UiT
THE ARCTIC UNIVERSITY OF NORWAY

Department of Medical Biology, Faculty of Health Sciences

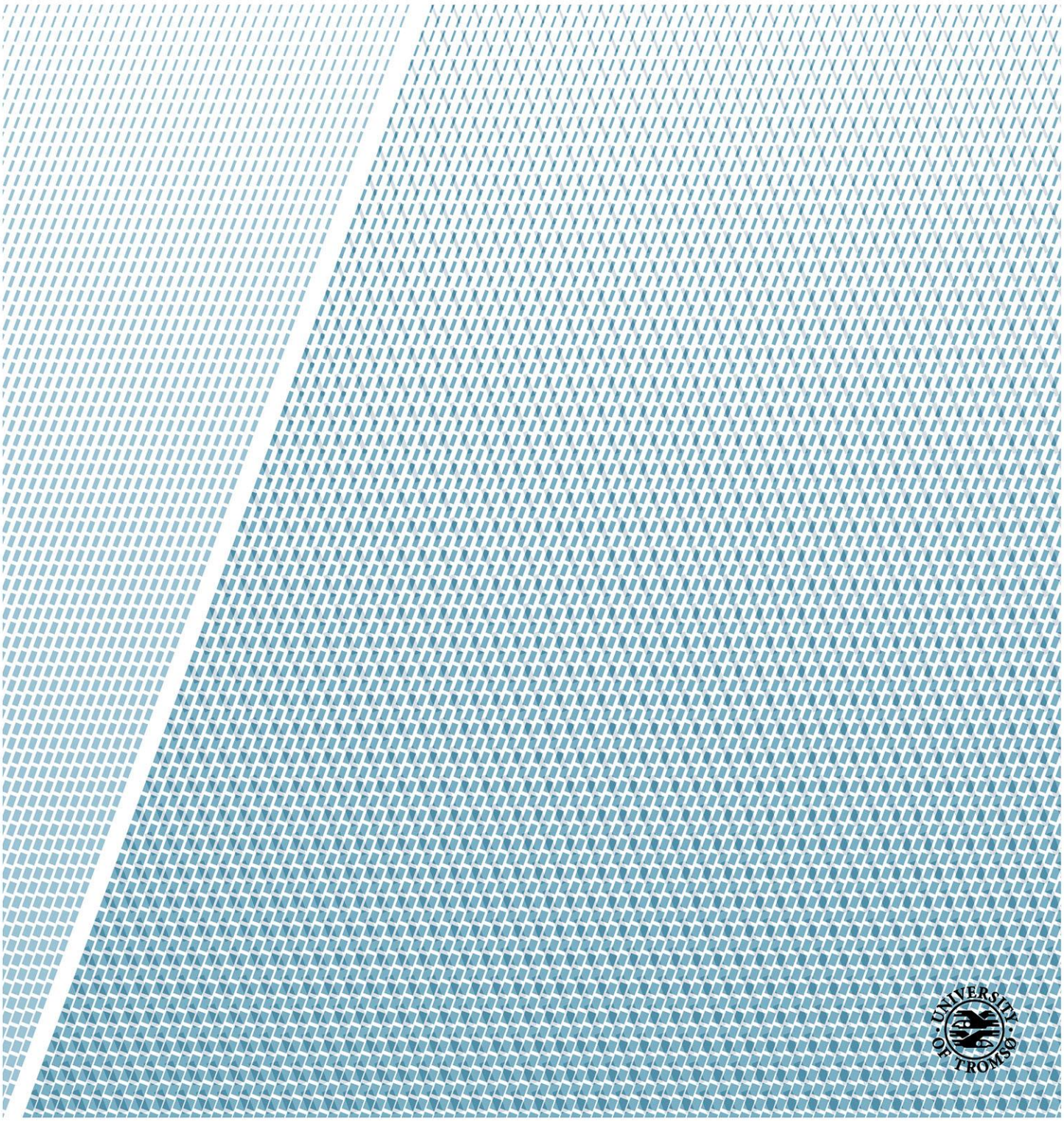
Forensic DNA genotyping by means of next generation sequencing

Analysis of Autosomal STRs of a Norwegian population sample using the ForenSeq FGx system

—

Sandra Buadu

Master thesis in Biomedicine MBI-3911. May 2018



Acknowledgments

This work for this master thesis was carried out at the Centre of Forensic Genetics, The Arctic University of Norway (UiT) from august 2017 to may 2018 under the primary supervision of Kirstin Janssen.

I would first like to direct thanks my supervisor, Kirstin Janssen, for all the support, guidance and proofreading of the thesis in this period. I really appreciate all the tremendous help I have received before and during this process of writing the thesis, such as laboratory guidance and especially for characterizing the allele frequencies for me.

And to the employees at the Forensic Genetics Centre, especially Gunn-Hege Olsen and Thomas Berg my co-supervisors, I would like to thank you all for your kindness during these two years and for great laboratory guidance and always being willing to answer my questions.

I would also like to thank the department of Clinical Pathology, UNN for being so kind in lending me The Qubit Fluorometer for analysis.

And a special thanks to Illumina, Richard Kessel, for help and advice with any issues that might have come during this process.

Finally, I would like to thank my fellow classmates for two wonderful years of master studies.

Sandra Buadu

Tromsø, May 2018

TABLE OF CONTENTS

I ABBREVIATIONS

AIMs	Ancestry Informative Markers SNPs
CE	Capillary electrophoresis
CODIS	Combined DNA Index System
C_T- value	Threshold Cycle Value
ddNTPs	dideoxynucleotide triphosphates
DI	Degradation Index
dNTP	deoxynucleotide triphosphates
DPA	Norwegian Data Protection Authority
DPMA	DNA primer mix A
DPMB	DNA primer mix B
ENFSI	European Network of Forensic Science Institutes
ESS	European Standard Set
FGC	Centre of Forensic Genetics
GD	Genetic diversity
Hobs	Observed Heterozygosity
HSC	Human Sequencing Control
HT1	Hybridization Buffer
HWE	Hardy-Weinberg Equilibrium
IPC	Internal PCR Control
LA	Large Autosomal Target
LE	Linkage Equilibrium
LNA1	Library Normalization Additives 1
LNB1	Library Normalization Beads 1
LNS2	Library Normalization Storage Buffer 2
LNW1	Library Normalization Wash buffer 1
LR	Likelihood Ratio
MPS	Massive Parallel Sequencing
mtDNA	Mitochondrial DNA
NGS	Next Generation Sequencing
PCA	Principal Component Analysis

PCR	Polymerase Chain Reaction
PD	Power of discrimination
PE	Power of exclusion
PIC	Polymorphism Information Content
PM	Match probability
RFLP	Restriction Fragment Length Polymorphism
RMNE	Random Man Not Excluded
RMP	Random Match Probability
RSB	Resuspension Buffer
SA	Small Autosomal Target
SBS	Sequencing by Synthesis
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
STRAF	STR Analysis for Forensics
TPI	Typical paternity index
UAS	Universal Analysis Software
VNTR	Variable Number of Tandem Repeat

II ABSTRACT

Population databases containing allele frequencies of the genetic markers used for DNA-profiling are necessary for forensic geneticist to be able to perform statistical calculations on the statistical weight of DNA-evidence. However, allele frequencies differ from population to population, it is therefore important to establish population databases for specific graphical areas or population groups. The Center of Forensic Genetics is currently using a method based on PCR and capillary electrophoresis for DNA-profiling, but wants to establish a method based on deep sequencing.

The purpose of this study was therefore to establish a population database for a Norwegian population with allele frequencies of autosomal STR markers used for DNA-profiling with the new NGS-based method. Samples from a previously established biobank were used to obtain DNA profiles for all 231 forensic genetic markers included in the ForenSeq DNA Signature Prep kit. Validation data of the ForenSeq FGx system was also assessed, focusing only on the 27 autosomal STR included in the kit.

The Norwegian population database for autosomal STRs was established, with frequencies of both length-based and sequence-based allele variants. There is an increase in the number of sequence-based allele variants compared to length-based allele variants for many markers, meaning the power of discrimination can be raised when using the same number of markers when using a deep sequencing method. A reproducibility and sensitivity study of the ForenSeq FGx system was also conducted. They showed that the system produces 100% reproducible genotypes when sequencing the same samples more than once. The sensitivity study showed that with the ForenSeq FGx system a complete DNA-profile can be obtained for 125 pg DNA, and that approximately 50 % of alleles can still be called with as little as 15.625 pg DNA. To check if the two methods, Signature Prep and the NGM Select, could produce the same autosomal genotype results, a concordance study was performed. Almost full concordance was found between the two methods, only three alleles had discordances. The discordance is probably because of two different primers being used for the kits. They might been bound differently an therefor called different autosomal genotypes.

1 INTRODUCTION

In a small town in Northern Norway, a young girl has been found murdered and left in a field. After some fieldwork, the crime scene technicians and police have a prime suspect named Mr Petersen, a 44-year-old man who lives not far from the crime scene. Upon further investigation, they now know that his DNA profile matches the mixed DNA profile obtained from skin cells underneath the young girl's fingernails. The question remains, did in fact MR Petersen contribute this DNA sample. How many other people could have a DNA-profile matching the DNA-profile of the stain? And what is the probability for that? To answer these questions, the statistical weight of the evidence is calculated using allele frequencies for the genetic markers included in the DNA profile from a relevant population database.

In 2016, the Centre of Forensic Genetics (FGC), Institute of Medical Biology, Faculty of Health Sciences (UiT) started a research biobank and began to collect data to build a Norwegian population database with allele frequencies of the genetic markers used for DNA-profiling. This study will complete the database so that the obtained allele frequencies can be used to calculate the statistical weight of DNA-evidence to answer the questions above mentioned.

The human body consist of approximately 100 trillion cells, and each of these contains DNA with genetic information unique to each individual (1). The unique information found within each individuals DNA can be used in correlation to criminal cases. DNA analyses of biological samples are conducted with the purpose of characterizing single noncoding sites in the DNA of an individual. A DNA profile is the collection of these characterized sites in the DNA. The DNA profiles become unique to each individual if enough sites (genetic markers) are included into the profile, therefor individuals can be identified though their DNA profile.

To determine if an individual might be involved in a crime, their DNA profile can be compared to the DNA profile obtained from a crime scene sample. An individual may be connected to a case due to different involvements. The individual might for example be the perpetrator, the victim or the police officer working on the particular case (2).

Today, the Short Tandem Repeats (STRs) are the genetic markers most widely used in forensic genetics. However, other methods have been previously used until the end of the 1990s, such as ABO blood groups and DNA fingerprinting with Restriction Fragment Length Polymorphism (RFLP).

In 1990, the Austrian researcher Karl Landsteiner at the University of Vienna discovered that blood from different people would occasionally clump together. This lead to his eventual identification of the four blood types A, B, AB and O. The ABO blood group system was the first genetic evidence used to identify individuals in court in 1915. Professor Leone Lattes at the institute of Forensic Medicine in Turin developed methods for typing dried bloodstains with antibodies for the ABO blood groups. His method spread throughout Europe and to the United States, and over the next decades the ABO typing method was used in forensic cases and paternity disputes. However, large amounts of blood were needed for samples to be analysed, and the marker had low discriminatory power considering that there are only a few blood types in a population. In addition, the genetic markers were very susceptible to environmental degradation (1).

In 1984, the British geneticist Sir Alec Jeffreys discovered a region in the chromosomes that were built of blocks of repetitive DNA, like a barcode. The blocks where present in all humans and specific in length to each individual. This meant that they could be used to distinguish between two people, similar to fingerprints. These areas were therefore titled DNA fingerprints. The repetitive blocks of DNA were later known as Variable Number of Tandem Repeats (VNTRs). Jeffreys used a method called Restriction Fragment Length Polymorphism (RFLP) to analyse them (3, 4).

Since the DNA fingerprints were relatively unique between individuals, Sir Alec Jeffreys thought they could be useful in criminal cases. The first criminal case was solved in 1983 by DNA fingerprinting. A 15-year-old girl was found raped and murdered in Leicestershire. Although a semen sample was retrieved from the body, all investigation came to a halt and the case went cold. Three years later, another 15-year-old girl was found raped and murdered with a semen sample present on the body. Initially, the prime suspect was 17-year-old Richard Burkland, who under questioning admitted he was responsible for the second crime and had knowledge of the body. However, with Sir Alec Jeffreys' DNA-fingerprinting method, it was discovered that the two semen samples from the cases in 1983 and now in 1987, came from the same individual and that the DNA did not match Burkland. Therefore, the Leicestershire Constabulary and the Forensic Science Service conducted a large-scale search to find the perpetrator. Approximately 5000 local men were asked to give a blood and saliva samples, but none of the samples matched the semen samples. Later it was discovered that a man named Colin Pitchfork had payed someone to give a DNA sample in his name. He was arrested was the discovery that his DNA-fingerprint perfectly matched the DNA-fingerprints of the semen

samples found at the crime scenes. He confessed to the two crimes and was sentenced to life imprisonment. This was the first criminal case solved with DNA evidence. DNA profiling as a valuable tool for solving crimes was hereby established (1, 5).

The first case in Norway was solved by DNA-analysis shortly after, in 1989. 17-year-old Inger Lise Olsen was raped and murdered in Mysen. After several weeks the police did not yet have a suspect. They then asked male residents to provide a blood sample in large-scale search. Based on these samples an 18-year-old man was convicted on the basis of a DNA match (6, 7).

The RFLP method used by Jeffreys consists of DNA samples being fragmented by restriction enzymes. The enzymes recognize specific nucleotide sequences in the DNA samples and cut the DNA strands. Then the fragments are separated according to their lengths by gel electrophoresis, see figure 1. RFLP as a method requires large amounts of DNA. This factor made DNA fingerprinting difficult in cases with low amount and/or degraded DNA samples. With the invention of Polymerase Chain Reaction (PCR) by Kary Mullis (8), only a drop of blood is needed to successfully obtain a complete DNA profile.

DNA itself is a very stable molecule, so it can easily be typed accurately even if blood and bodily fluids are degraded. RFLP of DNA samples was replaced by PCR of Short Tandem Repeat markers (STRs). The obtained DNA fragments were visualized by electrophoresis which has been further developed into today's capillary electrophoresis (CE) (3, 9-11). Today's DNA profiling method was largely developed in thanks to these two independent breakthroughs in molecular biology by Professor Sir Alec Jeffreys and Kary Mullis. One of the most exciting DNA profiling methods today is Next Generation Sequencing (NGS), a method that will be further explained later on in the thesis.

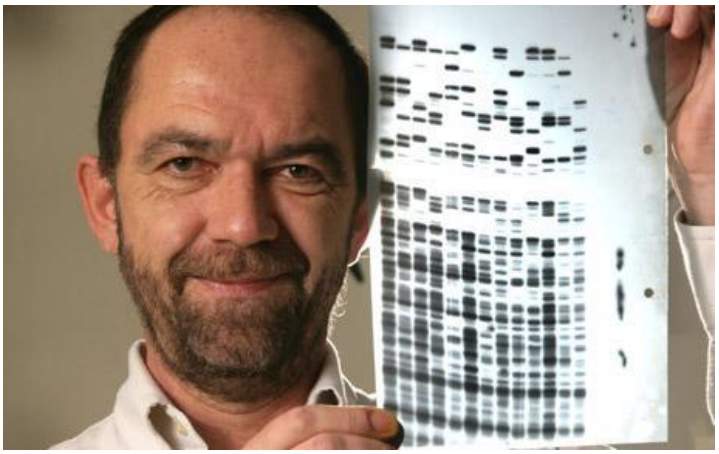


Figure 1. *The human DNA fingerprint, achieved by analysing specific DNA segments with Restriction Fragment Length Polymorphism. In this picture, Sir Alec Jeffreys is holding up a film that has been exposed over the gel. Modified from: (3).*

1.1 STRS AND DNA TYPING

The human genome consists of about 3 million base pairs, and although it is large, approximately 5% contains genetically relevant information also known as the gene-coding DNA. The other ~95% contains non-coding DNA and is in some parts built up of repeated sequences. Although the human genome is largely similar between individuals there are still parts with enough diversity to be able to distinguish between people. Short Tandem Repeats are simple repeated blocks of DNA, which are highly polymorphic spots in the non-coding DNA regions. STRs consist of short DNA-motives of 2-7 base pairs in length that are typically repeated 5-50 times and they are located around the centromere of the chromosomes (1, 12-14). STRs used in forensic genetics are mainly tetranucleotides, which are sequence motives of 4 base pairs. The chosen STR-markers are spread over as many chromosomes as possible to ensure that they are inherited independently. The independence of the STR is important when performing statistical analysis. If the STRs are closely linked, they may not be randomly distributed throughout a population.

STRs have different qualities that make them suitable for human identification. The STR alleles vary among individuals and some are more common than others. Allele frequencies vary between populations, making STRs quite effective in separating individuals also between populations. Forensic DNA evidence can often be quite degraded and it can be challenging to obtain PCR amplification products from them. The STRs however, are small in size, which

makes them ideal targets for use in forensic genetic. The data obtained by analysing STR is rather stable and predictable because STR alleles have low mutation rates (15, 16).

Each individual inherits one STR copy from each parent, therefore the locus will show two possible alleles (17). If the two copies have the same repeat numbers, the individual will be homozygote for that marker. If the two copies inherited have different repeat numbers, the individual is heterozygote for that particular marker. Homozygote and heterozygote STR loci are shown under STR loci 1 and 2 in figure 2, respectively. By examining enough STR markers each individual will obtain a specific DNA profile (Fig. 3) which will distinguish them from others (13). A DNA profile consists of all the allele lengths numbers of each included STR marker. In figure 3, the DNA profile is visualized by an electropherogram. Each STR marker has one or two peaks (heterozygous or homozygous), which visualises the alleles. The number beneath these indicates the allele number, which is equivalent to the number of times the STR tetranucleotide motif is repeated. For example, the individual in figure 3 is heterozygous for STR marker Penta E with allele 12 and 13.

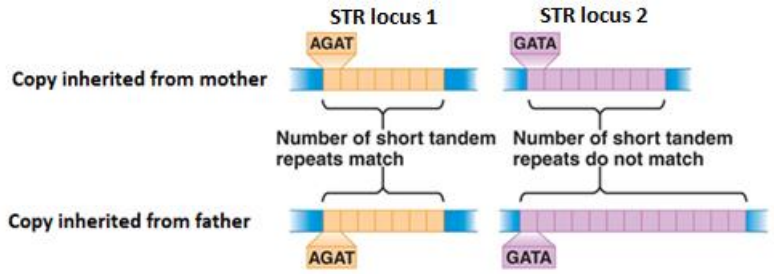


Figure 2. *Two different STR loci in the DNA profile of an individual, one with an equal number of repeats and the other with different number of repeats in the two copies inherited from the individual's parents. This individual has inherited two alleles 7 at STR locus 1, so it is homozygous at this locus. STR locus 2 contains alleles 8 and 13, so the individual is heterozygous at this locus. Modified from: (18).*

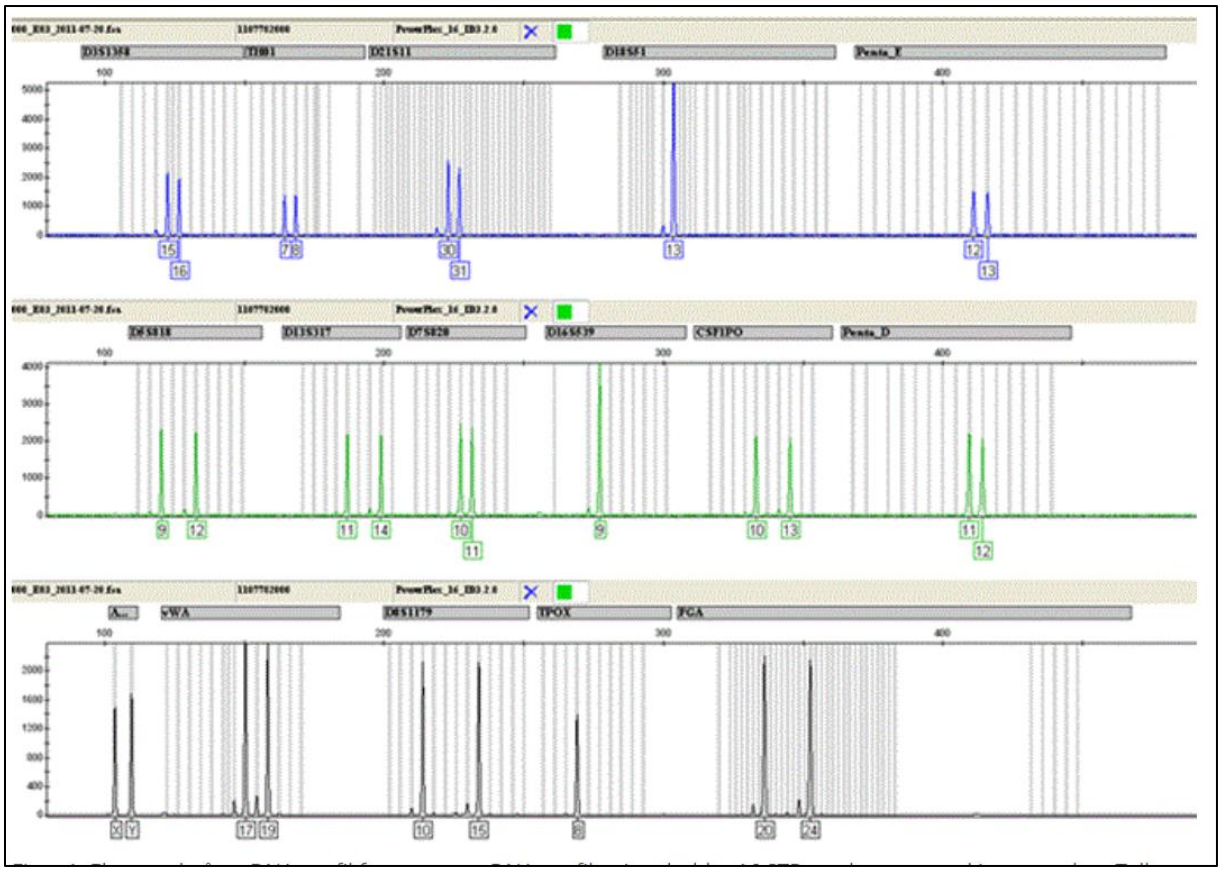


Figure 3. *An electropherogram of a DNA profile containing 15 STR markers and one sex marker. Modified from (19).*

Replication errors known as replication slippage can lead to mutations in STRs. During replication the DNA polymerase dissociates (slips) from the DNA template and anneals to homologous sequences nearby. This error is prone to happen in regions of repetitive DNA, leading to deletions or insertions of a repeat in the sequence (20, 21). Therefore, STRs have usually length polymorphism, but point mutations altering the sequence within one of the repeats or the flanking region may also occur. Any changes that do not alter the length of the fragment are not possible to detected by electrophoresis, but only by sequencing. NGS is a method that can detect these alterations, by characterizing sequence variations within alleles. Table 1 shows different variants of the D12S391 marker with the same allele number, but different sequences. The detection of these sequence differences can be helpful in further identifying individuals and raising the power of discrimination.

