Butler J.M., Vallone P.M., Gettings K.B., Borsuk L.A., Ruitberg C.M., R. D. J. (2017). *NIST Short Tandem Repeat DNA Internet Database, National Institute of Standards and Technology*. https://doi.org/10.18434/T44G6P

Gettings, K. B., Aponte, R. A., Vallone, P. M., & Butler, J. M. (2015). STR allele sequence variation: Current knowledge and future issues. *Forensic Science International: Genetics*, *18*, 118–130. https://doi.org/https://doi.org/10.1016/j.fsigen.2015.06.005

Goodwin, W., Linacre, A., & Hadi, S. (2007). *An Introduction to Forensic Genetics*. John Wiley \& Sons. https://books.google.cz/books?id=7wnivFnxel8C

Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics and Development*, *44*, 9–16. https://doi.org/10.1016/j.gde.2017.01.012

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Science International. Genetics*, *17*, 33–34. https://doi.org/10.1016/j.fsigen.2015.03.006

Härmälä, S. K., Butcher, R., & Roberts, C. H. (2017). *Copy Number Variation Analysis by Droplet Digital PCR BT - Functional Genomics: Methods and Protocols* (M. Kaufmann, C. Klinger, & A. Savelsbergh (Eds.); pp. 135–149). Springer New York. https://doi.org/10.1007/978-1-4939-7231-9_9

Henrichsen, C. N., Chaignat, E., & Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics*, *18*(R1), R1–R8. https://doi.org/10.1093/hmg/ddp011

Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., Bright, I. J., Lucero, M. Y., Hiddessen, A. L., Legler, T. C., Kitano, T. K., Hodel, M. R., Petersen, J. F., Wyatt, P. W., Steenblock, E. R., Shah, P. H., Bousse, L. J., Troup, C. B., Mellen, J. C., … Colston, B. W. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, *83*(22), 8604–8610. https://doi.org/10.1021/ac202028g

Hindson, C. M., Chevillet, J. R., Briggs, H. A., Gallichotte, E. N., Ruf, I. K., Hindson, B. J., Vessella, R. L., & Tewari, M. (2013). Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nature Methods*, *10*(10), 1003–1005. https://doi.org/10.1038/nmeth.2633

Iskow, R. C., Gokcumen, O., & Lee, C. (2012). Exploring the role of copy number variants in human adaptation. *Trends in Genetics : TIG*, *28*(6), 245–257. https://doi.org/10.1016/j.tig.2012.03.002

Jiao, H., Ren, H., Yang, Y., Ni, B., Zhou, H., Chen, W., Cao, Y., Chen, C., Huang, Y., & Yan, J. (2018). Tri-allelic patterns of STRs and partially homologous non-sister chromatid crossover observed in a parentage test. *Legal Medicine*, *30*, 34–37.

Karlin-Neumann, G., & Bizouarn, F. (2018). *Digital PCR : Methods and Protocols* (G. Karlin-Neumann & F. Bizouarn (Eds.); First). Humana New York, NY. https://doi.org/https://doi.org/10.1007/978-1-4939-7778-9

Lauer, S., & Gresham, D. (2019). An evolving view of copy number variants. *Current*

Genetics, *65*(6), 1287–1295. https://doi.org/10.1007/s00294-019-00980-0

Lucena-Aguilar, G., Sánchez-López, A. M., Barberán-Aceituno, C., Carrillo-Ávila, J. A., López-Guerrero, J. A., & Aguilar-Quesada, R. (2016). DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis. *Biopreservation and Biobanking*, *14*(4), 264–270. https://doi.org/10.1089/bio.2015.0064

Madsen, B. E., Villesen, P., & Wiuf, C. (2010). *Short Tandem Repeats and Genetic Variation BT - Genetic Variation: Methods and Protocols* (M. R. Barnes & G. Breen (Eds.); pp. 297–306). Humana Press. https://doi.org/10.1007/978-1-60327-367-1_16

Nowakowska, B. (2017). Clinical interpretation of copy number variants in the human genome. *Journal of Applied Genetics*, *58*(4), 449–457. https://doi.org/10.1007/s13353-017-0407-4

Phillips, C., Gettings, K. B., King, J. L., Ballard, D., Bodner, M., Borsuk, L., & Parson, W. (2018). "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Science International: Genetics*, *34*, 162–169. https://doi.org/https://doi.org/10.1016/j.fsigen.2018.02.017

Picanço, J. B., Raimann, P. E., Paskulin, G. A., Alvarez, L., Amorim, A., Batista dos Santos, S. E., & Alho, C. S. (2014). Tri-allelic pattern at the TPOX locus: A familial study. *Gene*, *535*(2), 353–358. https://doi.org/https://doi.org/10.1016/j.gene.2013.10.019