Table 1. *Example of different alleles in the STR marker D12S391 with the same fragment length (allele number 21), but different sequences.*

D12S391[21]AGAT[11]AGAC[9]AGAT[1]
D12S391[21]AGAT[11]AGAC[10]
D12S391[21]AGAT[12]AGAC[8]AGAT[1]
D12S391[21]AGAT[12]AGAC[9]
D12S391[21]AGAT[13]AGAC[7]AGAT[1]
D12S391[21]AGAT[13]AGAC[8]
D12S391[21]AGAT[13]GGAC[1]AGAC[7]
D12S391[21]AGAT[14]AGAC[6]AGAT[1]

DNA profiles can be stored in a DNA database. The DNA database contains a collection of computer files with DNA profiles obtained from crime scenes or DNA profiles that are connection to these. In Norway, there are currently three different DNA databases/registers. The investigation register contains the DNA profiles of individuals and trace sample evidence under investigation. Second, the identity register containing the DNA profiles of convicted individuals. Third, the trace sample register containing the DNA profiles of crime scene samples not yet identified. DNA databases have been and can be very useful in solving cases. They can be used to connect serial crimes, as well as resolve cases in which there initially have been no suspects. Unjustly incarcerated or charged individuals can be exonerated when the real offender might show up in the database later in connection to another crime. The DNA profiles in the investigation register are either deleted if the individual investigated is acquitted or transferred to the identity register if they are convicted. Unsolved trace samples are transferred to the trace sample register, and the different registers can be searched against each other. The databases make the connection between cases and DNA profiles (22).

To be able to compare unknown crime scene DNA profiles and search databases between countries and laboratories, there is a need to agree on a set of common STR markers. In order to achieve this, the Combined DNA Index System (CODIS) was established by the FBI Laboratory in1996. The CODIS consists of the 13 most common autosomal STR markers used for identification purposes. All 13 STRs have a high power of discrimination, which makes them suitable for forensic casework. These loci are internationally recognized as the standard for human identification. The 13 CODIS loci are CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11 (15).

In 1999, the DNA working group of the European Network of Forensic Science Institutes (ENFSI) established the European Standard Set (ESS), which consisted of 7 core loci also found in CODIS: TH01, vWA, FGA, D21S11, D3S1358, D8S1179 and D18S51. In order to raise the power of discrimination, Gill *et al.* suggested the addition of five new loci; D10S1248, D22S1045, D2S441, D12S391 and D1S1656. This increased the ESS core STR loci set from 7 to 12 (23-26). The number of ESS markers has recently been increased further with D2S1338, D8S1179, D16S539, D19S433 and D1S165 (27).

By expanding the number of markers used, the discrimination power increases. When the discrimination power increases it becomes easier to differentiate between individuals because the DNA profile become more complex. The robustness of the results can also improve by raising the power of discrimination. In addition to this, different markers have higher sensitivity when analysing degraded/ smaller amount of DNA (25).

As previously mentioned, the marker most widely used for DNA profiles are usually STRs. However, there are other markers suitable for identification purposes and additional analysis, such as single nucleotide polymorphisms (SNPs), markers on the Y-chromosome or mitochondrial DNA.

The Y chromosome is different from the other chromosomes because it is inherited (almost) intact from the father to the son. This means that the chromosome can be traced back in the male lineage in families, giving the genetic markers on the chromosome a valuable role in forensic genetics. Y-STRs are especially useful in separating male DNA in mixtures where there is a great excess of female DNA, as for instance in vaginal samples in rape cases.

Like the Y chromosome, mitochondrial DNA (mtDNA) is passed from generation to generation, but solely in the maternal lineage. This enables its DNA to be traced backwards in families. There are also multiple copies of mtDNA in each cell, making it not only useful for ancestral research but also suitable for cases involving extremely low amounts of DNA or degraded samples. Mitochondrial DNA is also small in size, has a high mutation rate and a lack of recombination making them suitable for DNA analysis.

ALLELE FREQUENCIES AND POPULATION DATABASES

Population databases are collections of allele frequencies from groups of representative samples from a population. Allele frequencies are equivalent to the probability for a specific allele to be found in that specific population. Allele frequencies differ from population to population, which is why it is important to make population databases for each country or geographical area.[1,2] For example, if a suspect matches the obtained DNA profile from a crime scene, different calculations can be performed to assess the rarity of this specific profile in a specific population or population group. The occurrence of a certain DNA profile might be more common in for example the subpopulation of European ancestry compared to the subpopulations of Hispanic or African ancestry in the United States.

To reliably estimate the allele frequencies in a large population it is sufficient to obtain a sample size greater than n=100 (28). Allele frequencies are obtained by dividing the number of times the allele is observed in the population by the total number of allele copies examined in that particular genetic locus. Allele frequencies are a direct reflection of the genetic diversity within a population. The probability of a specific DNA profile occurring in a given population can be calculated using the allele frequencies in a population database. Changes to the allele frequencies over time may indicate genetic drift or new mutations occurring in a population (1, 12, 29).

When an individual's DNA profile is found to match an obtained DNA profile from a crime scene, the individual may be considered the suspect or perpetrator of the crime. How many other people, apart from the perpetrator, could have contributed to this sample? Statistical calculations, such as Random Match Probability (RMP), Random Man Not Excluded (RMNE) or Likelihood Ratio (LR) can be used to answer that (1). These methods have all in common that the they in some way or another consider the probabilities of possible genotypes/allele combinations for each of the loci in a population. These probabilities can be calculated using the allele frequencies in the population database.

1.3 NEXT GENERATION SEQUENCING

New NGS technology is about to replace CE-based methods for DNA analysis. Before we can understand the sequencing methods used today, it is important to understand where and how it all began. DNA sequencing was first described in 1977 by Sanger *et al.* (30) and Maxim Gilbert (31). Sanger sequencing by CE is known as the gold standard for DNA sequencing. DNA sequencing allows to decode the nucleotide sequence of a DNA sample.

The Sanger method utilizes a DNA primer, DNA polymerase, deoxynucleotide triphosphates (dNTP) and dideoxynucleotide triphosphates (ddNTPs) with different fluorescent labels for each nucleotide. The DNA sample is denaturated yielding a template strand, where the primer anneals. The DNA polymerase starts to build the complementary strand with the dNTPs. At different stages one ddNTP is added to the reaction mix and its addition results in the halt of the DNA synthesis at random places on the DNA-strand. The result is a range of different DNA fragments with various0020lengths. The fragments are analysed by capillary electrophoresis, sending the fragments through polymer-filled capillaries. In addition, the fragments pass through a detection cell with a laser measuring the fluorescent strength of all the DNA fragments passing. Thus, in addition to separating the fragments by size they are separated by fluorescence. By combining fluorescence and fragment size, the DNA sequence is assembled. Originally, this reaction was divided into four different tubes, where only a single type of ddNTP would be added, one for each nucleotide. The fragments were separated on an acrylamide gel, where the gel bands had to be read manually. With further improvements over several years the method has been made more efficient and accurate which now allows a computer to read the sequence (32).

New sequencing technologies called Next Generation Sequencing or massive parallel sequencing (MPS) have been developed since then. NGS has revolutionized genomic research with its high speed, scale and throughput. It enables researchers to perform analysis and applications in biology like never before. Using Sanger technology, it took around a decade to sequence the entire human genome in the Human Genome Project, but this can be achieved within a few days using NGS (33). For instance, a single run on the Genome Analyzer (Illumina) in 2005 would produce approximately one Giga base of data, whereas in 2014 the amount was increased a 1000-fold to 1.8 Terra bases. The cost of sequencing has also dropped considerably, from 3 billion dollars (The Human Genome Project) to approximately 1000 dollars for whole genome sequencing today. In addition to lower costs and higher output the

input has also risen with multiplexing. Multiplexing allows to pool and sequence many libraries simultaneously. This is possible by using unique index sequences that are added to each DNA fragment during library preparation. The indexes help identify and sort the reads before final data analysis (1, 34-36).

1.3.1 NGS AND FORENSIC GENETICS

Several NGS platforms and methods have become available during recent years, allowing for large-scale production of genomic sequences. In addition, the number of human genomes sequenced is rapidly increasing. NGS technology enables to sequence several thousand copies of short DNA fragments in multiple individuals simultaneously (37). In forensic genetics the MiSeq FGx Forensic Genomic System from Illumina, the Ion Torrent PGM and the S5 from Thermo Fisher Scientific are the most common NGS platforms used (32). Within genomic research, NGS allows for complex genetic studies that are not technically or economically practical with the Sanger sequencing method alone (35).

Sequencing of the whole genome is known as whole-genome sequencing or shotgun sequencing where no prerequisites are made. However, in forensic genetics there is usually no need to study entire genomes. It is more interesting, that a range of specific parts (targets) of the genome can be analysed and sequenced in parallel. The sequencing of specific genes is known as targeted sequencing, which is the method used in this study (38). In forensic genetics, targeted sequencing is the preferred method over shotgun sequencing. As it allows for a more precise analysis, saving not only time but sample. In comparison to other fields, forensic genetic DNA analysis is challenging because of the low amounts of DNA and/or degraded DNA in the samples obtained, because there is also a need for high accuracy and reproducibility and samples may contain DNA from more than one individual (39). NGS can not only replace Sanger sequencing, but it can also be used for markers that have previously been analysed by other methods, such as STRs with CE, SNPs with Snapshot and RNAs with quantitative PCR.

Some NGS platforms are similar in method and workflow, even though the technical differences and sequencing biochemistry might be different. The platforms method broadly consists of library preparation, sequencing, imaging and data analysis (40). Methods such as hybridization- and amplicon-based enrichments are used for targeted sequencing. Amplicon-based enrichment utilizes a primer mix with tagged oligos for each of the target sequences,

which is mixed with the sample. In a second PCR reaction, the sequences are amplified and the tags are attached to the adapter sequences and indexes. Hybridization-based enrichments utilizes DNA “baits” that represent the target sequences. The sequences hybridize with the “bait” molecule to pull them down for sequencing. Adapter sequences are used for clonal amplification, and indexes are used to identify the sample. Indexes allow for mixing many samples together in one tube, so that they can be sequenced simultaneously. The number of samples that can be analysed together depends on several factors: the number of indexes used, the capacity of the NGS platform, the sequencing depth, and the number and sizes of the targeted regions (37, 38, 41, 42).

After library preparation, the amplicons are clonally amplified in clusters to create measurable amounts for sequencing. The amplification of fragments can be done by bridge PCR, a process in which the PCR amplicons are clustered on a planar substrate (such as a flow cell). Each cluster is a sequencing target and is sequenced in parallel on a chosen NGS platform (42). Sequencing of the clusters is done in real-time using either pyrosequencing, semiconductor sequencing, sequencing by synthesis or sequencing by ligation. Sequencing capacity is the total number of clusters or reads sequenced per run, which varies significantly between different platforms. For example, the MiSeq FGx Forensic Genomic System from Illumina can sequence 5 to 400 million clusters in 2 to 55 hours, and the HiSeq 2500 System from Illumina can sequence up to four billion clusters in five to eleven days. Figures 4, 5 and 6 show schematic drawings of the first three methods used for next generation sequencing. In pyro- and semi-conductor sequencing, the nucleotides are added sequentially. Attached nucleotides generate light signal, which is then detected and interpreted (41).

However, it can be difficult with these methods to distinguish exactly how many nucleotides there are in sequences with more than five homopolymers (nucleotides of the same type) (38). This problem is not relevant in sequencing by synthesis. With this method, all four nucleotides are added at the same time, but only the complementary nucleotide actually attaches to the sequence. All the nucleotides are blocked in the 3’ end, stopping further elongation of the molecule. Sequencing by ligation utilizes DNA ligase and probes labelled with fluorescence fragmented genomic DNA. Florescent signals from different clusters are produced during several rounds with ligation and cleavage (41, 43).

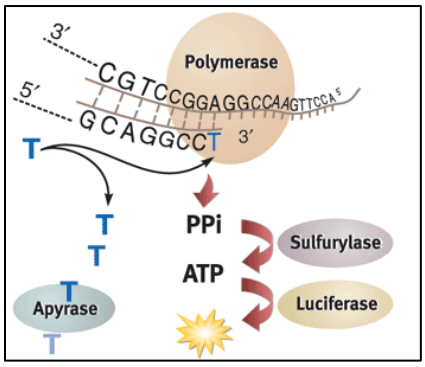


Figure 4. *Overview of pyrosequencing. The addition of one nucleotide at a time releases pyrophosphate that is converted into ATP by sulfurylase and then into light by luciferase. Modified from (44).*

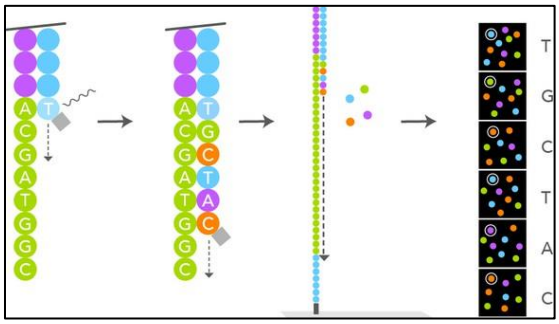


Figure 5. *Sequencing by synthesis. Fluorescently labelled nucleotides are added to the template DNA strand. Addition of a new nucleotides releases a specific fluorescence signal, and the DNA sequence can be typed. Since the nucleotides have terminated ends, they must be removed for another nucleotide to attach by DNA polymerase. After each nucleotide addition, a camera detects the emitted light that corresponds to a base and the DNA is sequenced. Modified from (45).*

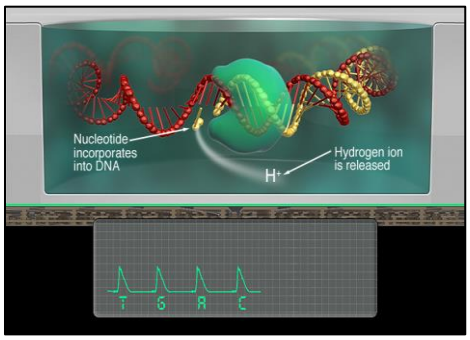


Figure 6. *Semiconductor sequencing technology. The addition of a nucleotide to a DNA strand releases a hydrogen ion. The ion changes the pH value in the solution, which is detected by an ion sensor. The pH change is specific to each nucleotide when the change is noticed the nucleotide will be called and added to the sequence. Modified from (7).*

Illumina has developed a specialized sequencing system for use in forensic genetics, the MiSeq FGx Forensic Genomics System. This instrument has two modes, research and forensic, and the latter is used in this study. Sequencing with the research mode is the same as the standard MiSeq instrument (46). The workflow of this system for forensic use consists of all steps from library preparation of input DNA to the processing of sequenced data. This has been the first fully validated sequencing system intended for forensic genomics applications. The MiSeq FGx Forensic Genomics System Workflow is illustrated in figure 7 (47). The Signature Prep kit contains the necessary reagents to prepare the libraries, and includes primers for targeted PCR amplification of 230 STR- and SNP-regions in the human genome relevant in forensic genetics, (48). Two primer sets are provided, DNA primer mix A (DPMA) and DNA primer mix B (DPMB). Primer mix A contains primer pairs for 58 STRs (27 autosomal STRs, 7 X-STRs and 24 Y haplotype markers) and 94 identity-informative SNPs. Primer mix B is the one used in this study, containing primers for the same markers as DPMA in addition to 56 biogeographical or ancestry-informative and 24 phenotype-informative SNPs (46, 47). In this study, only sequencing results for the 27 autosomal STR-markers have been processed further.

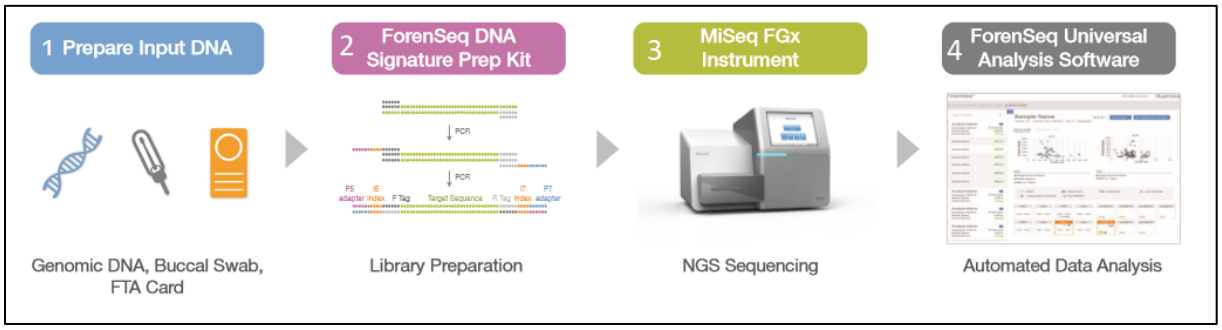


Figure 7. *The MiSeq FGx Forensic Genomics System Workflow consisting of four steps: 1. Preparation of input DNA, 2. Library preparation, 3. Sequencing and 4. Data analysis. Modified from (47).*

Several studies have shown that sequencing by ligation has the lowest error rate among NGS methods, followed by sequencing by synthesis, semi-conductor sequencing and pyrosequencing (49-51). Because the error rates are from genome sequencing studies, they can be misleading. The errors are unevenly distributed and are often related to specific sequence elements, for example sequencing of homopolymer regions. It is therefore too simple to state that sequencing by ligations is the best technology for forensic genetic application. To properly evaluate the quality of a NGS platform/assay, the genotypes must be validated against existing methods.

The standard STR-analysis method today is usually carried out by PCR and CE. The new technology for STR-analysis is based on deep sequencing (NGS). Individual assays for autosomal STRs and SNPs, Y chromosome STRs and SNPs, X-chromosome STRs, indels, mtDNA, ancestry-informative SNPs (AIMs) and phenotype-informative SNPs are all examples of the main forensic markers typed with PCR-CE. Although PCR-CE can be performed in one day and NGS takes 2-3 days, only 30-40 SNP-markers and even fewer STR-markers can be analysed with CE based methods at a time. The advantage of NGS is that several more markers can be analysed in one assay. In addition, different types of markers can be combined in one assay, as for example SNPs and STRs in the ForenSeq DNA Signature Prep kit (Illumina) used in this study. Although not relevant in this study, it allows for the analysis of an individual's STR-profile and the SNPs to elaborate on this particular person's physical appearance and ancestry. This will both save time and the amount of DNA sample used. These factors are extremely important in casework if additional DNA analysis is needed as an investigative lead and/or if the sample volume is low from the start (41).

Another advantage is the ability to analyse degraded DNA because of the short amplicons in most of the markers used. Apart from fragment length, sequence variation within the repeat and flanking region of the STRs can be obtained, increasing the discrimination power. It is also desirable to gain more knowledge about sequence variation in different populations (41, 52, 53).

1.3.2 SEQUENCE VARIATION

As previously mentioned, STRs are widely used as genetic markers for forensic DNA analysis. DNA databases around the world contain millions of valuable STR profiles, which is why STRs will be the preferable genetic markers for DNA profile analysis also in the future (54, 55). However, this does not exclude the ability to extract more information than is normally done from STRs today. Apart from fragment length, sequence variation within these markers can be explored. In Hussing *et al.* 2018 (56) they found no differences in STR typing results between CE and the sequencing method using length-based alleles. But with sequenced-based alleles we can explore the nucleotide composition of each marker. Meaning that the additional information of the markers can give further discrimination power when analysing samples. This can be very helpful if there are mixtures and especially if the two sample donors are closely related.

Traditionally, STR analysis is performed by a size-based DNA separation either using gel electrophoresis or CE. However, PCR product length alone does not identify the eventual variations found within the STRs (57, 58). On the other hand, NGS can identify the variation within the STRs in addition to the lengths traditionally examined. With the ability to examine the sequence variation within STR alleles, the number of effective alleles may increase (59). In addition, this can lead to the separation of samples where two individuals may have the same length-based STR allele (see figure 8) (60),. This was explored by Novroski NMM *et al.* (53) who used the MiSeq FGx Forensic Genomics System (Illumina), STRait Razor and in-house Excel workbooks to characterize the genetic variation within STR repeats and flanking regions of 27 autosomal, 7 X-chromosome and 24 Y-chromosome STR markers in 777 unrelated individuals from four different population groups (61).

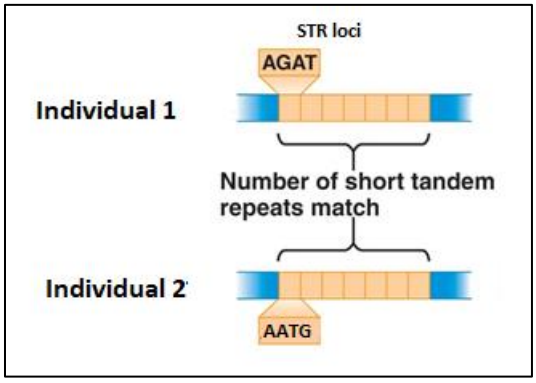


Figure 8. *Two individuals having the same STR repeat number/length-based allele for a specific STR locus, but in fact with two different sequences within the STRs. Modified from: (18).*

The Centre of Forensic Genetics is currently using the NGM SElect kit (Thermo Fisher Scientific) as a standard method for DNA profiling of STR markers. This method is based on PCR of the included autosomal STR markers and fragment length analysis by capillary electrophoresis. However, they are considering to test and implement a sequence-based DNA profiling method as well, using the MiSeq FGx Forensic Genomic System, consisting of the ForenSeq DNA Signature Prep kit, the MiSeq FGx system and the Universal Analysis Software (UAS).

FGC started this work in August 2016. After some initial problems with library preparation, the method is working reliably now (2). In forensic genetics, there is a constant dedication to improve the technology and methods used in order to obtain more and better information from samples containing little DNA. The implementation of a new method demands thorough testing and validation. The manufacturer performs a comprehensive developmental validation, but each laboratory has to perform an internal validation to ensure that the new method is reliable according to the manufacturer's documentation. Validations are usually carried out according to international recommendations and guidelines (62, 63), testing the reproducibility and sensitivity of the method and also concordance with different methods if possible. FGC plans to validate the Signature Prep kit, and some of the data obtained during this master project will contribute here.

To calculate the statistical weight of DNA evidence, a population database must be established for all allele frequencies of the STR-markers included in the specific analysis method used. There is already a Norwegian population database available, but it is currently limited to 10 autosomal STR markers (https://strider.online/frequencies). If the MiSeq FGx Forensic Genomic System is to be implemented in the future, there is a need to establish the allele frequencies for the extra STR markers. FGC has established a research biobank in 2016 containing more than 500 blood samples, with the purpose, among others, to establish the Norwegian population database with allele frequencies for autosomal STR-markers. Approximately half of the samples were already sequenced with the ForenSeq DNA Signature Prep kit in a previous project looking at phenotype-informative SNPs (46). The other half of the biobank samples still needs to be sequenced to obtain a full dataset for autosomal STR markers. Sequencing results for other markers in the kit will be used in future projects at FGC.

1.4 AIMS OF STUDY

- Finalize the ongoing sequencing project on the Norwegian population sample from the research biobank established at FGC. The aim is to obtain as many complete DNA profiles for all 231 forensic genetic markers included in the ForenSeq DNA Signature Prep kit as possible. This is done to enable the biobank to be utilized also in future projects.

- Establish validation data for the MiSeq FGx Forensic Genomic System with focus on the 27 autosomal STR-markers included in the Signature Prep kit:

    - Performance of the MiSeq FGx Forensic Genomic System using data from representative runs with libraries prepared using two different reagent lots of the Signature Prep kit
    - Reproducibility of genotyping results
    - Genotype concordance between the 15-overlapping autosomal STR-markers included in both the Signature Prep and the NGM SElect kit, which is the current standard method for DNA-analyses at FGC
    - Sensitivity: establish the amount of DNA in a sample for which a full DNA profile for autosomal STR-markers can be obtained

- Establish the population database of autosomal STR allele frequencies for Norway based on fragment length, which can be used to calculate the statistical weight of DNA-evidence in criminal cases in Norway.

- Define autosomal STR-alleles based on sequence variation using the MiSeq FGx Forensic Genomic System

2 MATERIALS AND METHODS

The DNA samples used in this study were taken from a research biobank established in 2016 by the Centre of Forensic Genetics, Institute of Medical Biology, Faculty of Health Sciences, UiT The Arctic University of Norway. There is a total of 540 DNA samples in the Biobank. Two previous master students in the research group helped to collect and process the samples (38, 46). Therefore, DNA extraction and quantitation of most of the samples used in this study was already carried out. I extracted the DNA and performed quantitation analysis on the remaining samples. I prepared libraries and sequenced approximately half of the biobank samples. The other half of the samples was already sequenced previously (38, 46), and the sequencing data was used in this study.

The biobank samples used in this study were all analysed using the reagent kits listed in Table 2. Unless otherwise specified in the protocols, all reagents were vortexed and spun down quickly before use. During this study, different lots of the ForenSeq DNA Signature Prep kit and the MiSeq FGx Reagent kit have been used. Validation of the MiSeq FGx Forensic Genomic System was also assessed, focusing only on the 27 autosomal STRs in the Signature Prep kit. By finalizing the sequencing of all the biobank samples, a Norwegian population database of the autosomal STR allele frequencies based on fragment length was consequently built.

Table 2. *Overview of all reagent kits used in this study.*

Kit	Distributor	Purpose
QIAamp DNA investigator kit	Qiagen	DNA extraction from blood
DNA Quantifiler Trio DNA Quantification kit	Applied Biosystems, ThermoFisher Scientific	Quantitation of human DNA in DNA extracts
Qubit ds DNA HS assay kit	Invitrogen, ThermoFisher Scietific	Quantitation of purified libraries for Next Generation Sequencing
ForenSeq DNA Signature Prep kit	Illumina Inc.	DNA Library preparation for Next Generation Sequencing
MiSeq FGx Reagent kit	Illumina Inc.	Sequencing reagents and flow cell for the MiSeq FGx.

SAMPLES

Research biobanks can contain human tissue samples gathered for clinical and research purposes such as microscopy. These biobanks can be based on population studies or patient based studies. In addition, biobanks can contain medical information about the donors. The information and samples can be stores for a long period of time to ensure that long term future research projects can be accomplished. Extra information regarding the donors and samples is registered and documented and may consist of questionnaires, pictures, observations and/or measurements. Tissue samples from the donors may consist of blood, saliva, skin biopsies or even whole organs if the donor has passed away (64, 65).

The DNA samples used in this study were taken from the research biobank established in 2016/2017 at the Centre of Forensic Genetics**.** All data and samples were collected anonymously. Storage and usage of the information in the biobank for this project was approved by the Norwegian Data Protection Authority (DPA) (Reference DPA: 15/00367-3/CGN). Blood sample collection of consenting volunteers was done at various public institutions in Northern Norway, such as the Faculty of Health Science and the Faculty of Law at the University of Tromsø, the police units in Tromsø and Bodø as well as the Norwegian Police University College in Bodø. The samples were stored in EDTA tubes, containing 500-1000 µl of blood each. The sample donors were between 20 and 69 years of age, 73.4% of them were women and 26.6% were men. They all signed an informed consent document. Additionally, digital images and/or colour measurements were taken from each donor's eye, hair and skin. Donors also answered a questionnaire containing questions about gender, height, phenotypic traits and heritage (e.g. the birthplace of all their grandparents) (38, 46). However, the additional information obtained by these questions were not relevant for this study. The purpose of establishing this research biobank was to obtain a representative sample of the population of Northern Norway, i.e. the people that live here, for studies in forensic genetics. This is the reason why the biobank also includes samples from individuals that are not born and raised in this particular part of Norway or even have foreign background. Further information about sample collection and methods is given in (38, 46).

Additionally, two types of Control DNA, 2800M from the ForenSeq DNA Signature Prep kit and 007 from the AmpFLSTR NGM SElect PCR Amplification Kit (Applied Biosystems™, ThermoFisher Scientific), were used to test how little DNA is needed to still obtain a complete DNA profile when using the ForenSeq DNA Signature Prep kit (sensitivity study, see below).

DNA EXTRACTION

DNA from most of the samples in the biobank were already extracted in 2016 using either the DNA DSP Midi kit on the QIAsymphony robot (Qiagen) or the PrepFiler Express Forensic DNA Extraction Kit on the AutoMate Express instrument (ThermoFisher Scientific) (38, 46). Because of extraction failure a few samples (n = 8) needed to be reextracted in this study. For this purpose, the QIAamp DNA investigator kit (Qiagen) with manual pipetting was used, following the manufacturer's protocol (Qiagen, 2012) (66). This specific kit was chosen because DNA yields are known to be higher than with the PrepFiler Express kit. This is not so important for this study, but DNA extracts are also planned to be used in other projects later, requiring higher amounts of DNA. Furthermore, for so few samples it had not been profitable to use the QIAsymphony robot.

DNA extraction is a process in which DNA is separated from other cellular components (1). The QIAamp DNA investigator kit is for isolating both genomic and mitochondrial total DNA from small volumes of whole blood. In addition to the process being efficient, the kit is also designed to reduce contamination between samples. The protocol for the QIAamp DNA investigator kit consist of four steps: 1. Lysing the cells with proteinase K and two different buffers, 2. Binding the DNA to the membrane in the QIAamp MinElute spin column by centrifugation, 3. Cleaning the DNA with ethanol and washing buffers in several steps, and 4. Eluding the DNA with an elution buffer. The result is DNA, which is free of nucleases, proteins and other PCR inhibitors making the sample ready for immediate use.

DNA QUANTIFICATION

As previously mentioned, DNA from most of the biobank samples was already extracted and quantified (38, 46). Therefore, only a subset of biobank samples (n = 8) had to be re-quantified. Furthermore, dilutions of Control DNA used in the sensitivity study (n = 12) were quantified and checked if they contained the expected DNA-concentration. For further details of the sensitivity study see below.

The amount of DNA in each sample was determined by using the DNA Quantifiler Trio DNA Quantification Kit (ThermoFisher Scientific) on the 7500 Fast Real-time PCR system (Applied Biosystems), following the manufacturer's protocol (ThermoFisher Scientific) (67). The purpose of this step is to measure the amount of amplifiable human DNA in a sample so that the amount of DNA used in further analysis steps can be controlled. All the biobank samples were diluted 1:20 before quantification to avoid concentrations outside the range of the standard curve. The standard curve is based on DNA dilutions with known concentrations run in duplicates. Samples between 0.005 and 50 ng/µl can be reliably quantified with this kit.

Three different TaqMan probes, which are short sequences that bind to specific human target loci, are included in the reaction mix. The small and large autosomal targets (SA and LA) are both found in multiple copies on different autosomal chromosomes, and the Y-chromosome target (Y) is located on the Y-chromosome. The probe for the small autosomal DNA target is used to quantify the amount of amplifiable human DNA in a sample. Because large DNA fragments degrade first, the state of degradation in a sample can be measured by comparing the ratio between the small and large autosomal DNA fragments, given as degradation index (DI). The Y chromosome target is used to measure the amount of male DNA in a sample, which may be highly relevant for mixed samples. The information on a sample's DI and amount of male DNA were not further considered in this study.

The quantification system also contains a synthetic internal PCR control (IPC) that detects if PCR inhibitors are present in a DNA sample. The IPC consists of a synthetic DNA template present in each sample and is less amplified in the presence of inhibitors. It can also provide confirmation that all assay components are functioning as expected. Thus, the IPC allows users to distinguish between negative sample results and samples that may be affected by the analysis set up or PCR inhibitors.

In figure 9, a schematic drawing of the quantification method based on a 5'-nuclease assay is shown: 1) The TaqMan probes attach to their complementary sequences (target) between the forward and revers primer sites. They contain a specific fluorescent 5´- reporter and a non-fluorescent 3´-quencher. As long as the probe is intact, the quencher and reporter are close enough for the quencher to absorb the light that the reporter is emitting. 2) Polymerization of the complementary DNA strand and strand displacement of the TaqMan probe starts. 3) During amplification of the target sequences, the Taq DNA polymerase enzyme cleaves the probe and separates the reporter from the Quencher. The result is a non-supressed reporter now able to emit light that can be measured. Taq DNA polymerase has 5' to 3' exonuclease activity, which means the nucleotides on the probe will be cleaved from the 5'end to the 3' end. 4) The polymerization continues. However, because the 3'end of the probe is blocked, there is no extension of the probe during PCR.

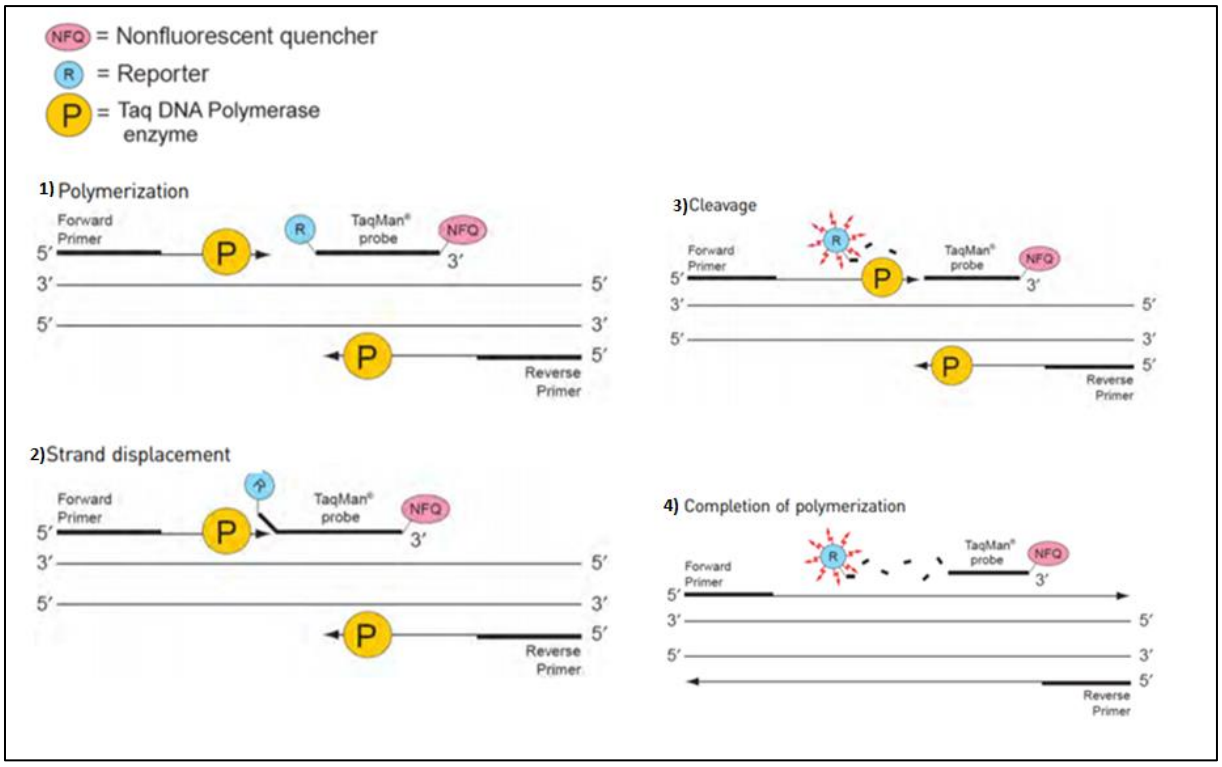


Figure 9. *Overview of the 5' Nuclease assay. 1. The Tag DNA polymerase enzyme anneals to the forward primer and starts polymerization of the complementary DNA strand. 2. Strand displacement of the TaqMan probe begins. 3. The polymerase cleaves the probe. Separation of the reporter and quencher increases the fluorescence signal. 4. Polymerization of the strand continues. Modified from: (67).*

The fluorescent signal increases proportionally to the amount of DNA amplified in each sample. When reporters are cleaved from the TaqMan probe they emit a fluorescent signal. Naturally, the increase in amplified DNA is proportionate to the increase of the fluorescent signal. The strength of the fluorescent signal will at some point exceed a pre-defined threshold, and the sample's fluorescence signal is compared to the standard curve. The number of PCR cycles the samples need to reach the fluorescence threshold is measured in a threshold cycle value (C_T-value). The quicker the sample reaches this C_T-value, the higher the DNA concentration in the sample. The standard curve is based on samples with known concentrations between 0.005 and 50 ng/µl. By using the standard curve, the amount of DNA in each sample can be read by knowing their C_T-value. Figure 10 shows a standard curve plotted against the C_T-value on the x-axis and the DNA quantity on the y-axis.

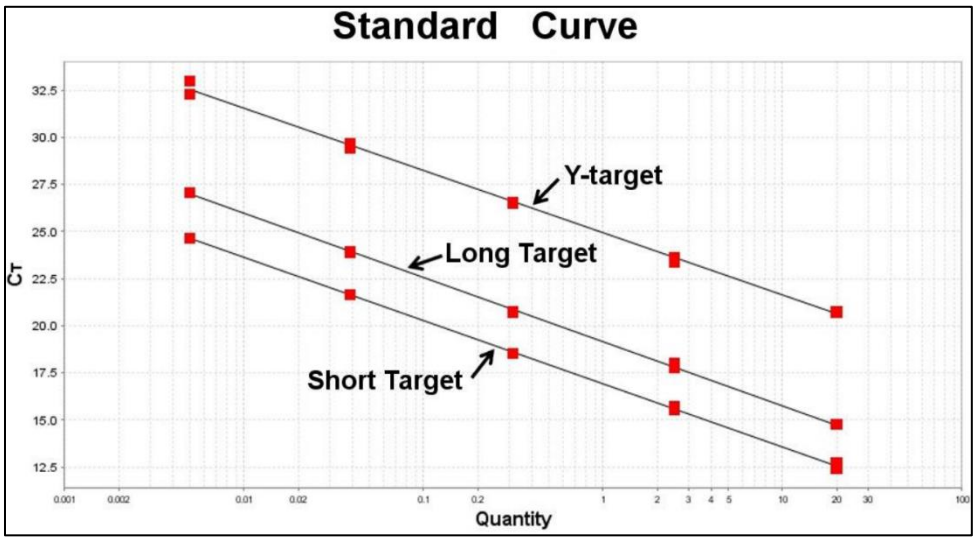


Figure 10. *Standard curve based on the C_T-value of the standard samples quantified in duplicates. Modified from: (68).*

2.4 NORMALIZATION AND DILUTION

In this step, the DNA in all the samples is normalized to 0.2 ng/µl, this is done to ensure the same DNA amount in each sample when the library is being made. Using the quantification results, the biobank samples were normalized by diluting them with TE-buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0) to a concentration of 0.2 ng/µl DNA. 5 µl of each sample is added to the library to have a final DNA amount of 1 ng, which is the recommended input amount for gDNA for the Signature Prep kit.

For the sensitivity study, two types of Control DNA were used to make a series of 6 dilutions with a factor two. The two types of Control DNA were 2800M and 007, from the ForenSeq DNA Signature Prep kit the AmpFLSTR NGM SElect PCR Amplification Kit, respectively. Using the DNA concentration provided by the manufacturers as a starting point, the samples were diluted with TE-buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0). This was done in triplet series to reach the concentrations given in table 3. 2800M and 007 had different starting concentrations. Therefore, the 007 dilution series started at 500 pg and the 2800M dilution series started at 1000 pg. To be able to compare the sequencing and quantification result the triplet DNA samples were marked with numbers from 1-7 to reflect their position in the dilution series. The sample were quantified to ensure that the dilutions was done accurately. All normalized and diluted samples were stored in the fridge (~4°C) until library preparation was conducted.

Table 3. *Amount of Control DNA for the libraries prepared for the sensitivity study. Control DNA 2800M from the ForenSeq DNA Signature Prep kit and 007 from the AmpFLSTR™ NGM Select™ PCR Amplification Kit were diluted with TE-Buffer. By using the manufacturer's stated DNA concertation, the dilution series was calculated so that 5 microliters of each sample into the library would result in the DNA concentration as stated in the table. All dilutions were prepared in triplets.*

Control DNA 007	DNA Amount	Control DNA 2800M
-	1000 pg	**1-2008M**
2-007	500 pg	**2-2008M**
3-007	250 pg	**3-2008M**
4-007	125 pg	**4-2008M**
5-007	62.5 pg	**5-2008M**
6-007	31.25 pg	**6-2008M**
7-007	15.625 pg	-

2.5 LIBRARY PREPARATION, FORENSEQ DNA SIGNATURE PREP KIT:

All the libraries were prepared by using the ForenSeq DNA Signature Prep kit, following the manufacturer's protocol with minor exceptions (Illumina, 2015) (69). Samples were all diluted with TE-buffer and not nuclease free water as stated in the protocol.

When working with the beads, several steps were taken to ensure that they were well mixed when in use. For example, they were never centrifuged after vortexing. The beads were also vortexed regularly in-between pipetting steps. The largest available pipette tips was always used to ensure an even distribution of beads, although volumes were so small to allow for smaller tips. These three steps were taken in order to inhibit the beads sinking to the bottom of the tube and prevent uneven distribution of beads to each sample. This was done during library purification and normalization. Each setup usually contained 30 samples in addition to two controls, according to the manufacturer's recommendation for samples of good quality in combination with the DPMB.

For the sensitivity study, the dilution series of different control DNAs (2800M and 007) were prepped and sequenced in separate setups that were filled up with biobank samples to reach the maximum of 32 reactions.

Library preparation consists of various steps shown in Figure 11: Amplifying and tagging the target sequences, enriching the targets, purifying libraries, normalizing and pooling the libraries, and finally diluting and denaturing the pooled libraries. The DNA sample is mixed with tagged oligos that are linked to copies of the targets by PCR. This forms DNA templates with the regions of interest flanked by universal primer sequences. Index adapters are then attached to the tags and amplified with another PCR. Thereafter, the library is purified, normalized and pooled and is ready to be sequenced.

Figure 11. *Overview of the Library Prep workflow using the ForenSeq DNA Signature Prep kit. Modified from (Illumina, 2015) (69).*

2.5.1 AMPLIFY AND TAG TARGETS, PCR1

In this step, the genomic DNA in the samples are amplified and tagged using the ForenSeq oligonucleotide primer mix (Illumina). The primer mix is specific to different DNA sequences up- and downstream of STR and SNP markers included in the kit. Two different DNA primer mixes, DPMA and DPMB are included in the kit. Only DPMB is used in this study, including primers for 58 STRs (including 27 autosomal STRs, 7 X and 24 Y STRs), 94 identity-informative SNPs, 56 ancestry-informative SNPs and 22 phenotype-informative SNPs. Thus, results for a range of different markers can be obtained in one analysis.

All the DNA samples were vortexed and spun down before use. The master mix was made by pipetting PCR1 reaction mix, FEM enzyme mix, and DPMB together, according to protocol. 10 µl of the master mix was added to each of the 32 wells in a MicroAmp Optical 96-Well Reaction Plate (Applied Biosystems), then 5 µl of the normalized human DNA samples was added to their respective wells. A positive control (2800M) and negative control was also used in each setup. The plate was then sealed and spun down before placing it on the Veriti™ Thermal Cycler (Applied Biosystems, ThermoFisher Scientific) for the first PCR (PCR1). The cycling conditions were set according to the manufacturer's protocol for a 9700 thermal cycler in 9600 emulation mode (70). PCR1 lasts for about 3.5 hours. The PCR1 program entails: 98°C for 3 min, 8 cycles of: 96°C for 45s, 80°C for 30s, 54°C for 2 min with specified ramping mode (8%) , 68°C for 2 min also with specified ramping mode, 10 cycles of: 96°C for 30s, 68°C for 3 min with specified ramping mode, then 68°C for 10 minutes before the final hold at 10°C (69).

2.5.2 TARGET ENRICHMENT, PCR2

In this step, the DNA is amplified, and sequences required for cluster amplification are added. Index adapters 1 (i7) and 2 (i5) are added giving each sample a unique combination of index adapters. This is necessary for the sequencing system to be able to separate the data from different samples in the library after being pooled together. The ForenSeq DNA Signature Prep kit includes eight different index adapters 2 and twelve different index adapters 1. Because each setup consists of 32 samples including controls, only four of the twelve different index adapters 1 are used at a time. To prevent any contamination between runs, the use of the same four index adapters 1 was avoided in adjacent runs.

After PCR1 was finished, 4 µl of index 1 and 2 were added to each well according to the sample sheet. 27 µl of PCR2 reaction mix was then added to the wells, and the plate was sealed again and placed on the Veriti™ Thermal Cycler for the second PCR (PCR2). The cycling condition were set according to the manufacturer's protocol for a 9700 thermal cycler in 9600 emulation mode (70). The PCR2 program entails: 98°C for 30 seconds, 15 cycles of: 98°C for 20s, 66°C for 30s, 68°C for 90s, then 68°C for 10 minutes and a final hold at 10°C (69). When working with the index adapters it is important to prevent cross contamination between them. Therefore, gloves were changed frequently, and the index tubes were sealed with fresh caps each time after used.

2.5.3 PURIFY LIBRARIES

After PCR2, the libraries were purified using Sample Purification Beads (SPB). The goal is to separate the amplified DNA from the PCR reaction components, such as unbound adapters. When using SPB it is important to mix the solution thoroughly by vortexing and pipetting up and down . This is to insure an even distribution of the beads added. A magnetic stand was used to trap the amplified DNA between the magnetic beads. These form a pellet as long as the reaction plate is placed on the magnet. The pellet was then washed twice with freshly prepared 80% ethanol before the amplified DNA was resuspended from the beads using a resuspension buffer (RSB). The purified libraries were now ready to be normalized.

2.5.4 QUBIT® DSDNA HS ASSAY

Before normalization of the libraries, the amount of DNA in each sample was quantified with the Qubit® dsDNA HS (High Sensitivity) Assay Kit following the manufacturer's protocol for the Qubit® 2.0 Fluorometer (Invitrogen, ThermoFisher Scientific, revision B.0) (71). The samples were quantified in order to determine if every step has worked as expected to this point. In addition, the Qubit measurements give an indication of how well the sample may perform under sequencing.