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., & Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*, *44*(5), 548–559. https://doi.org/https://doi.org/10.1016/j.bj.2021.02.003

Repnikova, E. A., Rosenfeld, J. A., Bailes, A., Weber, C., Erdman, L., McKinney, A., Ramsey, S., Hashimoto, S., Lamb Thrush, D., Astbury, C., Reshmi, S. C., Shaffer, L. G., Gastier-Foster, J. M., & Pyatt, R. E. (2013). Characterization of copy number variation in genomic regions containing STR loci using array comparative genomic hybridization. *Forensic Science International: Genetics*, *7*(5), 475–481. https://doi.org/https://doi.org/10.1016/j.fsigen.2013.05.008

Ruitberg, C. M., Reeder, D. J., & Butler, J. M. (2001). STRBase: a short tandem repeat DNA

database for the human identity testing community. *Nucleic Acids Research*, *29*(1), 320–322. https://doi.org/10.1093/nar/29.1.320

Saitou, M., & Gokcumen, O. (2020). An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. *Journal of Molecular Evolution*, *88*(1), 104–119. https://doi.org/10.1007/s00239-019-09911-6

Schneider, P. M. (2009). Expansion of the European Standard Set of DNA Database Loci—the Current Situation. *Profiles DNA*, *12*.

Shrivastava, P., Jain, T., & Kumawat, R. K. (2021). Direct PCR amplification from saliva sample using non-direct multiplex STR kits for forensic DNA typing. *Scientific Reports*, *11*(1), 7112. https://doi.org/10.1038/s41598-021-86633-0

Šimková, H., Faltus, V., Marvan, R., Pexa, T., Stenzl, V., Brouček, J., Hořínek, A., Mazura, I., & Zvárová, J. (2009). Allele frequency data for 17 short tandem repeats in a Czech population sample. *Forensic Science International: Genetics*, *4*(1), e15–e17. https://doi.org/https://doi.org/10.1016/j.fsigen.2009.01.003

Sismani, C., Koufaris, C., & Voskarides, K. (2015). *Copy Number Variation in Human Health, Disease and Evolution BT - Genomic Elements in Health, Disease and Evolution: Junk DNA* (K. Felekkis & K. Voskarides (Eds.); pp. 129–154). Springer New York. https://doi.org/10.1007/978-1-4939-3070-8_6

Statistics, S. S. (2022). *Mann-Whitney U Test Calculator*. https://www.socscistatistics.com/tests/mannwhitney/

Tan, C., Chen, X., Wang, F., Wang, D., Cao, Z., Zhu, X., Lu, C., Yang, W., Gao, N., Gao, H., Guo, Y., & Zhu, L. (2019). A multiplex droplet digital PCR assay for non-invasive prenatal testing of fetal aneuploidies. *The Analyst*, *144*(7), 2239–2247. https://doi.org/10.1039/c8an02018c

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115. https://doi.org/10.1093/nar/gks596

Vidal, C., & Cassar, M. (2008). A case of tri-allelic pattern at locus D3S1358 on chromosome 3p21 inherited from paternal grandmother. *Forensic Science*

International: Genetics, *2*(4), 372–375. https://doi.org/https://doi.org/10.1016/j.fsigen.2008.06.002

Wang, L.-F., Yang, Y., Zhang, X.-N., Quan, X.-L., & Wu, Y.-M. (2015). Tri-allelic pattern of short tandem repeats identifies the murderer among identical twins and suggests an embryonic mutational origin. *Forensic Science International: Genetics*, *16*, 239–245. https://doi.org/https://doi.org/10.1016/j.fsigen.2015.01.010

Wyner, N., Barash, M., & McNevin, D. (2020). Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype . In *Frontiers in Genetics* (Vol. 11, p. 884). https://www.frontiersin.org/article/10.3389/fgene.2020.00884

Yang, Q., Shen, Y., Shao, C., Liu, Y., Xu, H., Zhou, Y., Liu, Z., Sun, K., Tang, Q., & Xie, J. (2020). Genetic analysis of tri-allelic patterns at the CODIS STR loci. *Molecular Genetics and Genomics*, *295*(5), 1263–1268. https://doi.org/10.1007/s00438-020-01701-w

Yang, Q., Yao, Y., Shao, C., Zhou, Y., Li, H., Li, C., Tang, Q., & Xie, J. (2021). Calculation of the Paternity Index for STR with tri-allelic patterns in paternity testing. *Forensic Science International*, *324*, 110832. https://doi.org/https://doi.org/10.1016/j.forsciint.2021.110832

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*(3), 172–183. https://doi.org/10.1038/nrg3871