The kit contains a dilution buffer, two DNA standards and an assay reagent. A master mix was prepared with the dilution buffer and assay reagent. The master mix was distributed to the assay

tubes that contained a final volume of 200 µl after adding 2 µl of DNA sample or 10 µl of standard.

First, the two standard tubes were read to create a standard curve that the DNA samples could be measured against. Then, the sample tubes were placed into the Qubit® Fluorometer and read by measuring the emitted fluorescence light from fluorophores attached to the dsDNA in the sample. The instrument measures the concentration of the sample after it is diluted in the assay tube. Therefore, the actual concentration needs to be calculated using the following equation:

Concentration of the sample (ng/µl) = QF value * (200/volume of sample added)

The QF value is the value given by the Qubit® 2.0 Fluorometer. The volume of sample added to the assay tube in this study was always 2 µl.

2.5.5 NORMALIZE LIBRARIES

The normalization process prepares for cluster generation by normalizing the DNA concentrations among samples so that they are equally represented in the sequencing run. This step is necessary to achieve a pooled library with the same amount of DNA from each sample. By doing this, the samples can achieve consistent cluster density to optimize the resolution of the individual samples when pooled together.

First, a master mix is made of Library Normalization Additives 1 (LNA1) and Library Normalization Beads 1 (LNB1). Here it is important to remember that the LNB1 has to be vortexed thoroughly in order to have an even distribution of the beads in the master mix. 20 µl of the purified library were added to the master mix and shaken on a BioShake XP (Quantifoil Instruments GmbH) for 30 minutes. The library was then washes twice with Library Normalization Wash buffer 1 (LNW1) before 0.1 N HP3 was added and mixed well. HP3 was added to break the hydrogen bonds between DNA bases, converting a DNA strand from double-stranded to single-stranded molecules. However, it is important to work fast after adding the HP3, because otherwise the DNA will start to degrade. Lastly, the Library Normalization Storage Buffer 2 (LNS2) was added. The library was now normalized and ready for the next step.

2.5.6 POOL, DILUTE AND DENATURE LIBRARIES

In this step, the normalized libraries were pooled together in a single tube, adding 5 µl of each sample. The pooled samples were sequenced together on the same flow cell. Finally, the pooled library was diluted with the Hybridization Buffer (HT1) provided in the MiSeqFGx Reagent kit, and 2 µl of Human Sequencing Control (HSC) was added, making a total of 600 µl. The HSC is a DNA library of 23 ForenSeq STRs functioning as a positive sequencing control (48).

According to the manufacturer, it is recommended to use 7 µl of pooled library in the final dilution. Based on previous experience with this kit, 14 µl of pooled library was used for most of the sequencing runs (46). However, for a few runs, only 10 or 12 µl of pooled library was used to avoid over clustering of the flow cell. The decision was based on higher than usual Qubit values of the purified libraries that may be caused by variation between lots.

When denaturing the diluted pooled library, it was heated for exactly 2 minutes, then immediately placed on ice and kept there until the MiSeq FGx instrument was ready to be loaded.

2.6 MISEQ SEQUENCING, MISEQ FGX REAGENT KIT:

Developed in collaboration with the forensic community, Illumina created the MiSeq FGx™ Forensic Genomics System, which is the first fully validated NGS system specifically designed for forensic genomics applications. The MiSeq FGx system enables users to analyse forensic DNA samples from DNA-to-data. The system contains the UAS, which can analyse STR and SNPs for human identification. A major advantage of NGS in forensic genomics is the ability to analyse and separate alleles that are identical in size, but different by sequence. This provides an important and precise method for human identification.

The software enables run setup, sample management, analysis, report generation as well as mixed DNA sample detection. The software also provides several features not relevant for this study, such as population statistics, automated sample comparison as well as an estimation of a sample donor's biogeographical ancestry, hair and eye colour (47, 72, 73).

Following the instructions in the MiSeq FGx Instrument Reference Guide (73), the flow cell was thoroughly washed, and the reagent cartridge was loaded with the denatured diluted pooled library. Each sequencing run takes approximately 30 hours, containing 398 cycles. A total of 16 sequencing runs were conducted with the 17 prepared library set ups, including some reruns due to human and machine errors.

The Illumina sequencing workflow consist of four steps, sample preparation (explained above), cluster generation, sequencing and data analysis (74).

2.6.1 CLUSTER GENERATION

Cluster generation occurs on the flow cell, which is a glass slide with lanes. Each lane is a channel with a lawn coated with two types of oligos. When the library samples run through the flow cell, hybridization occurs randomly between the DNA fragment adapters in the sample fragments and the complementary oligos attached to the flow cells. Then, a polymerase creates a complimentary copy of the original template strand. This is the first step in cluster generation. The double stranded molecule is then denatured, and the template is washed away as it is not covalently bound to the adapter sequence on the flow cell. Clusters are generated by a process called bridge amplification. The strands are clonally amplified by folding over, and the adapter region hybridizes to the second type of oligo on the flow cell. Polymerases elongate the second oligo by generating a complementary strand, thus forming a double stranded bridge. The two copies are denatured resulting in two single stranded copies of the molecule that are covalently bound to the flow cell on each their oligo. This bridge amplification process happens simultaneously for millions of clusters and is repeated. After bridge amplification, all the reverse strands are cleaved and washed away, leaving only the forward strands that have their 3’ end blocked to prevent any unwanted priming. Figure 12 shows an overview of the basic steps of cluster generation (74).

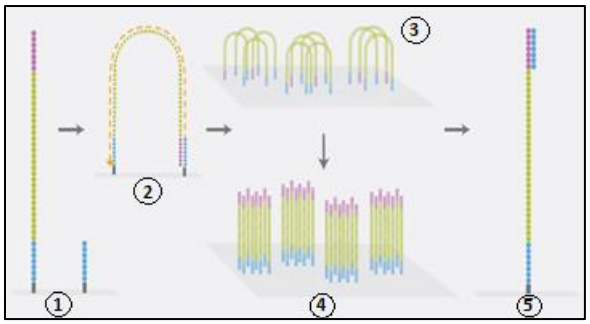


Figure 12. *Overview of cluster generation steps. 1. The original DNA template is used to make a complementary template copy. 2. Bridge amplification where the template copy flips over and hybridizes to the neighbouring complementary oligo, creating a bridge. 3. The template is copied creating a double stranded bridge. 4. The bridge is denatured. 5. The sequence primers hybridize, making the samples ready for sequencing. Modified from: (75).*

2.6.2 SEQUENCING

The sequencing technology applied by Illumina involves a process called Sequencing by Synthesis (SBS), using fluorescently labelled dNTPs with a reversibly terminated end that are sequentially added by a DNA polymerase. The individual dNTPs carry a specific fluorophore that helps the software to identify each base added during sequencing. Sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle, all the four nucleotides compete to be added to the growing chain, but only one of them will match based on the complementary nucleotide in the template strand. After the nucleotide is added the others are washed away. The clusters are excited by a light source, and their specific florescence is emitted. The number of cycles determines the length of the read whilst the emitted wavelength and signal intensity determine the base call. The fluorophore and terminated ends on the dNTPs are cleaved and washed away, enabling the addition of a new nucleotide in the following sequencing cycle.

After the first read, the read product is washed away. In this step, the index 1 read primer is introduced and hybridized to the template, and the read is generated similar to the first read. After completion of the index read the product is washed away. The 3’ ends of the template are unblocked, and the template fragment folds over and binds to the second type oligo attached to the flow cell, creating a bridge. Index 2 is read in the same manner as index 1. The polymerase extends the second oligo forming another double stranded bride. The double stranded bridge is cleaved, and the 3’ ends are blocked. The original forward strand is then cleaved and washed

away, leaving only the reverse strand. The second read begins with the introduction of the read 2 sequencing primer. The sequencing steps are repeated as in read 1 until the desired read length is achieved. This process generates millions of reads that represent all the sequenced fragments.

Finally, the sequences from the pooled sample libraries are separated based on the unique indexes introduced during the sample preparation. For each sample, reads with similar stretches of base calls are locally clustered. Forward and revers reads are paired to make a contiguous sequence, which are then aligned back to the reference genome by the software (47, 74, 76, 77).

2.6.3 DATA ANALYSIS

Before running the MiSeq FGx instrument, a file containing the names and different index combinations for each sample for the particular run was uploaded onto the server and transferred to the instrument. After the run had finished, the data was automatically analysed with the ForenSeq Universal Analysis Software (78). The software is specifically developed for analyses of data obtained with the ForenSeq DNA Signature Prep Kit. It is installed on a stand-alone server, and the data is viewed through a web browser (Google Chrome).

After a sequencing run, the results were evaluated. When first looking at the overall quality of the two control samples and quality metrics, they would either be marked with green, orange or black. Green indicates that all metrics are within the acceptable range set by the manufacturer. Orange indicates that at least one metric is not within the acceptable range. Black indicated that no metrics are assessed with the analysis. The quality metrics include cluster density, clusters passing filter, phasing, pre-phasing, number of reads and index quality.

Cluster density (K/mm^2) is the number of clusters per square millimetre on the flow cell for the run. The recommended target cluster density range is between 400– 1650 K/mm^2. Clusters Passing Filter (%) is the percentage of all clusters that pass the Illumina chastity filter, which measures quality. Data only appears after cycle 25, and the filter detects low quality base calls. The recommended cluster passing filter is a above 80%. However, cluster density and cluster passing filter values that are outside of the recommended range can still produce results that are sufficient for analysis.

During the sequencing reaction, each DNA strand in a cluster is extended by 1 base per cycle. A strand can come out of phase with the current cycle, either falling a base behind (phasing) or

running a base ahead (pre-phasing). Phasing and pre-phasing estimates the fraction of molecules that became phased or pre-phased in each cycle within Read 1 and Read 2. Low percentages indicate good run statistics. The recommended phasing and pre-phasing values are $\leq 0.25\%$ and $\leq 0.15\%$, respectively.

A check list was made, containing different criteria that had to be met in order for each sample to be considered as a complete DNA profile (see Appendix I). For example, if an allele was 10-15% of the alternate allele's size, and in stutter position, it was considered a stutter**.** A stutter is a PCR artefact, which is produced when the DNA polymerase slips and creates a minor product band which is smaller than the main fragment (usually 4bp shorter) (79). After all the checkpoints were run though, a sample was re-sequenced if it did not have a complete STR or SNP profile, i.e. there were any drop outs or if it was uncertain if a marker was homozygous or heterozygous, i.e. there was one allele call that was below the interpretation threshold. The goal was to achieve complete profiles for all the markers even though only results for the autosomal STR markers were considered further in this study. It should also be mentioned that all thresholds pre-set by Illumina were used, such as the analytical threshold, interpretation threshold and stutter filter.

Further assessment of the sequencing result included the comparison of two reagent lots, checking the reproducibility of the kit as well as the concordance between the Signature Prep and NGM Select kit. As previously mentioned, several lots of the Signature Prep kit were used during this study. Based on previous experiences (38, 46) there could be lot to lot differences. Therefore, the run metrics and marker performance of two representative runs from each of two different reagent lots was compared. Marker performance included, STR coverage, allele balance within each STR, markers flagged for imbalance by the UAS and stutters which had to be manually removed because the UAS called them as alleles. The average coverage of each STR was plotted in Excel and standard deviations were calculated using the STDAV formula within the Excel program. Allele balance was calculated by dividing the allele with the lowest amount of reads with the highest. If the software calculates an intralocus balance below its predefined threshold, the imbalance quality control indicator is triggered. The allele balance was also plotted in Excel with their corresponding standard deviation values. The UAS generates a file containing sequencing information of all the samples run, including if any marker has been flagged for imbalance. By going through this file, all the STR markers flagged for imbalance could be identified. Lastly, all sample were correlated after sequencing. and the markers which had stutters marked as alleles were corrected and noted.

Because many samples were re-sequenced during this study, there were several parallel sequencing results available for a range of individuals. This data was used to compare the STR results and check the reproducibility of the MiSeq FGx Forensic Genomic System. The samples run more than once were checked for each marker, and any discrepancies where noted and evaluated. To be included in the comparison, the overall sequencing run had to be good, and samples had to have more than 85000 reads. Lastly, the concordance of STR-genotypes in the 15 overlapping markers between the two kits were evaluated. The NGM SElect genotype data was already available at the Centre of Forensic Genetics.

The dataset was analysed using STR Analysis for Forensics (STRAF), an online tool that performs forensics and population genetics analysis of STR data, to obtain the length-based allele frequencies for the Norwegian population sample. In addition, population genetic parameters were also obtained, such as Genetic Diversity (GD) and Polymorphism Information Content (PIC, the probability of being able to assume which allele a parent has passed on to the child), Match Probability (PM, the probability of a match between two unrelated individuals), The Power of Discrimination (PD, the probability to discriminate between two unrelated individuals), The power of exclusion (PE) and the Typical Paternity Index (TPI), which reflects the mean allele frequency for random non-excluded men at a given locus (80).

A Principal Component Analysis (PCA) module is also included in the STRAF-tool. PCA is used to control the population substructure and quality. The PCA detects genotype clusters caused by population substructure (81). A single cluster on the PCA is normally expected for forensic population studies based on STR data. The single cluster expectation can also serve as a quality control of the data. If outliers are identified on the PCA plot, a check of these samples is warranted (80).

The dataset can also be tested for Hardy-Weinberg Equilibrium (HWE) and Linkage Equilibrium (LE) with the STRAF-tool. HWE is used to check if alleles within a locus are independent, while LE is used to check if loci are independent. These two tests are usually used when establishing a new population database.

The HWE can calculate genetic variation of a population at equilibrium. The HWE equation is based on the Hardy-Weinberg equilibrium, which states that the amount of genetic variation found in a population will remain constant between generations if no disturbing factors such as no net mutation, natural selection or migration occur. Further assumptions of the model are random mating and a large population size. If these factors are fulfilled, the genotype

frequencies in a population will remain unchanged over following generations, and the population is in Hardy-Weinberg equilibrium. The model can be applied to the genotype frequency of a single gene.

The Hardy-Weinberg model for a marker with two alleles p and q entails two equations: one that calculates allele frequencies ($p + q = 1$), and the other that calculates expected genotype frequencies ($p^2 + 2pq + q^2 = 1$). So, if p and q allele frequencies are known or measured in a population, then the three expected genotype frequencies p^2, 2pq and q^2 can be calculated using the Hardy-Weinberg equation (1, 30). A marker is in HWE if the observed and expected genotype frequencies are not significantly different. For STR markers it is more complicated to calculate HWE because there are more than two possible alleles for each marker.

Allele frequencies can be calculated based on the observed genotype frequencies in the DNA profile data set of the Norwegian population. Then, the allele frequencies are used to calculate the expected genotype frequencies. The observed and expected genotype frequencies are compared in the HWE test. If they differ from one another, then there is a deviation from HWE, which means that some genotypes are more common than expected. The reason why could be natural selection, mutation or migration within the population tested.

Linkage equilibrium (LD) is a term used for two genes/ loci that are inherited independently in each generation, for example, two STR loci that are in random association and are originally located on two different chromosomes. In contrast, if two genes are in linkage disequilibrium, certain alleles from each gene are inherited together more often that they would be by chance. The genes can be closely located on the same chromosome. Alternatively, linkage disequilibrium could be caused by an interaction between combinations of alleles at the two loci affecting viability of potential offspring (82).

To ensure that an allele is reliable in statistical calculations of the evidential weight, a minimum allele frequency is used for rare alleles. The minimum allele frequency is 5/2N, where N is the number of individuals sampled from the population. Allele frequencies can be inaccurate if the allele is so rare that it is represented only a few times in the database. Some alleles might be so rare that they are not represented in the database at all. Each allele is recommended to at least be observed 5 times before it can be included in the database. However, for rarer alleles, the frequency is adjusted to the minimum frequency (1).

After determining the length based allele frequencies using STRAF, the sequence based alleles were characterized using Excel, based on the nomenclature described in Devesse *et al*.

2018 (52, 53, 83, 84). The allele frequencies along with their corresponding sequence is listed in Appendix IV. Some of the markers had a flanking region next to the sequence characterized, these are written in a light colour, and include D13S317, D18S51, D19S433, D1S1656, D5S818, D7S820 and vWA.

3 RESULTS

During this study, a total of 287 individuals were sequenced. Some of them had to be re-sequenced. In total, 16 library setups were prepared, each containing 30 samples plus controls. One of the goals of this study was to sequence all of the 540 individuals in the research biobank to build a population database for allele frequencies of autosomal STR-markers. After correlating the samples, complete data for all the autosomal, X- and Y-STRs was obtained for most of the samples. However, for 20 samples there are still some SNPs that yielded inconclusive or no results. These SNPs are known to be difficult to type, considering their low number of reads. These samples will be re-sequenced at a later stage.

Since a significant quantity of samples was sequenced more than once, there was a good opportunity to compare their respective results to each other and evaluate the reproducibility of the MiSeq FGx Forensic Genomic System. Of the 287 individuals sequenced during this study, 263 (91.6%) were sequenced more than once. There were no deviations in autosomal STR genotypes between replicates, so the reproducibility was 100 %.

3.1 PERFORMANCE OF THE FORENSEQ DNA SIGNATURE PREP KIT

Lot performance of the ForenSeq DNA Signature Prep kit was done with data from two representative, technically good-quality runs from each lot. Two runs from each lot were

chosen, and the sequencing of these runs was performed under the same conditions and according to the protocol. 1 ng DNA was amplified with DPMB from the ForenSeq DNA Signature Prep kit. All samples were quantified with the Qubit dsDNA HS Assay (Qiagen). 14 µl of pooled library was used for all four sequencing runs.

First, some of the run metrics of the two lots were compared (table X). The optimal cluster density range is expected to be between 400-1650 K/mm^2. From the four runs, a cluster density between approximately 1100 and 1500 K/mm^2 was obtained, which is well within the set parameters set by the manufacturer (73) and also the expected range based on earlier run experience.

Qubit measurement of the samples was done to determine the DNA concentration in each sample after library purification to indicate how well the sample had performed so far. The average DNA quantity after purification is relatively similar between lots, ranging from 5.7 to 7.15 ng/µl with overlapping standard deviations (table x). The sample coverage ranged between 100 000 and 672 000 reads, except for one sample from run X which had less than 85 000 reads. There was no obvious difference in the average sample coverage between lots (table 4). The index CV indicates how even the number of reads are between the samples. The index CV ranged from between 32.5 and 43.7 %.

Table 4. *Performance of two different lots of the ForenSeq DNA Signature Prep kit.*

	Lot 1 (RGT8232152)		**Lot 2 (RGT 1518597)**	
DNA-quantity of purified libraries (ng/µl) - Qubit	2.1 – 12.1		0.48 – 13.6	
Average Qubit quantity (ng/µl)	7 ± 3.07	7.15 ± 3.26	6.49 ± 0.96	5.47 ± 2.33
Run number	1	2	3	4
Cluster density K/mm^2	1163	1201	1532	1185
Index CV (%)	36.5	43.7	32.5	38.1
Average # Reads	287633 ± 89387	350517 ± 123332	480200 ± 127644	383571 ± 107138
Maximum # reads/sample	491 000	592 000	672 000	603 000

STR marker performance was evaluated, including aspects such as locus coverage, allele balance within markers, markers flagged for imbalances and stutters not removed by the stutter filter in the UAS.

The average number of reads varies greatly between markers. Most of the STR markers have an average allele balance close to 0.8, with variable standard deviations (Fig. 13). STR-markers D1S1656, D5S818 and D22S1045 have slightly lower allele balances. Their average allele balance for lot 1 and 2 was respectively 0.78, 0.73, 0.61 and 0.55, 0.56, 0.61. STR-markers D1S1656 and D5S818, and also D19S433 have higher standard deviations than the other markers. This means the alleles are more variable, which could indicate that they are more variable than the other markers. Sometimes they may be good, but other times they yield bad results. In general, there is not a big difference between the two lots, but for D1S1656 and D5S818 the allele balance was even poorer in lot 2 than in lot 1. Table 5 shows an overview of the markers flagged for allele imbalance by the UAS, using predefined thresholds by Illumina. The markers flagged most often for imbalance were D1S1656, D5S818 and D22S1045 (n>30). There is a clear correlation between markers with most imbalance flags and markers with low allele balance (Fig.13). There were more imbalances observed in lot 1than in lot 2.

The marker with the lowest average amount of reads for lot 1 is vWA and D19 for lot 2. The marker with the highest average amount of reads from lot 1 is D20S482 and TH01 for lot 2. In addition to this, the standard deviation for markers such as TH01 and D6S1043 are quite large compared to the other ones.

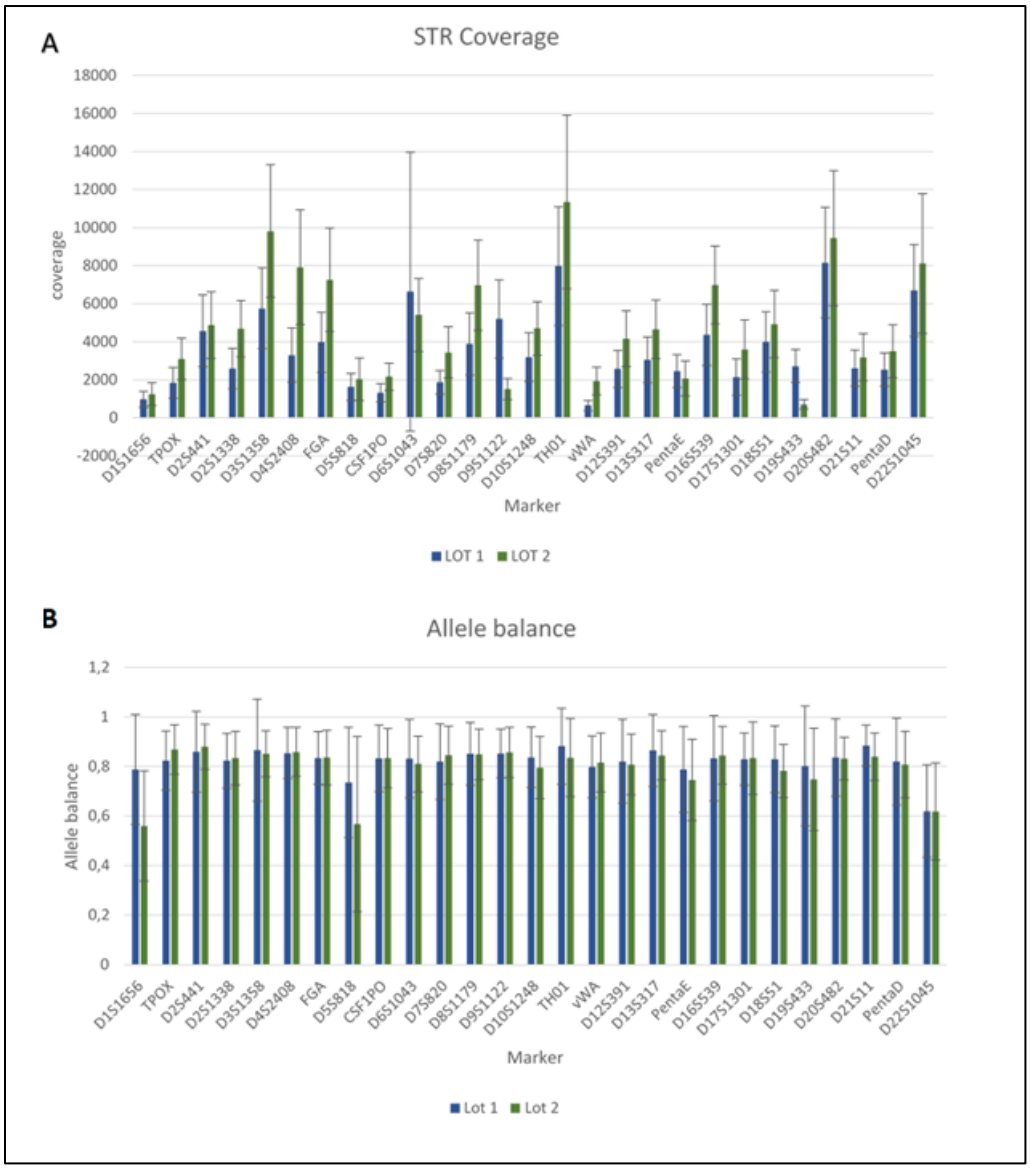


Figure 13. *A: The average #reads ± standard deviation for each STR marker. B: The average allele balance of STR markers ± standard deviation.*

Table 5. *All STR markers flagged for imbalance by the UAS filter.*

Marker	LOT 1: flagged for imbalance	LOT 2: flagged for imbalance	Total flagged per marker
D1S1656	4	29	**33**
D5S818	13	18	**31**
D22S1045	17	13	**30**
PentaE	5	10	**15**
D18S51	6	3	**9**

vWA	4	5	**9**
PentaD	3	3	**6**
D12S391	3	2	**5**
D19S433	4	1	**5**
D7S820	1	4	**5**
CSF1PO	2	2	**4**
D6S1043	2	1	**3**
TPOX	2	1	**3**
D10S1043	1	1	**2**
D16S539	2	0	**2**
D17S1301		2	**2**
D2S1338	1	1	**2**
D8S1179	1	1	**2**
FGA	0	2	**2**
D10S1248	1	0	**1**
D21S11	0	1	**1**
D4S2408	0	1	**1**
D9S1122	0	1	**1**
TH01	0	1	**1**
Total	**72**	**103**	

Table 6 shows the number of stutters that were not filtered away by the UAS and had to be manually removed from the DNA profile. For lot 1, the markers with most unfiltered stutters were D6S1043, D17S1301, D22S1045, D9S1122 and D18S51. For lot 2, the markers with most unfiltered stutters were D19S433 and D20S482.

Table 6. *Number of stutters that were manually removed because they were not filtered away by the UAS.*

Marker	Lot 1	Lot 2	Total per marker
D6S1043	4	2	**6**
D17S1301	3	2	**5**
D22S1045	3	1	**4**
D9S1122	3	1	**4**
D18S51	3		**3**
D19S433	2	4	**6**
D20S482	1	3	**4**
D21S11	1		**1**
D7S820	1		**1**
vWA	1		**1**
D10S1248		1	**1**

D2S1338		1	**1**
Penta E		1	**1**
Total per lot	**22**	**16**	

CONCORDANCE BETWEEN THE SIGNATURE PREP AND THE NGM SELECT KIT (THERMOFISHER SCIENTIFIC)

3.2

Concordance between the Signature Prep kit and the NGM SElect kit from ThermoFisher Scientific (CE based method) a method used by FGC, was checked to establish if two kits with different methods could produce the same autosomal genotype results.

The STR markers analysed in these kits vary, so only the 15 overlapping markers could be considered. The 15 markers used were: D10S1248, D12S391, D16S539, D18S51, D19S433, D1S1656, D21S11, D22S1045, D2S1338, D2S441, D3S1358, D8S1179, FGA, TH01 and vWA. The genotype results from NGM SElect were previously obtained from the same biobank samples by FGC (unpublished data). The comparison of genotypes obtained with the Signature Prep kit and genotypes obtained with the NGM SElect Express kit showed three discordances. After comparing 32400 alleles using the NGM SElect and the Signature Prep kits (2 kits × 15 loci × 2 alleles/locus × 540 samples), genotype concordance was found in 99.99% (32397 out of 32400) of the STR alleles compared. Discordance was found in in three different individuals and in a total of two different markers (see table 7).

Sample 1 showed a discordance in marker D22S1045. When using the Signature Prep kit, allele 11 (>7000 reads) was typed instead of allele 19 which was typed with the NGM Select kit. Allele 19 was however, present with (with 95 reads) but was not called by UAS and went unnoticed as a possible artefact instead of an allele. Sample 2 showed a discordance in marker D1S1656 where allele 13 (with 360 reads) was typed with Signature and allele 12 was typed with the NGM SElect. In this case as well, allele 12 was present (50 reads, stutter position) but was not called by the UAS. Lastly, sample 3 showed a discordance in D22S1045 where allele 8 (with 11000 reads) was typed with Signature and allele 12 was typed with NGM SElect. In this case Allele 12 was not at all present with the Signature Prep kit.

Table 7. *Genotype concordance between the Signature Prep and NGM SElect kit. The three individuals listed showed discrepancies. Sample 1 showed discordance in D22S1045, allele 11 was typed with Signature Prep in contrast to NGM Select typing allele 19. Sample 2 showed discordance in D1S1656, the Signature Prep kit typed allele 13 compared to allele 12 typed with NGM Select kit. And lastly, sample 3 showed discordance in D22S1045, allele 8 was typed with Signature Prep and allele 12 was typed with the NGM Select kit.*

Sample	Marker	Kit	Genotype
1	D22S1045	NGM	11,19
		Signature	11,11
2	D1S1656	NGM	12,13
		Signature	13,13
3	D22S1045	NGM	8,12
		Signature	8,8

3.3 POPULATION DATABASE

The length-based alleles obtained from the biobank where analysed with the computer programme STR Analysis for Forensics (STRAF). This was done to obtain the autosomal STR allele frequencies for the Norwegian population, to perform various population genetic analysis.

With STRAF analysis the autosomal STR allele frequencies were calculated based on the length-based allele frequencies (see figure 14) was obtained. The average number of alleles per marker is 11.3. The markers with the highest number of length-based allele variants are FGA (19 alleles), D18S51 (18 alleles) and D1S1656 (18 alleles). The markers with the lowest number of length-based allele variants are D4S2408 (5 alleles), TPOX (7 alleles) and TH01 (7 alleles) (Fig. 14). The data file containing the allele frequencies used for this figure can be found in Appendix III.

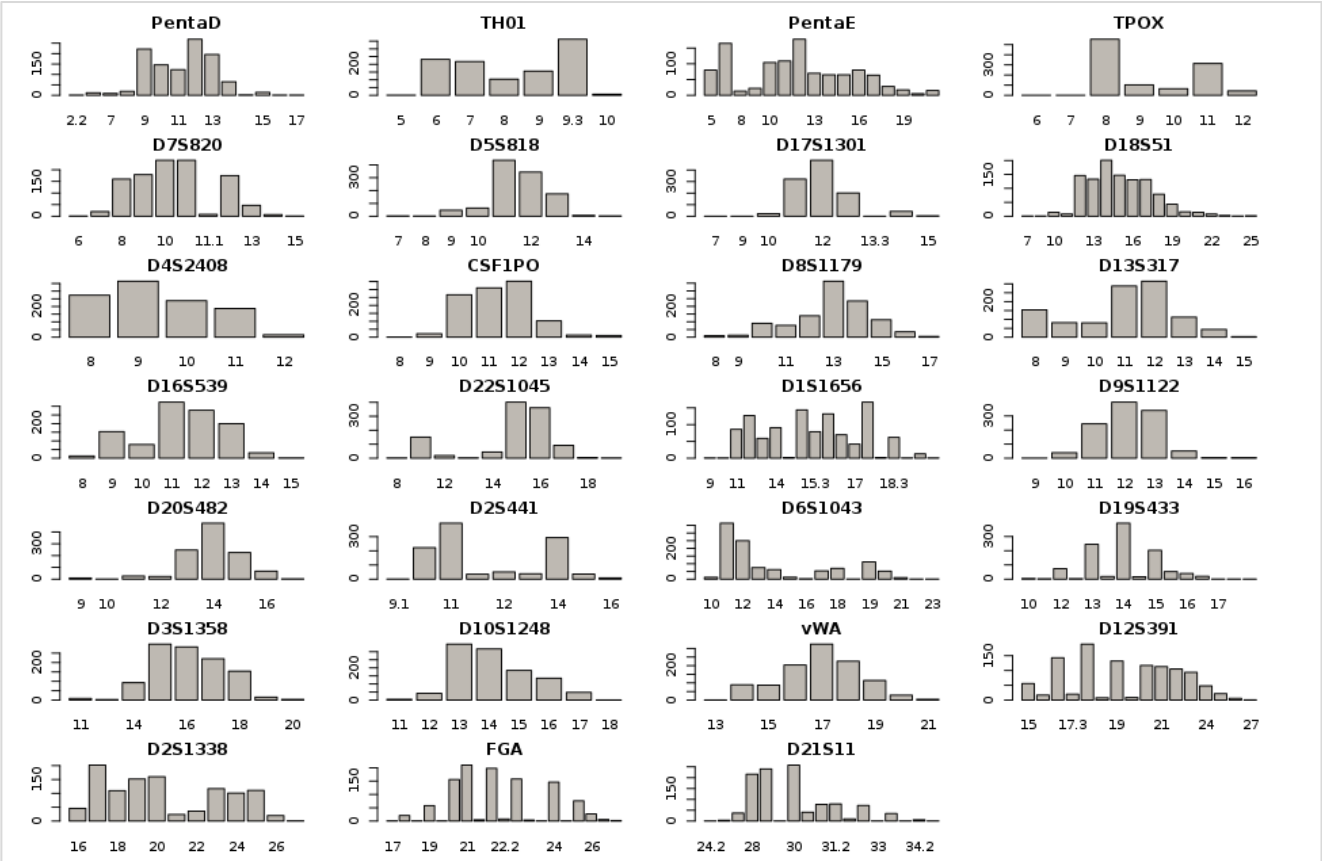


Figure 14. *Plot distribution of LB-allele frequencies per each STR-locus calculated using STRAF.*

Different forensic parameters for the Norwegian population were calculated using STRAF. The results (see table 8) for each marker includes values for; Expected Heterozygosity/ Genetic diversity (GD), Polymorphism Information Content (PIC), match probability (PM), power of discrimination (PD), Observed Heterozygosity (Hobs), power of exclusion (PE), the typical paternity index (TPI) and p-value of the Hardy-Weinberg equilibrium (pHW).

pHW is calculated using the expected and observed heterozygosity values. PentaD, vWA and D21S11 are the only three markers which have a pHW value <0.05, which means they are not in HW equilibrium.

Table 8. *An overview of the studied population's different forensic statistics.*

locus	N	Nall	GD	PIC	PM	PD	Hobs	PE	TPI	pHW
C5F1PO	1080	8	0.7412	0.6950	0.1174	0.8826	0.7630	0.5322	2.1094	0.1350
D1051248	1080	8	0.7629	0.7249	0.1014	0.8986	0.7907	0.5820	2.3894	0.1510
D125391	1080	16	0.8922	0.8815	0.0231	0.9769	0.8963	0.7879	4.8214	0.8040
D135317	1080	8	0.8002	0.7722	0.0693	0.9307	0.7981	0.5956	2.4771	0.4830
D165539	1080	8	0.7840	0.7503	0.0802	0.9198	0.7778	0.5585	2.2500	0.8560
D1751301	1080	9	0.6731	0.6151	0.1671	0.8329	0.6963	0.4226	1.6463	0.7670
D18551	1080	18	0.8769	0.8634	0.0296	0.9704	0.8889	0.7728	4.5000	0.6440
D195433	1080	15	0.7717	0.7400	0.0832	0.9168	0.7630	0.5322	2.1094	0.4350
D1516564	1080	18	0.8994	0.8898	0.0206	0.9794	0.8907	0.7765	4.5763	0.2860
D205482	1080	9	0.7060	0.6606	0.1330	0.8670	0.7000	0.4283	1.6667	0.6350
D21511	1080	17	0.8356	0.8150	0.0467	0.9533	0.8204	0.6374	2.7835	0.0270
D2251045	1080	10	0.7223	0.6760	0.1252	0.8748	0.7167	0.4545	1.7647	0.1220
D251338	1080	12	0.8789	0.8658	0.0280	0.9720	0.8556	0.7058	3.4615	0.2640
D25441	1080	9	0.7454	0.7046	0.1055	0.8945	0.7315	0.4786	1.8621	0.4600
D351358	1080	9	0.7864	0.7524	0.0820	0.9180	0.7815	0.5651	2.2881	0.5640
D452408	1080	5	0.7433	0.6969	0.1147	0.8853	0.7556	0.5193	2.0455	0.8300
D55818	1080	9	0.7048	0.6542	0.1418	0.8582	0.7056	0.4369	1.6981	0.2600
D651043	1080	15	0.8053	0.7823	0.0612	0.9388	0.7944	0.5888	2.4324	0.4190
D75820	1080	11	0.8232	0.7977	0.0573	0.9427	0.8278	0.6516	2.9032	0.7860
D851179	1080	10	0.7996	0.7738	0.0659	0.9341	0.8093	0.6163	2.6214	0.3960
D951122	1080	8	0.7117	0.6583	0.1446	0.8554	0.7352	0.4848	1.8881	0.2550
FGA	1080	19	0.8599	0.8430	0.0377	0.9623	0.8704	0.7354	3.8571	0.1240
PentaD	1080	14	0.8280	0.8046	0.0550	0.9450	0.8278	0.6516	2.9032	0.0200
PentaE	1080	16	0.9032	0.8945	0.0184	0.9816	0.8870	0.7690	4.4262	0.6330
TH01	1080	7	0.7713	0.7349	0.0877	0.9123	0.7648	0.5355	2.1260	0.9420
TPOX	1080	7	0.6396	0.5846	0.1943	0.8057	0.6574	0.3655	1.4595	0.2500
vWA	1080	9	0.8056	0.7783	0.0678	0.9322	0.8259	0.6480	2.8723	0.0360

A heat map of pairwise LD values shows that most of the loci are independent of each other (see figure 15). Values under 0.05 are significant, which means they are in linkage disequilibrium (53). The lowest values are shown in a darker colour in the heat map, for example FGA, TPOX and FGA, PentaD both have a p-value of 0.0. A matrix with pairwise p-values is available in Appendix II. The PCA projection (see figure 16) obtained by STRAF yielded a plot based on the populations length-based alleles. It shows that the individuals in the population sample clustered relatively close together with a slight distribution. There are two

outliers (marked 1 and 2 in figure 16). These samples were collected from individuals of African heritage.

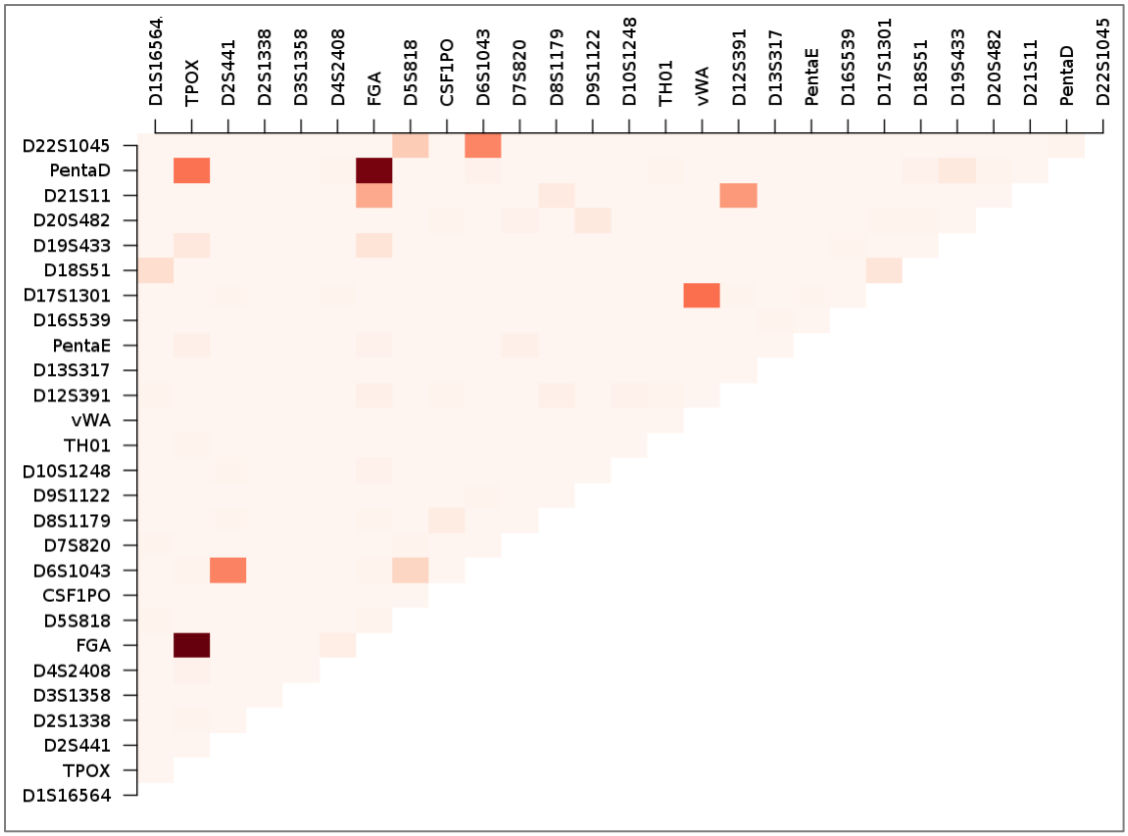


Figure 15. *Pairwise Linkage Disequilibrium p-values matrix plotted in a heat map. The darker the colour the lower the p-value for the corresponding loci. Values close to 0 indicate a high degree of linkage between the two loci.*

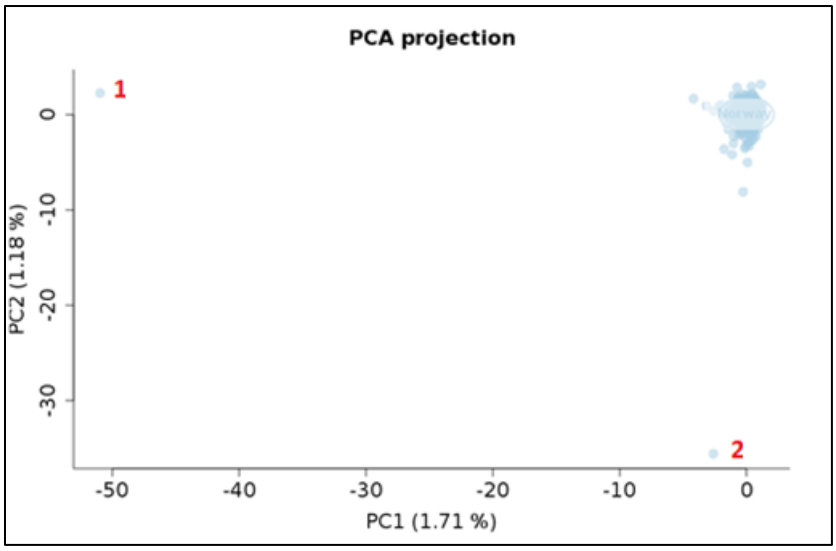


Figure 16*. PCA projection of Norwegian populations length based alleles obtained by STRAF. All the sequenced (length-based allele) samples are plotted in a graph, showing two outliers, marked 1 and 2. Both these individuals, had upon further examination, African heritage.*

3.4 SEQUENCE VARIATION

Characterization of the sequence-based alleles lead to multiple novel alleles being observed. Markers such as CSF1PO, D18S51, D22S1045, D16S539, D17S1301 and especially D21S11 showed an increase in sequence variants. D211S11 was the marker with the highest sequence variation with

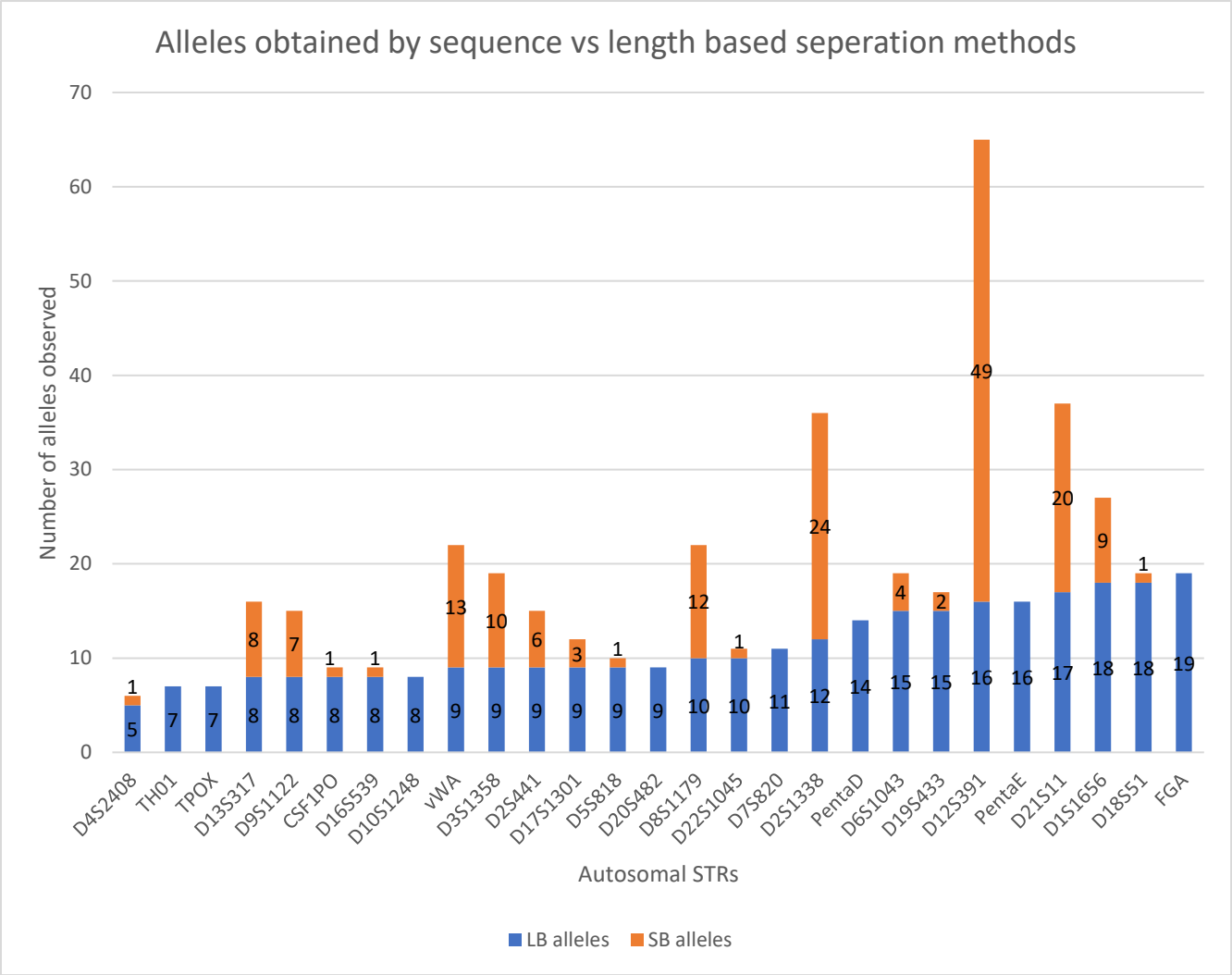
The sequence based allele frequencies showed many rare alleles (1 out of 1080 allele copies) in the Norwegian population. Some of the rare alleles obtained are found in markers: CSF1PO, D13S317 and All of the sequence-based allele frequencies are listed in Appendix IV, next to its corresponding sequence. The novel alleles found in this study are marked red in Appendix., Unfortunately, because of time restraint, not all markers where checked to see if there were any novel alleles. These markers include, FGA, D12, D19 and vWa.

Some of the characterized alleles their sequence variants are described below.

1.

2.

The number of individual alleles obtained by length and sequence were compared, see figure 17. This was done to determine if some markers may become more informative by sequencing compared to length based. Eight of the twenty-seven autosomal STRs (TH01, TPOX, D10S1248, D20S482, D7S820, Penta D, Penta E and FGA) did not show an increase of alleles between the two methods (see figure 17). However, there are several markers that double, and even triple, their number of alleles obtained by a sequence based separation method compared to length based methods. These markers include; D13S317, vWA, D3S1358, D8S1179, D2S1338 and D21S11. The most polymorphic marker in this study is D12S391, the number of alleles for this marker increased with 49 to a total of 65 alleles.



Figur 17. *Allele increase for 27 autosomal STR markers targeted by the ForenSeq™ kit when comparing sequence-based to length-based alleles. The number of alleles obtained by length based separation methods are shown as the blue pole. Whilst the orange pole above represents the additional alleles obtained by sequence based separation methods.*

3.1

SENSITIVITY STUDY

For the sensitivity study two control DNA, 2800M and 007, were chosen to make the dilutions. The samples were also quantified, to ensure that they contained the correct concentration before library preparation and sequencing. Table 9, shows the quantified and measured DNA concentration compared to the intended concentration (Dilution). All the DNA concentrations

measured are approximately the same as the intended dilution. With small DNA concentrations, manually pipetting can affect the outcome considerably. This makes it very easy for the dilutions to vary and become a little more uneven than intended.

Table 9. *Control DNA samples, 2800M and 007, with intended dilutions beside the quantified DNA concentration.*

Control DNA 007			Control DNA 2800M	
Sample Name	Quantified Concentration (ng/µl)	Dilution (ng/µl)	Quantified Concentration (ng/µl)	Sample Name
		0.2	0.238	1-2008M
2-007	0.152	0.1	0.125	2-2008M
3-007	0.077	0.05	0.0542	3-2008M
4-007	0.0301	0.025	0.0321	4-2008M
5-007	0.0199	0.0125	0.0191	5-2008M
6-007	0.0101	0.00625	0.009	6-2008M
7-007	0.0036	0.003125		

The sample coverage is visualized in figure 18. All the samples were sequenced in triplets to ensure more reliable results. There is a clear parallel between the DNA concentrations and the sequencing coverage. The decrease in coverage is gradual as the concentrations gets lower, but it is not halved as one would expect when the concentration is halved. At each dilution step when the DNA concentration is lowered with 50%, the sample coverage only drops with approximately 25%.

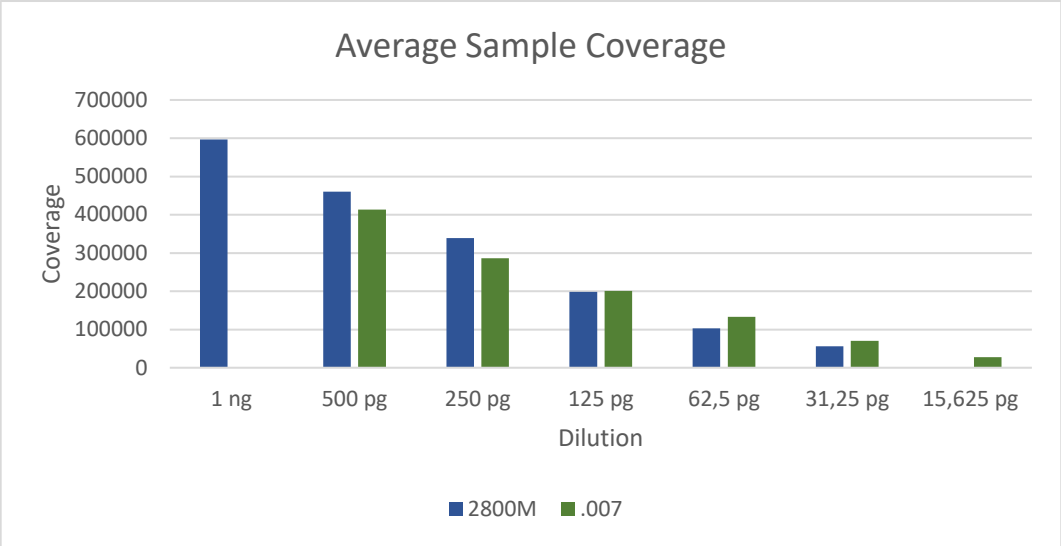


Figure 18. *Average sample #read (coverage) for control DNA samples analysed with the MiSeq FGx Forensic Genomic System. Six steps of triplet dilutions of control DNA 2800M and 007 were used, ranging from 1 ng to 15.6 pg. However, the two control DNAs had different starting concentrations in the dilution series.*

Apart from the system being able to read the samples with low concentrations, it was also important to examine whether or not Illuminas sequencing system could produce full DNA profiles at lower concentrations. This is especially important in forensic genetics, when sample obtained from crime scenes often can be low in amount and or very degraded. Figure 19 shows the allelic coverage of each sample. Between 1 ng and 250 pg DNA input, both control DNAs yielded a 100% DNA profile with no drop outs. And for the 2800M control DNA at 125 pg, a complete DNA profile was obtained. However, for two of the three 007 control DNA samples at 125 pg, one of the alleles had dropped out, giving an allele count of 98.8%. Samples with DNA amounts of 62.5 and 31.25 pg started to lose more alleles with both DNA controls. However, over 80% of the alleles were present in these samples. At 15.6 pg, only approximately half of the alleles where present in the DNA profile.

D19S433 was the marker which most often had allelic dropouts (n=13), 4 and 9 times for respectively 2800M and 007 control DNA. And despite full profiles being achieved at 125 pg DNA, approximately 12 markers were flagged for artefacts such as stutters or imbalances. As the DNA concentration decreases, also the number of flags for imbalances and number of stutters that had to be removed manually increases (data not shown).

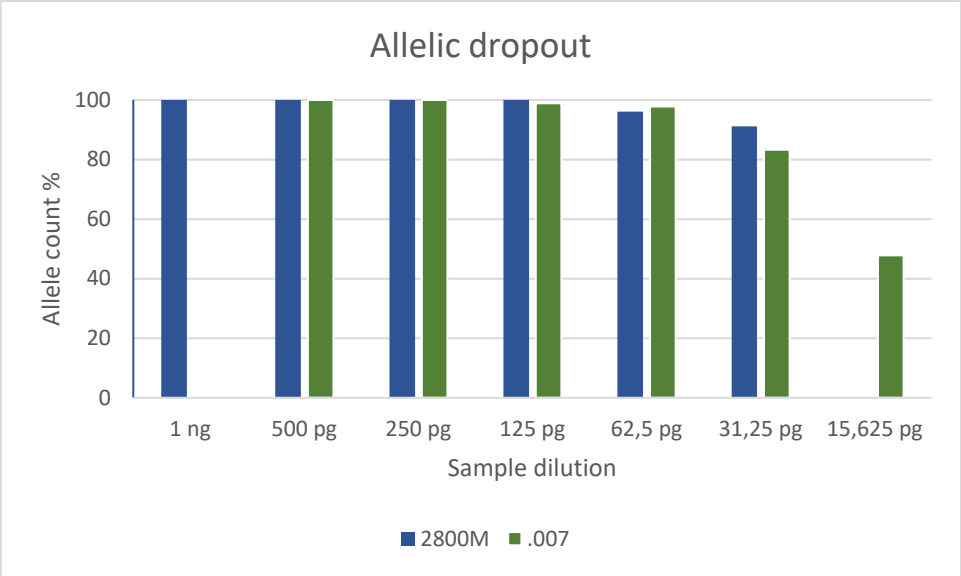


Figure 19. *Allelic counts for a dilution series of two DNA controls, 2800M and 007, analysed in triplets. Complete profiles with no alleles missing have an allele count of 100%.*

4 DISCUSSION

4.1 TECHNICAL RESULTS

When obtaining complete DNA-profiles for all the samples in the Norwegian population sample from the biobank, many samples had to be re-sequenced because certain STR or SNP markers had incomplete or missing results. This also means that samples that already had complete STR-profiles but were incomplete for even a single SNP were re-sequenced despite that this information was not necessary for this study focussing on the autosomal STR markers only. In the end, sequencing results from a previous study (38, 46) and this study were combined for data analyses. If samples had low total coverage, this could often be traced back to errors done in the lab, such as evaporation of the samples during PCR or possibly manual errors such as pipetting. When dealing with so many samples and pipetting such small amounts errors can easily occur. In addition to this, one of the reagent lots of the ForenSeq DNA Signature Prep kit used gave incomplete results for certain SNP markers. It took some runs to understand the problem. In the end, the manufacturer replaced this low quality reagent lot, and the affected samples were re-sequenced with new reagents.

One of the aims of this study was to finalize the sequencing project on the Norwegian population and to obtain complete DNA profiled for all 231 forensic genetic markers. This goal was almost completed. There are 20 samples left that need to be re-sequenced, because they have missing or inconclusive results in some SNP markers. Sequencing of the research biobank samples was stopped, because all the information on autosomal STRs needed to conclude this study was obtained, and time was running out to complete this thesis.

The standard procedure was further modified by adding the Qubit measuring step. This was done to measure the DNA quantity of each sample after the library had been purified and before normalization. The purpose of adding this step was to estimate how the samples/lots were going to behave in sequencing. For example, if a sample has a high DNA quantity at this point and the sequencing result is poor the error has probably occurred during normalization or pooling the libraries, as long as the sequencing run itself is approved. By storing the measured DNA quantities, the performance of different lots can be compared. When receiving new kit lots, the Qubit measurements of the purified libraries can be used to evaluate if PCR performance is

similar compared to the previous reagent lot. In this study, Qubit measurements have been very helpful and should therefore be considered as a permanent part of the standard procedure.

4.1.1 PERFORMANCE OF THE FORENSEQ DNA SIGNATURE PREP KIT

Based on previous experience by FGC (38, 46), the amount of pooled normalized library used for sequencing was raised from the recommended 7 µl to 14 µl. This was done to ensure sequencing results within the parameters set by the manufacturer, with highest possible cluster density yet avoiding overload, and also to ensure the maximum possible number of reads per sample and marker. Increased amount of pooled normalized library for sequencing has also been used in other studies, such as Devesse *et al.* 2018 (52) who used 12µl , as well as Novroski *et al.* 2016 (53) who used 10µl.

The cluster density values obtained by all runs were within the parameters and are therefore considered as representative values for the other sequencing runs performed in this study. In studies conducted by Churchill *et al.* 2016 and Xavier *et al.* 2017, cluster densities of 716 K/mm^2 and 607 K/mm^2 were obtained when using the recommended amount of pooled library (7µl) (85, 86). These cluster density values are within but at the low end of the optimal range set by the manufacturer. Although the cluster density obtained by these two studies are approved by the manufacturer, there can be drawn a correlation between amount of input library and cluster density. The cluster density in this study was between 1100-1500 K/mm^2 using 14 µl pooled library compared to 600-700 K/mm^2 by Churchill *et al.* 2016 and Xavier *et al.* 2017, using 7 µl. When comparing these results, it seems as if by doubling the input amount the cluster density also doubles. Individual differences between instruments may also contribute to the difference in cluster density values when using the same amount of pooled normalized library. Further research could be performed to obtain the optimal input amount, since this is not already specified by the manufacturer.

A total of 4 sequencing runs were chosen to evaluate the performance of two reagent lots of the ForenSeq DNA Signature Prep kit. In general, the obtained values for Qubit measurements, cluster density, and the number of reads/coverage per sample from the 4 runs were quite similar. Index CV value is only affected if there are samples within a sequencing run which fail in

anyway. This means that the Index CV, 43.7 % for one of the runs in lot 1 had a failed sample. All the other Index CV values are between 32.5 and 38.1 %.

STR marker performance was evaluated, involving locus coverage, allele balance within markers, markers flagged with imbalance and stutters that were not filtered by the software. Locus coverage of the STR were quite variable. However, the STR marker with the lowest coverage for lot 1 and 2 was VWA and D19S433, respectively. The STR markers with the highest coverage for lot 1 and 2 was D20S482 and TH01, respectively. The three markers with the lowest allele balance were D1S1656, D5S818 and D22S1045. In addition, they had large standard deviations, which indicated that they were more variable between individuals than the other autosomal STRs.

D19S433 has quite a high standard deviation for its allele balance, although the allele balance itself is similar to the rest of the markers. This could possibly be explained by the low coverage in lot 2, which could have led to a more variable result.

The markers flagged frequently for imbalance correlate well with the markers with low allele balance. Lot 1 had a total of 72 flagged markers, whilst lot 2 had a total of 103. This indicates that lot 2 may have worse allele balances than lot 1, which means there could be variations between different lots and that it could be worth checking when receiving new lots for analysis. There are several people contributing with DNA to the stain, DNA-profile is more difficult to interpret if markers have poor allele balances.

Some alleles designated as alleles by the UAS slipped through the stutter filter and were called as alleles by the software. These had to be manually removed from the DNA-profile. Overall, there was no obvious difference in the number of unfiltered stutters observed in each lot, but there may be differences in individual markers such as D6S1043, D19S433 and D17S1301. Interestingly, the markers with most unfiltered stutters in lot 1 seem to be less effected in lot 2 and vice versa. However, the sample size may be too small to make a definitive conclusion.

Even though the mistake alleles are relatively few, a person is still needed to go thoroughly through all the results after sequencing. It is however, hard to conclude anything because the sample size of this particular study is relatively small. Increased levels of stuttering and imbalances detected between lots may indicate that the PCR in the affected lot is working less optimally. Unfiltered stutters also make it harder to interpret mixed DNA samples with minor contributors. If there already is little DNA present, unfiltered stutters are harder to separate from

the true alleles. An idea could be to perform a more thorough stutter study, similar to Hussing *et al.* 2018 (87). A new study could lead to the adjustment of the stutter filter values in the UAS, making the results more reliable.

REPRODUCIBILITY OF THE MISEQ FGX FORENSIC GENOMIC SYSTEM

4.2 Since so many samples were sequenced more than once there was sufficient data available to test the reproducibility of the MiSeq FGx Forensic Genomic System. Sample run results with low coverage were disregarded, as well as samples that had inconclusive results for any of the autosomal STR-markers, meaning having one allele call below the interpretation threshold set in the Universal Analysis software. Additionally, all profiles were manually inspected and evaluated for potential stutters and other artefacts. The final approved STR-profiles showed that results obtained with the ForenSeq DNA Genomic System are 100% reproducible. Jäger *et al.* 2017 (93) also evaluated reproducibility in their study. Using DPMB they analysed 231 loci 5 times using five different researchers. Although, they did not include STR locus D22S1045 in their study, because the primer set was still in development during this time, they yielded a 100% reproducible result for the other markers.

Reproducibility is an important aspect of the validation process that is carried out whenever new reagents and instruments are tested before implementing them in forensic casework (62, 63). Reproducibility of methods is so crucial because the obtained results are more trustworthy, bring a level of transparency to the method and raise the credibility and understanding of what was done with a sample when for example presenting the results in court.

4.3 CONCORDANCE BETWEEN THE SIGNATURE PREP KIT AND THE NGM SELECT KIT (THERMOFISHER SCIENTIFIC)

Only the overlapping markers between two methods can be used to perform a concordance evaluation. At FGC, the PCR-CE based NGM SElect kit is the current method used for DNA-

profiling. The goal was to see if concordant results for length-based allele calls could be produced with the Signature Prep kit. A concordance of 99.99% (32397 out of 32400 alleles) was achieved. There were only two STR markers that showed discordance between the two kits, D22S1045 and D1S1656.

In a study conducted by Hussing *et al.* 2018, 30 samples were typed with the AmpFLSTR NGM SElect Express PCR Amplification Kit and the ForenSeq kit. Out of their 450 STR genotypes tested, 448 (99.6%) were found to be concordant. The two discordances were drop-ins in D21S11 in the PCR-CE assay (88). In another study, by Garcia *et al.* 2011, concordance was done between three kits. The Identifiler-NGM, PowerPlex ESX17 and the Investigator ESS systems. They yielded a 99.99% concordance. The only discordances were due to null alleles in the vWA marker (89). Gettings *et al.* 2016 (90) also evaluated the concordance between length-based CE and NGS genotypes for 183 samples (4392 loci in total). They also observed ~99% concordance between the two methods.

It seems common that rare discordances occur when using different reagent kits. Discordances may be caused by using different primers in the kits compared. If a primer binding site for one of the markers mutates in one of the kits, this may lead to less amplification product or no product at all, i.e. a null allele, in the other kit, as was found in the study by Garcia *et al.* 2011.

4.4 SENSITIVITY STUDY

In forensic genetics and in DNA analysis the ability to yield usable DNA profiles/results from degraded or damaged samples is critical. Evidence samples from biological stains are often in poor condition for several reasons, for example caused by environmental conditions. Damaged and degraded DNA does not perform well in PCR amplification because it hinders polymerization (91, 92). This in turn can lead to partial or no DNA profiles at all. In addition, PCR artefacts can potentially yield incorrect genotypes in the DNA profile. A solution to this problem has been developed and utilized in for example the MiSeq FGx Forensic Genomics System from Illumina. They say having employed steps to raise the analysis sensitivity. PCR cycle number has been increased, amplicon sizes have been reduced, and the amplicons are purified after PCR (93-97). As of now, NGS is a technology known to exhibit the highest sensitivity.

Therefore, the purpose of this sensitivity study was to explore how little DNA could be used in the ForenSeq DNA Signature Prep kit and still yield complete DNA profiles. The results show, that complete DNA-profiles (no allelic dropouts) are still detected with 125 pg DNA input. The only exception is for two of the 007 control DNA triplets, where one allele fell out, making their profile 98.8% complete. 62.5 pg and 31.25 pg samples had over 80% full DNA-profiles, for both control DNAs. Even with 15,625 pg DNA (007 control DNA), approximately 50% of the alleles were called. Jäger *et al.* 2017 (98) also performed a sensitivity study using DPMB and quadruple samples ranging from 1 ng down to 7.82 pg DNA. Their results did however show even higher sensitivity than performed in this study. DNA inputs ranging from 1ng down to 62.5 pg yielded complete profiles. Whilst 15.625 pg and 7.82 pg DNA yielded a 70% and 50% allele call, respectively.

Despite complete DNA-profiles being achieved with 125 pg DNA in this study, approximately 12 markers were flagged for artefacts such as stutters or imbalances. As DNA concentrations decrease, the number of stutters that are called as alleles and imbalances increases (data not shown). However, after manual assessment, the genotypes are found to be correct. This could however, have an impact on mixed DNA samples. The increase in allele calls will make it harder to accurately interpret the results.

Fattorini *et al.* 2017 (99) performed a sensitivity study with the commercial kit PowerPlex® ESI 17 system and the 3500 Genetic Analyser Sequencer (Thermo Fisher Scientific) for CE analysis. When using the DNA amount of 55 pg DNA, no results were produced.

At a first glance, the results in this study show that the sensitivity of the ForenSeq DNA Signature Prep kit may be equivalent to CE-based kits, i.e. producing complete DNA-profiles with 125 pg (100). However, the Signature Prep kit has an advantage, especially in forensic genetics cases where the samples used may be of low quantity and quality. As little as 15 pg DNA can still yield 50% allele calls. With CE, the DNA-profiles of so little DNA would only contain a few or none alleles. Thus, the ForenSeq DNA Signature Prep is more sensitive than CE-methods. Some trace evidence sample that effectively would not yield any results using CE could in fact produce useful result with the Signature Prep kit.

4.5 POPULATION DATABASE

There were only three markers, which did not fulfil the HWE: D21S11, vWA and Penta D. They all had p values less than 0.05. Most of the loci were not in LD, but there were a couple of loci that were not entirely unlinked, such as FGA, Penta D and TPOX.

The importance of having unlinked loci is so that when allele and genotype frequencies are being estimated correctly. When for dealing with example linked loci or the markers differ from HWE it may indicate genotyping errors, the populations structure might not be as expected, homologous regions in the genome, then is to prevent errors in genotyping.

The PCA is used to determine the population's substructure and quality control. The PCA in the figure shows a slightly scattered cluster with two obvious outliers only. This is indicating that there is little population substructure in the Norwegian population sample despite having included…. The outlier samples were checked back in the database and their heritage confirmed as African.

The STRAF analysis, including the PCA plot, were not separated into more populations, even though the population used was including individuals with more than strictly Norwegian heritage. The number of individuals from other origin is not sufficient to separate the dataset into several subpopulations. Due to the compound population that lives here in Northern Norway, specifically the university, students and employees from all over the world come to study and work in Tromsø. Therefore, the individuals in the research databank are from different places in the world, explaining the slight scattering in the PCA plot.

The first length-based allele frequencies were calculated. These can be used for further calculations/research and to calculate the statistical weight of DNA evidence to present to the court. The most polymorphic marker based on length-based alleles is FGA with 19 alleles, whilst the least polymorphic marker is D4S2408 with only 5 alleles in the Norwegian population sample. Devesse *et al.* 2018 (52) characterized two populations, the White British and the British Chinese. Their most polymorphic length-based marker differs? between the two populations, D12S391 and FGA for the White British and the British Chinese population, respectively. Their least polymorphic markers were TPOX and D4S2408 for the White British population and TPOX for the British Chinese population.

4.6 SEQUENCE VARIATION

Characterization of sequence-based alleles and allele frequencies was obtained, but unfortunately novel alleles remain to be characterized. this needs to be done and comparison to other populations as well. Optimally the new alleles variants should be compared to other studies conducted, but only Devesse *et al.* was obtained. For future projects this part of the study should be finalized. And properly characterized for novel alleles.

By characterizing the sequence-based alleles and comparing them to the length-based alleles, an increase in number of alleles is found in many markers. Eight of the twenty-seven markers did not show an increase of allele numbers between the two methods. Compared to Devesse *et al.* 2018 (52), the results in this study differ slightly. They found that the alleles increased in number by sequence variation for all markers except Penta D, Penta E, D22S1045, D16S539, TPOX, TH01, D10S1248, and D19S433. However, Novroski *et al.* 2016 (53) found sequence variation in all markers except for TPOX. This might be due to their sample size (n=777) and the fact that they analysed four different populations groups (52).

In this study the most polymorphic marker based on a sequence based allele separation method is D12S391, which is concordant with results by Devesse *et al.* 2018 (52). This Autosomal STR marker showed an increase of 49 alleles. Because sequencing increased the number of alleles for many of the markers compared to CE, and for some considerably, sequencing helps to raise the power of discrimination when comparing the DNA-profiles of reference persons with the DNA-profile of a biological stain collected at a crime scene. This means that individuals with identical length-based alleles can have different sequence-based alleles, which makes it possible to more easily separate them in DNA mixtures.

Each population can contain different sequences, as shown in this study where novel sequences were discovered. By several studies analysing different populations, we build our understanding of different allelic compositions around the world. And it enables us to compare the populations genetic data with other populations, when performing forensic genetics analysis.

5 CONCLUSION AND FUTURE PERSPECTIVES

The aim of obtaining full DNA profiles for all 231 forensic genetic markers included in the ForenSeq DNA Signature Prep kit for all samples in the Norwegian population from the research biobank was almost finalized in this study. A full autosomal STR-profile was obtained for all samples. There are approximately 20 samples left that need to be re-sequenced to complete the full 231 genetic markers for all the DNA profiles.

The validation data for the MiSeq FGx Forensic Genomic System was achieved. The data shows that the system has relatively good performance between two representative lots, with small differences, especially with STR marker performance which may indicate lot to lot variation.

The analyses also show that the ForenSeq DNA Signature Prep kit can produce reproducible genotyping results for autosomal STR-markers. The concordance in genotypes between this kit and the NGM Select kit was 99.99%, which is in line with others when comparing different reagent kits.

The sensitivity study showed that a complete DNA-profile could be obtained with 125 pg DNA. Because partial profiles with 50% of alleles were still obtainable with 15.625 pg DNA, the ForenSeq DNA Signature Prep kit is more sensitive than more established CE-based methods.

The Norwegian population database of autosomal STR allele frequencies for Norway based on fragment length was obtained during this study, which can now be used to calculate the statistical weight of evidence in cases concerning the Norwegian population.

Allele frequencies of the SB alleles are obtained but novel variants need to be further examined. Due to lack of time the only study these sequences could be compared to was Devesse et al. but they in turn compared their data to many other studies. Unfortunately, time ran out on this part of the thesis, and the aim of Defining autosomal STR-alleles based on sequence variation using the MiSeq FGx Forensic Genomic System was not completed.

By completing the Norwegian population database and obtaining the length based allele frequencies, the question raised in the start of this thesis can be answered with a few

calculations. If we recall, Mr. Petersen was accused of killing a young girl. The question was: what is the probability that the obtained DNA was deposited by Mr. Petersen. And what is the probability that another member of the population could have contributed to the mixed DNA sample obtained from under the victims' fingernails. By using the newly obtained allele frequencies, statistical calculations can now be used to calculate the weight of the evidence to help determine the answers to these questions.

In the future, studies involving the analysis of other biological trace samples, as for example fingerprints or skin cells and also DNA mixtures from several contributors should be conducted to test and validate the performance of the kit further.

Another study that could be conducted in the future would be a study analysing the sequence variation within the flanking regions of the markers. Within the markers there can be novel variant which also may help raise the power of discrimination when it comes to trace samples in a case.

In this study, the autosomal STRs were the focus, future projects could look at the other markers in the kit and design studies around them, similar to the study performed by Hussing *et al.* 2015 (56) which focused on the eye colours of the individuals they analysed.

And lastly, to further build on this study a future project might perform a similar study to Vilsen *et al.* 2018 (87), which was a stutter analysis of STRs using the MiSeq FGx Forensic Genomics System. To learn more about the stutters that form during sequencing and the rate at which they do so, and maybe learn different ways in how to identify stutters and call them correctly. This would then minimize the need of manual checking/correction of sequencing results.

6 REFERENCES

1. Butler JM. Fundamentals of Forensic DNA Typing. National Institute of Standards and Technology. Gaithersburg, Maryland, USA: Elsevier; 2010. 500 p.

2. Olaisen B. Rettsgenetikk: Store medisinske leksikon; 2009 [cited 2018 31.03]. Available from: https://sml.snl.no/rettsgenetikk

3. Sir Alec Jeffreys The Forensics Library: The Forensics Library; [cited 2018 01.04]. Available from: http://aboutforensics.co.uk/sir-alec-jeffreys/.

4. The Discovery of DNA Fingerprinting: DNA Forensics; [cited 2018 01.04]. Available from: http://www.dnaforensics.com/DNAFingerprinting.aspx.

5. Saad R. Discovery, development, and current applications of DNA identity testing. Proceedings (Baylor University Medical Center). 2005;18(2):130-3.

6. Falch-Nilsen K. Metoden som løser alt. 2013.

7. Ommundsen EG. Felles av genetiske avtrykk. dagsavisen. 2013.

8. Valones MAA, Guimarães RL, Brandão LAC, de Souza PRE, de Albuquerque Tavares Carvalho A, Crovela S. Principles and applications of polymerase chain reaction in medical diagnostic fields: a review. Brazilian Journal of Microbiology. 2009;40(1):1-11.

9. Ginther C, Issel-Tarver L, King MC. Identifying individuals by sequencing mitochondrial DNA from teeth. Nature genetics. 1992;2(2):135-8.

10. Sullivan KM, Hopgood R, Gill P. Identification of human remains by amplification and automated sequencing of mitochondrial DNA. International journal of legal medicine. 1992;105(2):83-6.

11. Genetic fingerprinting explained University of Leicester; [cited 2018 01.04]. Available from: https://www2.le.ac.uk/departments/genetics/jeffreys/explained.

12. Panneerchelvam S, Norazmi MN. Forensic DNA Profiling and Database. The Malaysian Journal of Medical Sciences : MJMS. 2003;10(2):20-6.

13. Børresen-Dale A-L. Dna Typing I Store medisinske leksikon2014 [cited 2018 26.02]. Available from: https://sml.snl.no/DNA-typing.

14. David Carmody JS, Siri Atma W. Greeley, Graeme I. Bell, Louis H. Philipson. Genetic Diagnosis of Endocrine Disorders (Second edition). 2 ed: Academic Press; 2016.

15. Forensic DNA Testing Systems, Short Tandem Repeats (STRs) [cited 2018 26.02]. Available from: http://www.forensicdnacenter.com/index.html

16. Advances in Forensics Provide Creative Tools for Solving Crimes Bulletin of the Connecticut Academy of Science and Engineering2004 [cited 2018 27.02]. 19,2:[Available from: http://www.ctcase.org/bulletin/19_2/forensics.html.

17. Autosomal DNA Profiling SecuriGene Technologies Inc [cited 2018 26.02]. Available from: https://www.securigene.com/dna-testing/autosomal-dna-profiling/.

18. Methods in DNA Analysis. Short Tandem Repeats (STR) / Microsatellites Kitalah Gigi: Kitalah Gigi; [Available from: http://kitalahgigi.blogspot.no/2014/11/methods-in-dna-analysis.html.

19. DNA-analyser, farskapstest og andre slektskapstester – faktaark Oslo Universitetssykehus2017 [cited 2018 25.04]. Available from: https://oslo-universitetssykehus.no/fag-og-forskning/nasjonale-og-regionale-tjenester/rettsmedisinske-fag/dna-analyser/dna-analyser-farskapstest-og-andre-slektskapstester-faktaark.

20. Bishop A, Schiestl R. Homologous Recombination and Its Role in Carcinogenesis2002. 75-85 p.

21. Viguera E, Canceill D, Ehrlich S. Replication slippage involves DNA polymerase pausing and dissociation. The EMBO Journal. 2001;20(10):2587-95.

22. Butler JM. Advanced Topics in Forensic DNA Typing: Methodology. 1 ed. National Institute of Standards and Technology, Gaithersburg, Maryland, USA: Academic Press; 2011. 704 p.

23. Schneider PM. Expansion of the European Standard Set of DNA Database Loci—the Current Situation2009.

24. Martin PD, Schmitter H, Schneider PM. A brief history of the formation of DNA databases in forensic science within Europe. Forensic Science International. 2001;119(2):225-31.

25. Gill P, Fereday L, Morling N, Schneider PM. The evolution of DNA databases—Recommendations for new European STR loci. Forensic Science International. 2006;156(2):242-4.

26. Welch LAG, P. Phillips, C. Ansell, R. Morling, N. Parson, W. Palo, J. U. Bastisch, I. European Network of Forensic Science Institutes (ENFSI): Evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers. Forensic Science International: Genetics. 2012;6(6):819-26.

27. Canturk KM, Gurkan C, Sevay H, Emre R. Evaluation of the genetic parameters for 10 common and five new ESS core autosomal STR loci in seven major geographic regions and the largest metropolitan province of Turkey. Annals of Human Biology. 2017;44(2):149-63.

28. Chakraborty R. Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Human biology. 1992;64(2):141-59.

29. Steele CD, Balding DJ. Choice of population database for forensic DNA profile analysis. Science & justice : journal of the Forensic Science Society. 2014;54(6):487-93.

30. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.

31. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(2):560-4.

32. Tucker T, Marra M, Friedman JM. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. American Journal of Human Genetics. 2009;85(2):142-54.

33. Introduction to NGS- Learn how the technology works and what it can do for you Illumina2018 [cited 2018 28.04]. Available from: https://emea.illumina.com/science/technology/next-generation-sequencing.html?langsel=/af/.

34. Behjati S, Tarpey PS. What is next generation sequencing? Archives of Disease in Childhood Education and Practice Edition. 2013;98(6):236-8.

35. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.

36. An introduction to Next-Generation Sequencing Technology Illumina2017 [cited 2018 28.04]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

37. Eduardoff M, Santos C, de la Puente M, Gross TE, Fondevila M, Strobl C, et al. Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM. Forensic science international Genetics. 2015;17:110-21.

38. Zakariassen M. Predicting visible traits in a Norwegian population A prototype SNP panel for massive parallel sequencing [Msc]: Arctic University of Tromsø; 2016.

39. Yang Y, Xie B, Yan J. Application of Next-generation Sequencing Technology in Forensic Science. Genomics, Proteomics & Bioinformatics. 2014;12(5):190-7.

40. Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics. 2010;11(1):31-46.

41. Borsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Science International: Genetics. 2015;18:78-89.

42. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008;26:1135.

43. SOLiD® Next-Generation Sequencing Chemistry ThermoFisher Scientific [cited 2018 27.02]. Available from: https://www.thermofisher.com/no/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing/solid-next-generation-sequencing-systems-reagents-accessories/solid-next-generation-sequencing-chemistry.html.

44. Technology Overview Centre for Genomic Sciences [cited 2018 01.03]. Available from: http://cgs.hku.hk/portal/index.php/pyrosequencing/technology-overview.

45. Sequencing-by-Synthesis: Explaining the Illumina Sequencing Technology BitesizeBio [cited 2018 01.03]. Available from: https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/.

46. Kirsebom MK. Forensic DNA phenotyping SNP-based prediction of eye and hair colour in the Norwegian population Arctic University of Tromsø; 2016.

47. Targeted Next-Generation Sequencing for Forensic Genomics. illumina Illumina2015 [cited 2018 27.02]. Available from: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/appspotlights/app_spotlight_forensics.pdf

48. ForenSeq DNA Signature Prep - Reference Guide Illumina2015 [Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/forenseq/forenseq-dna-signature-prep-guide-15049528-01.pdf.

49. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC. Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample. PLoS ONE. 2013;8(2):e55089.

50. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One. 2013;8(6):e66621.

51. Ozsolak F. Third Generation Sequencing Techniques and Applications to Drug Discovery. Expert Opinion on Drug Discovery. 2012;7(3):231-43.

52. Laurence Devesse DB, Lucinda Davenport Immy Riethorst, Gabriella Mason-Buck, Denise Syndercombe Court. Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups. Forensic Science International: Genetics. 2018;34:57-61.

53. Novroski NMM, King JL, Churchill JD, Seah LH, Budowle B. Characterization of genetic sequence variation of 58 STR loci in four major population groups. Forensic science international Genetics. 2016;25:214-26.

54. Butler JM. The future of forensic DNA analysis. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2015;370(1674).

55. Butler JM, Coble MD, Vallone PM. STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic science, medicine, and pathology. 2007;3(3):200-5.

56. Hussing C, Børsting C, Mogensen HS, Morling N. Testing of the Illumina ForenSeq kit. Forensic Science International: Genetics Supplement Series.5:e449-e50.

57. Pitterl F, Schmidt K, Huber G, Zimmermann B, Delport R, Amory S, et al. Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations. International journal of legal medicine. 2010;124(6):551-8.

58. Kline MC, Hill CR, Decker AE, Butler JM. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic science international Genetics. 2011;5(4):329-32.

59. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nature genetics. 2012;44(10):1161-5.

60. Gettings K, Kiesler K, A. Faith S, Montano E, Baker C, Young B, et al. Sequence variation of 22 autosomal STR loci detected by next generation sequencing2015.

61. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic science international Genetics. 2015;18:118-30.

62. Validation Guidelines for DNA Analysis Methods: Scientific Working Group on DNA Analysis Methods (SWGDAM); 2016 [updated 12.05.2016; cited 2018 25.04]. Available from: https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf.

63. Recommended Minimum Criteria for the Validation of Various- Aspects of the DNA Profiling Process ENFSI DNA WORKING GROUP2010 [cited 2018 05.05]. Available from: http://enfsi.eu/wp-content/uploads/2016/09/minimum_validation_guidelines_in_dna_profiling_-_v2010_0.pdf.

64. De Souza YG, Greenspan JS. Biobanking Past, Present and Future: Responsibilities and Benefits. AIDS (London, England). 2013;27(3):303-12.

65. What is a biobank? Genetic Alliance, Registry and BioBank [Available from: http://www.biobank.org/toolbox/start_a_biobank/basics_and_benefits.

66. QIAamp® DNA Investigator Handbook Protocol: Isolation of Total DNA from Small Volumes of Blood or Saliva: QIAGEN; 2012 [cited 2018 29.03]. Available from: https://www.qiagen.com/kr/resources/resourcedetail?id=dcc5a995-3743-4219-914d-94d6a28e49b3&lang=en.

67. Quantifiler™ HP and Trio DNA Quantification Kits USER GUIDE revision G Thermo Fisher Scientific 2017 [cited 2018 27.03]. Available from: https://tools.thermofisher.com/content/sfs/manuals/4485354.pdf.

68. INNOQUANT® HY. COMPATIBLE WITH COMMONLY USED REAL-TIME PCR INSTRUMENTS & SOFTWARE InnoGenomics [cited 2018 24.04]. Available from: http://innogenomics.com/products/innoquant-hy/.

69. ForenSeq™DNA Signature Prep Reference Guide: Illumina; 2015 [cited 2018 24.02]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/forenseq/forenseq-dna-signature-prep-guide-15049528-01.pdf.

70. Applied Biosystems Veriti™ Thermal Cycler User Guide Rev. E: Thermofisher; 2010 [cited 2018 30.03]. Available from: https://tools.thermofisher.com/content/sfs/manuals/cms_042832.pdf.

71. Qubit® dsDNA HS Assay Kits B.0 Thermo Fisher Scientific Inc; 2015 [cited 2018 23.03]. Available from: https://tools.thermofisher.com/content/sfs/manuals/Qubit_dsDNA_HS_Assay_UG.pdf.

72. Focused forensic power Illumina [cited 2018 24.04]. Available from: https://emea.illumina.com/systems/sequencing-platforms/miseq-fgx.html?langsel=/af/.

73. MiSeq FGx™ Instrument Reference Guide: Illumina; 2015 [cited 2018 20.03]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq-fgx/miseq-fgx-instrument-ref-guide-15050524-c.pdf.

74. Sequencing Technology Video [cited 2018 02.03]. Available from: https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html

75. NGS technology CeGaT [Available from: https://www.cegat.de/en/services/next-generation-sequencing/.

76. Illumina Sequencing Technology Illumina: Illumina; [cited 2018 21.03]. Available from: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

77. Sharma G. What is the mechanism that cleave reverse strand on flow cell and leaving forward strand as template of sequencing in Illumina's platform? ResearchGate2015 [cited 2018 24.04]. Available from: https://www.researchgate.net/post/What_is_the_mechanism_that_cleave_reverse_strand_on_flow_cell_and_leaving_forward_strand_as_template_of_sequencing_in_Illuminas_platform.

78. ForenSeq™ Universal Analysis Software Guide Illumina: Illumina; 2016 [cited 2018 20.03]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/forenseq-universal-analysis-software/forenseq-universal-analysis-software-guide-15053876-01.pdf.

79. Walsh PS, Fildes NJ, Reynolds R. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Research. 1996;24(14):2807-12.

80. Gouy A, Zieger M. STRAF—A convenient online tool for STR data evaluation in forensic genetics. Forensic Science International: Genetics. 2017;30:148-51.

81. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nature genetics. 2008;40(5):646-9.

82. Barker DF. What is the difference between Linkage Equilibrium and Linkage Disequilibrium? Research Gate 2013 [cited 2018 20.04]. Available from: https://www.researchgate.net/post/What_is_the_difference_between_Linkage_Equilibrium_and_Linkage_Disequilibrium.

83. Just RS, Moreno LI, Smerick JB, Irwin JA. Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. Forensic Science International: Genetics. 2017;28:1-9.

84. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, et al. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Science International: Genetics. 2016;22:54-63.

85. Xavier C, Parson W. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer. Forensic Science International: Genetics. 2017;28:188-94.

86. Churchill JD, Schmedes SE, King JL, Budowle B. Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. Forensic Science International: Genetics. 2016;20:20-9.

87. Vilsen SB, Tvedebrink T, Eriksen PS, Bøsting C, Hussing C, Mogensen HS, et al. Stutter analysis of complex STR MPS data. Forensic Science International: Genetics. 2018;35:107-12.

88. Hussing C, Huber C, Bytyci R, Mogensen HS, Morling N, Børsting C. Sequencing of 231 forensic genetic markers using the MiSeq FGx™ forensic genomics system – an evaluation of the assay and software. Forensic Sciences Research. 2018:1-13.

89. García O, Alonso J, Cano JA, García R, Luque GM, Martín P, et al. Population genetic data and concordance study for the kits Identifiler, NGM, PowerPlex ESX 17 System and Investigator ESSplex in Spain. Forensic Science International: Genetics. 2012;6(2):e78-e9.

90. Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, et al. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Science International: Genetics. 2016;21:15-21.

91. Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. PCR methods and applications. 1991;1(1):17-24.

92. Alaeddini R, Walsh SJ, Abbas A. Forensic implications of genetic analyses from degraded DNA--a review. Forensic science international Genetics. 2010;4(3):148-57.

93. Diegoli TM, Farr M, Cromartie C, Coble MD, Bille TW. An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA. Forensic science international Genetics. 2012;6(4):498-503.

94. Smith PJ, Ballantyne J. Simplified low-copy-number DNA analysis by post-PCR purification. Journal of forensic sciences. 2007;52(4):820-9.

95. Coble MD, Butler JM. Characterization of new miniSTR loci to aid analysis of degraded DNA. Journal of forensic sciences. 2005;50(1):43-53.

96. Butler JM, Shen Y, McCord BR. The development of reduced size STR amplicons as tools for analysis of degraded DNA. Journal of forensic sciences. 2003;48(5):1054-64.

97. Forster L, Thomson J, Kutranov S. Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number" (LCN) method for analysis of trace forensic DNA samples. Forensic science international Genetics. 2008;2(4):318-28.

98. Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, et al. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Science International: Genetics. 2017;28:52-70.

99. Fattorini P, Previdere C, Carboni I, Marrubini G, Sorcaburu-Cigliero S, Grignani P, et al. Performance of the ForenSeq(TM) DNA Signature Prep kit on highly degraded samples. Electrophoresis. 2017;38(8):1163-74.

100. Ewing MM, Thompson JM, McLaren RS, Purpero VM, Thomas KJ, Dobrowski PA, et al. Human DNA quantification and sample quality assessment: Developmental validation of the PowerQuant¨r) system. Forensic Science International: Genetics. 2016;23:166-77.

7 APPENDIX

Appendix I – Sequencing criteria for completed DNA profiles.

Kriterier for konferering av NGS-resultater

DNA-profilene skal være rene/fra en person

STRer:

Ekstra alleler i stutter posisjon fjernes hvis de er ca. på 10-15% av nabotoppen.

Hvis det er et allel som er under den stokastiske terskelen i markører der det kan forekommer to alleler, tas ikke allelet med, og det legges en kommentar i prøveoversikten.

For markører med manglende resultat legges en kommentar i prøveoversikten.

SNPer:

For markører med manglende resultat: en kommentar legges i prøveoversikten.

Hvis markøren har et allel ≤ 30 reads fjernes resultatet for denne markøren (pga. mulig dropout). En kommentar legges i prøveoversikten.

Hvis markøren har to alleler ≤ 30 reads eller et allel > 30 reads og et allel < 30 reads og allelbalansen er noenlunde, tas begge allelene med.

Hvis et av allelene er < 10 %, tas dette allelet ikke med, og det legges en kommentar i prøveoversikten.

Hvis et av allelene er mellom 10-20 % legges en kommentar i prøveoversikten.

Kommentarer i prøveoversikten:

S-nummer

Markørtype (aSTR, Y-STR, X-STR, iSNP, pSNP, aSNP)

Markørnavn

Eventuelle kommentarer (uten kommentar = mangler resultat for denne markøren)

Hvis det foreligger flere resultater for en prøve, sammenlignes disse for å se om alle markører har resultater sammenlagt.

Appendix II- Pairways Linkage Disequilibrium p-values calculated using STRAF.

	D1S16564	TPOX	D2S441	D2S1338	D3S1358	D4S2408	FGA	D5S818	CSF1PO	D6S1043	D7S820	D8S1179	D9S1122	D10S1248	TH01	vWA	D12S391	D13S317	PentaE	D16S539	D17S1301	D18S51	D19S433	D20S482	D21S11	PentaD	D22S1045
D22S1045	0.8896	0.8309	0.4305	0.3157	0.1177	0.4024	0.8718	0.0000	0.7694	0.0000	0.6800	0.3677	0.9944	0.5067	0.6483	0.4490	0.1244	0.1210	0.8737	0.6158	0.9972	0.9987	0.9805	0.0550	0.9140	0.0156	
PentaD	0.9815	0.0000	0.9830	0.1121	0.8560	0.0084	0.0000	0.1178	0.1849	0.0002	0.6604	0.9847	0.5869	0.4076	0.0175	0.6807	0.5415	0.9508	0.0470	0.4941	0.9100	0.0004	0.0000	0.0026	0.9977		
D21S11	0.7229	0.8821	0.5855	0.6006	0.0622	0.2684	0.0000	0.9840	0.1373	0.9895	0.9614	0.0000	1.0000	0.7896	0.9228	0.2472	0.0000	0.9996	0.0627	0.8722	0.9910	0.9445	0.8352	0.9527			
D20S482	0.9476	0.5100	0.9952	0.1275	0.9781	0.0821	0.2712	0.9655	0.0209	0.8800	0.0004	0.3020	0.0000	0.9351	0.3659	0.6902	0.0527	0.9253	0.3440	0.7425	0.0223	0.0054	0.1790				
D19S433	0.8163	0.0000	0.9984	0.8514	0.5900	0.5098	0.0000	0.8698	0.9278	0.3602	0.7443	0.9200	0.6308	0.1374	0.8153	0.0759	0.9855	0.4561	0.9124	0.0020	0.6204	0.1699					
D18S51	0.0000	0.5783	0.8817	0.2922	0.9581	0.6886	0.8273	0.9979	0.8943	0.1846	0.9932	0.2646	0.7964	0.8758	0.9751	0.7160	0.6532	0.3710	0.4425	0.8703	0.0000						
D17S1301	0.9968	0.8130	0.0030	0.4201	0.8884	0.0248	0.5768	0.8221	0.4485	0.1574	0.9919	0.4350	0.7206	0.6368	0.0625	0.0000	0.0177	0.7550	0.0204	0.9796							
D16S539	0.6438	0.6269	0.3210	0.2712	0.3826	0.7495	0.3979	0.2267	0.1358	0.9919	0.7769	0.6249	0.6538	0.7747	0.5142	0.7231	0.8327	0.0385	0.3057								
PentaE	0.2420	0.0000	0.1300	0.1761	0.0621	0.2239	0.0004	0.4803	0.7237	0.3870	0.0001	0.2864	0.6624	0.6793	0.1278	0.8471	0.8282	0.0991									
D13S317	0.9047	0.9925	0.3364	0.8688	0.7404	0.6331	0.6043	0.8964	0.8201	0.8904	0.7333	0.6550	0.7577	0.3822	0.7077	0.4981	0.4011										
D12S391	0.0329	0.0550	0.9549	0.7097	0.7771	0.3279	0.0000	0.8212	0.0296	0.9955	0.9997	0.0001	0.9839	0.0001	0.0334	0.9680											
vWA	0.0610	0.0509	0.9753	0.5985	0.9013	0.2743	0.9626	0.8937	0.7816	0.5896	0.0837	0.8921	0.1364	0.6627	0.0614												
TH01	0.9983	0.0117	0.7988	0.0666	0.9631	0.4970	0.6155	0.7733	0.7312	0.4935	0.3120	0.1105	0.2700	0.3593													
D10S1248	0.9762	0.2277	0.0421	0.3021	0.9969	0.0454	0.0001	0.0534	0.1422	0.9399	0.8404	0.5832	0.4837														
D9S1122	0.9686	0.4745	0.7784	0.6584	0.7606	0.6456	0.9994	0.4566	0.3571	0.0136	0.4579	0.0519															
D8S1179	0.1482	0.7674	0.0126	0.7143	0.7976	0.9449	0.0121	0.4940	0.0000	0.5370	0.5410																
D7S820	0.0090	0.3148	0.2429	0.2788	0.9783	0.1704	0.8610	0.0405	0.9934	0.9432																	
D6S1043	0.2890	0.0310	0.0000	0.1669	0.9429	0.1836	0.0123	0.0000	0.8047																		
CSF1PO	0.9831	0.9764	0.9127	0.3836	0.8267	0.1399	0.9999	0.5903																			
D5S818	0.0229	0.8852	0.9156	0.5950	0.9250	0.3713	0.0402																				
FGA	0.9625	0.0000	0.9357	0.0433	0.8436	0.0000																					
D4S2408	0.7622	0.0011	0.5167	0.7430	0.0679																						
D3S1358	0.6828	0.2741	0.3197	0.5248																							
D2S1338	0.8429	0.0067	0.5779																								
D2S441	0.2736	0.8013																									
TPOX	0.6459																										
D1S16564																											

Appendix III- Length based allele frequencies (obtained by STRAF).

	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D17S1301	D18S51	D19S433	D1S1656	D20S482	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D4S2408	D5S818	D6S1043	D7S820	D8S1179	D9S1122	FGA	PentaD	PentaE	TH01	TPOX	vWA
2.2																							0.001				
5																								0.074	0.001		
6																			0.001				0.011		0.215	0.001	
7						0.001	0.001										0.003		0.019				0.008	0.153	0.202	0.001	
8	0.001			0.143	0.012							0.002				0.254	0.002		0.148	0.009			0.018	0.012	0.096	0.512	
9	0.020			0.076	0.143	0.001	0.001		0.001	0.009						0.338	0.044		0.166	0.012	0.001		0.206	0.020	0.144	0.095	
9.1														0.001													
9.3																									0.335		
10	0.247			0.075	0.073	0.020	0.013	0.005	0.001	0.001				0.206		0.220	0.059	0.011	0.223	0.083	0.036		0.135	0.096	0.006	0.060	
11	0.287	0.005		0.267	0.300	0.298	0.007	0.003	0.080	0.025		0.140		0.366	0.008	0.173	0.404	0.337	0.223	0.071	0.226		0.114	0.101		0.291	
11.1																			0.007								
11.3														0.032													
12	0.326	0.039		0.292	0.256	0.449	0.135	0.069	0.118	0.021		0.018		0.047		0.015	0.319	0.231	0.162	0.129	0.369		0.249	0.165		0.040	
12.2								0.004																			
13	0.095	0.321		0.105	0.185	0.187	0.123	0.227	0.055	0.229		0.003		0.035	0.003		0.161	0.069	0.044	0.336	0.314		0.181	0.065			0.001
13.2								0.018																			
13.3						0.001																					
14	0.013	0.294		0.041	0.029	0.039	0.186	0.365	0.084	0.440		0.041		0.271	0.086		0.006	0.057	0.006	0.218	0.047		0.060	0.060			0.082
14.1																							0.002				
14.2								0.016																			
14.3									0.002																		
15	0.010	0.170	0.052	0.003	0.002	0.004	0.136	0.188	0.133	0.209		0.371		0.033	0.276		0.003	0.012	0.001	0.105	0.004		0.013	0.060			0.061
15.2								0.049																			
15.3									0.073																		
16		0.126	0.016				0.121	0.037	0.122	0.063		0.334	0.043	0.008	0.262			0.003		0.032	0.004		0.001	0.074			0.169

	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D17S1301	D18S51	D19S433	D1S1656	D20S482	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D4S2408	D5S818	D6S1043	D7S820	D8S1179	D9S1122	FGA	PentaD	PentaE	TH01	TPOX	vWA
16.2								0.019																			
16.3									0.065																		
17		0.044	0.132				0.122	0.001	0.039	0.003		0.085	0.187		0.204			0.050		0.005		0.001	0.001	0.059			0.301
17.2								0.001																			
17.3			0.019						0.155																		
18		0.001	0.175				0.073		0.002			0.005	0.102		0.143			0.065				0.019		0.026			0.209
18.2								0.001										0.001				0.001					
18.3			0.007						0.057																		
19			0.122				0.040		0.001			0.002	0.141		0.015			0.104				0.054		0.016			0.106
19.2																						0.001					
19.3			0.008						0.012																		
20			0.108				0.015						0.148		0.004			0.048				0.144		0.005			0.027
20.3									0.001																		
21			0.105				0.013						0.022					0.009				0.194		0.014			0.005
21.2																						0.005					
21.3																		0.001									
22			0.097				0.007						0.033									0.163					
22.2																						0.007					
23			0.087				0.003						0.108					0.001				0.146					
23.2																						0.004					
23.3																						0.001					
24			0.044				0.001						0.094									0.135					
24.2											0.001											0.001					
25			0.020				0.002						0.103									0.070					
26			0.006								0.004		0.019									0.025					

	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D17S1301	D18S51	D19S433	D1S1656	D20S482	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D4S2408	D5S818	D6S1043	D7S820	D8S1179	D9S1122	FGA	PentaD	PentaE	TH01	TPOX	vWA
27			0.001								0.034		0.001									0.006					
28											0.200											0.002					
29											0.222																
29.2											0.001																
30											0.239																
30.2											0.038																
31											0.071																
31.2											0.073																
32											0.009																
32.2											0.067																
33											0.001																
33.2											0.032																
34											0.001																
34.2											0.006																
35											0.001																

APPENDIX IV - Characterization of sequence based autosomal STR alleles including allele frequencies.

If a novel allele is characterized, it is marked red in the table, novel alleles are only checked up against Devesse *et al.* 2018 (52). Unfortunately, because of time restraint, not all markers where checked to see if there were any novel alleles. These markers include, FGA, D12, D19 and vWa.

STR	Allele	Bracket sequence	Frequency
CSF1PO*	8	[AGAT]8	0,00093
	9	[AGAT]9	0,02037
	10	[AGAT]10	0,24630
	11	[AGAT]3[ATAT][AGAT]7	0,00093
	11	[AGAT]11	0,28704
	12	[AGAT]12	0,32593
	13	[AGAT]13	0,09537
	14	[AGAT]14	0,01296
	15	[AGAT]14	0,01019
D10S1248	11	[GGAA]11	0,00463
	12	[GGAA]12	0,03889
	13	[GGAA]13	0,32130
	14	[GGAA]14	0,29444
	15	[GGAA]15	0,16944
	16	[GGAA]16	0,12593
	17	[GGAA]17	0,04444
	18	[GGAA]18	0,00093
D12S391	15	[AGAT]8[AGAC]6AGAT	0,05185
	16	[AGAT]9[AGAC]6AGAT	0,01574
	17	[AGAT]9[AGAC]7AGAT	0,00741
	17	[AGAT]10[AGAC]7	0,00093
	17	[AGAT]10[AGAC]6AGAT	0,11944
	17	[AGAT]11[AGAC]5AGAT	0,00463
	17,3	[AGAT]GAT[AGAT]8[AGAC]7AGAT	0,01852
	18	[AGAT]10[AGAC]7	0,00278
	18	[AGAT]11[AGAC]6AGAT	0,15926
	18	[AGAT]12[AGAC]5AGAT	0,01296
	18,3	[AGAT]GAT[AGAT]9[AGAC]7AGAT	0,00741
	19	[AGAT]11[AGAC]9	0,00093
	19	[AGAT]11[AGAC]7AGAT	0,02593
	19	[AGAT]12[AGAC]6AGAT	0,08889
	19	[AGAT]13[AGAC]5AGAT	0,00648

	19,3	[AGAT]5GAT[AGAT]7[AGAC]6AGAT	0,00278
	19,3	[AGAT]GAT[AGAT]10[AGAC]7AGAT	0,00556
	20	[AGAT]11[AGAC]9	0,02963
	20	[AGAT]11[AGAC]8AGAT	0,00185
	20	[AGAT]12[AGAC]8	0,00741
	20	[AGAT]12[AGAC]7AGAT	0,0287
	20	[AGAT]13[AGAC]7	0,00278
	20	[AGAT]13[AGAC]6AGAT	0,03426
	20	AGGT[AGAT]10[AGAC]9	0,0037
	21	[AGAT]11[AGAC]10	0,0037
	21	[AGAT]12[AGAC]9	0,05926
	21	[AGAT]12[AGAC]8AGAT	0,00648
	21	[AGAT]13[AGAC]8	0,01019
	21	[AGAT]13[AGAC]7AGAT	0,01481
	21	[AGAT]14[AGAC]7	0,00093
	21	[AGAT]14[AGAC]6AGAT	0,00926
	22	[AGAT]11[AGAC]10AGAT	0,00093
	22	[AGAT]12[AGAC]10	0,02037
	22	[AGAT]12[AGAC]9AGAT	0,00278
	22	[AGAT]13[AGAC]9	0,05278
	22	[AGAT]13[AGAC]8AGAT	0,00741
	22	[AGAT]14[AGAC]8	0,01019
	22	[AGAT]14[AGAC]7AGAT	0,00185
	22	[AGAT]15[AGAC]6AGAT	0,00093
	23	[AGAT]12[AGAC]11	0,00463
	23	[AGAT]12[AGAC]10AGAT	0,00278
	23	[AGAT]13[AGAC]10	0,00833
	23	[AGAT]13[AGAC]9AGAT	0,01019
	23	[AGAT]14[AGAC]9	0,03796
	23	[AGAT]14[AGAC]8AGAT	0,01944
	23	[AGAT]15[AGAC]8	0,00278
	23	[AGAT]16[AGAC]6AGAT	0,00093
	24	[AGAT]12[AGAC]12	0,00093
	24	[AGAT]13[AGAC]11	0,0037
	24	[AGAT]13[AGAC]10AGAT	0,00093
	24	[AGAT]14[AGAC]10	0,00926
	24	[AGAT]14[AGAC]9AGAT	0,00278
	24	[AGAT]15[AGAC]9	0,01481
	24	[AGAT]15[AGAC]8AGAT	0,00926
	24	[AGAT]16[AGAC]8	0,00093
	24	[AGGT]15[AGAC]8AGAT	0,00185
	25	[AGAT]14[AGAC]11	0,00463
	25	[AGAT]14[AGAC]10AGAT	0,00093
	25	[AGAT]15[AGAC]10	0,00278
	25	[AGAT]16[AGAC]9	0,0037
	25	[AGAT]16[AGAC]8AGAT	0,00741

	25	[AGGT][AGAT]15[AGAC]8AGAT	0,00093
	26	[AGAT]17[AGAC]9	0,0037
	26	[AGAT]17[AGAC]8AGAT	0,00185
	27	[AGAT]18[AGAC]8AGAT	0,00093

D13S317	8	[TATC]8 [AATC]2 [ATCT]3 TTC [TGTC]2	0,14259
	9	[TATC]9 [AATC]2 [ATCT]3 TTC [TGTC]2	0,07593
	10	[TATC]10 [AATC]2 [ATCT]3 TTC [TGTC]2	0,05741
	10	[TATC]10 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,01667
	10	[TATC]10 [TATC]2 [AATC][ATCT]2 TTC [TGTC]2	0,00093
	11	[TATC]11 [AATC]2 [ATCT]3 TTC [TGTC]2	0,10463
	11	[TATC]11 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,16111
	11	[TATC]11 [TATC]2 [ATCT]3 TTC [TGTC]2	0,00093
	12	[TATC]12 [AATC]2 [ATCT]3 TTC [TGTC]2	0,15370
	12	[TATC]12 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,13796
	13	[TATC]13 [AATC]2 [ATCT]3 TTC [TGTC]2	0,07222
	13	[TATC]13 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,03241
	14	[TATC]14 [AATC]2 [ATCT]3 TTC [TGTC]2	0,03056
	14	[TATC]14 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,01019
	15	[TATC]15 [AATC]2[ATCT]3 TTC [TGTC]2	0,00185
	15	[TATC]15 [TATC][AATC][ATCT]3 TTC [TGTC]2	0,00093

D16S539	8	[GATA]8	0,01204
	9	[GATA]9	0,14167
	10	[GATA]10	0,07315
	11	[GATA]11	0,30093
	12	[GATA]12	0,25648
	12	[GATA][GAGA][GATA]10	0,00093
	13	[GATA]13	0,18426
	14	[GATA]14	0,02870
	15	[GATA]15	0,00185

D17S1301	7	[AGAT]7	0,00093
	9	[AGAT]9	0,00093
	10	[AGAT]10	0,02035
	11	[AGAT]11	0,29602
	11	[AGAT]10[CGAT]	0,00093
	12	[AGAT]12	0,44496
	12	[AGAT]11[CGAT]	0,00185
	12	[AGAT]10[AGGT][AGAT]	0,00093
	13	[AGAT]13	0,18686
	13,3	[AGAT]3GAT[AGAT]10	0,00093
	14	[AGAT]14	0,04163
	15	[AGAT]15	0,00370

D18S51	7	[AGAA]7 AAAG AGAG AG	0,00093
	9	[AGAA]9 AAAG AGAG AG	0,00093
	10	[AGAA]10 AAAG AGAG AG	0,01296
	11	[AGAA]11 AAAG AGAG AG	0,00741
	12	[AGAA]12 AAAG AGAG AG	0,13519
	13	[AGAA]13 AAAG AGAG AG	0,12315
	14	[AGAA]14 AAAG AGAG AG	0,18241
	14	[AGAA][AGCA][AGAA]12 AAAG AGAG AG	0,00370
	15	[AGAA]15 AAAG AGAG AG	0,13611
	16	[AGAA]16 AAAG AGAG AG	0,12037
	17	[AGAA]17 AAAG AGAG AG	0,12222
	18	[AGAA]18 AAAG AGAG AG	0,07315
	19	[AGAA]19 AAAG AGAG AG	0,04074
	20	[AGAA]20 AAAG AGAG AG	0,01481
	21	[AGAA]21 AAAG AGAG AG	0,01296
	22	[AGAA]22 AAAG AGAG AG	0,00741
	23	[AGAA]23 AAAG AGAG AG	0,00278
	24	[AGAA]24 AAAG AGAG AG	0,00093
	25	[AGAA]25 AAAG AGAG AG	0,00185

D19S433*	10	[AAGG][AAAG][AAGG][TAGG][AAGG]8 AGAG AGGA AGAA AGAG AG	0,00463
	11	[AAGG][AAAG][AAGG][TAGG][AAGG]9 AGAG AGGA AGAA AGAG AG	0,00278
	12	[AAGG][AAAG][AAGG][TAGG][AAGG]7[AAAG][AAGG]2 AGAG AGGA AGAA AGAG AG	0,00093
	12	[AAGG][AAAG][AAGG][TAGG][AAGG]10 AGAG AGGA AGAA AGAG AG	0,06759
	12,2	[AAGG][AA][AAGG][TAGG][AAGG]11 AGAG AGGA AGAA AGAG AG	0,00370
	13	[AAGG][AAAG][AAGG][TAGG][AAGG]11 AGAG AGGA AGAA AGAG AG	0,22685
	13,2	[AAGG][AA][AAGG][TAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	0,01759
	14	[AAGG][AAAG][AAGG][TAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	0,00833
	14	[AAGG][AAAG][AAGG][AAGG][AAGG]12 AGAG AGGA AGAA AGAG AG	0,35556
	14,2	[AAGG][AA][AAGG][TAGG][AAGG]13 AGAG AGGA AGAA AGAG AG	0,01667
	15	[AAGG][AAAG][AAGG][TAGG][AAGG]13 AGAG AGGA AGAA AGAG AG	0,18796
	15,2	[AAGG][AA][AAGG][TAGG][AAGG]14 AGAG AGGA AGAA AGAG AG	0,04907
	16	[AAGG][AAAG][AAGG][TAGG][AAGG]14 AGAG AGGA AGAA AGAG AG	0,03704
	16,2	[AAGG][AA][AAGG][TAGG][AAGG]15 AGAG AGGA AGAA AGAG AG	0,01852
	17	[AAGG][AAAG][AAGG][TAGG][AAGG]15 AGAG AGGA AGAA AGAG AG	0,00093
	17,2	[AAGG][AA][AAGG][TAGG][AAGG]16 AGAG AGGA AGAA AGAG AG	0,00093
	18,2	[AAGG][AA][AAGG][TAGG][AAGG]17 AGAG AGGA AGAA AGAG AG	0,00093

D1S1656*	9	[TAGA]9 [TG]5	0,00093
	10	[TAGA]9 TAGG [TG]5	0,00093
	11	[TAGA]11 [TG]5	0,07315
	11	[TAGA]10 TAGG [TG]5	0,00556
	12	[TAGA]12 [TG]5	0,05278
	12	[TAGA]11 TAGG [TG]5	0,06481
	13	[TAGA]13 [TG]5	0,02870
	13	[TAGA]12 TAGG [TG]5	0,02315

	13	[TAGA]11 TAGC TAGA [TG]5	0,00278
	14	[TAGA]13 TAGG [TG]5	0,07407
	14	[TAGA]14 [TG]5	0,01019
	14	[TAGA]12 TAGC TAGA [TG]5	0,00093
	14,3	[TAGA]2 TGA [TAGA]11 TAGG [TG]5	0,00093
	14,3	[TAGA]4 TGA [TAGA]9 TAGG [TG]5	0,00093
	15	[TAGA]14 TAGG [TG]5	0,13056
	15	[TAGA]15 [TG]5	0,00185
	15,3	[TAGA]4 TGA [TAGA]10 TAGG [TG]5	0,04074
	15,3	[TAGA]3 TGA [TAGA]11 TAGG [TG]5	0,03148
	15,3	TAGA]2 TGA [TAGA]12 TAGG [TG]5	0,00093
	16	[TAGA]15 TAGG [TG]5	0,11667
	16	[TAGA]15 TAAG [TG]5	0,00556
	16,3	[TAGA]4 TGA [TAGA]11 TAGG [TG]5	0,06389
	16,3	[TAGA]3 TGA [TAGA]12 TAGG [TG]5	0,00093
	17	[TAGA]16 TAGG [TG]5	0,03889
	17,3	[TAGA]4 TGA [TAGA]12 TAGG [TG]5	0,15556
	18	[TAGA]17 TAGG [TG]5	0,00185
	18,3	[TAGA]4 TGA [TAGA]13 TAGG [TG]5	0,05741
	19	[TAGA]18 TAGG [TG]5	0,00093
	19,3	[TAGA]4 TGA [TAGA]14 TAGG [TG]5	0,01204
	20,3	[TAGA]4 TGA [TAGA]14 TAGG [TG]5	0,00093

D20S482	9	[AGAT]9	0,00926
	10	[AGAT]10	0,00093
	11	[AGAT]11	0,02500
	12	[AGAT]12	0,02222
	13	[AGAT]13	0,22685
	14	[AGAT]14	0,44167
	15	[AGAT]15	0,20833
	16	[AGAT]16	0,06296
	17	[AGAT]17	0,00278

D21S11	24,2	[TCTA]5[TCTG]6[TCTA]3TCA[TCTA]2TCCATA[TCTA]9	0,00093
	26	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]8	0,00370
	27	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]8	0,02685
	27	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]9	0,00741
	28	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]9	0,00370
	28	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]2TCA[TCTA]2TCCATA[TCTA]10	0,00093
	28	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0,19537
	28	[TCTA]4[TCTG]7[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]9	0,00093
	29	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0,04352
	29	[TCTA]5[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,00093
	29	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0,00370
	29	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,17407

	29,2	[TCTA]5[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10TATCTA	0,00093
	30	[TCTA]7[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0,00463
	30	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,11852
	30	[TCTA]6[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	0,00093
	30	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,05741
	30	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12	0,05648
	30,2	[TCTA]5[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11TATCTA	0,00926
	30,2	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10TATCTA	0,02870
	31	[TCTA]7[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,00741
	31	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12	0,02685
	31	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12	0,02407
	31	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]13	0,01296
	31,2	[TCTA]5[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12TATCTA	0,00093
	31,2	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11TATCTA	0,07222
	32	[TCTA]7[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12	0,00093
	32	[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]13	0,00093
	32	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]13	0,00556
	32	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]14	0,00185
	32,2	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12TATCTA	0,06667
	33	[TCTA]7[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]13	0,00093
	33,2	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]13TATCTA	0,03241
	34	[TCTA]10[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	0,00093
	34,2	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]14TATCTA	0,00556
	35	[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]12TA[TCTA]2TATCTA	0,00093

D22S1045	8	[ATT]5 ACT [ATT]2	0,00185
	11	[ATT]8 ACT [ATT]2	0,13981
	12	[ATT]9 ACT [ATT]2	0,01852
	13	[ATT]10 ACT [ATT]2	0,00278
	14	[ATT]11 ACT [ATT]2	0,03981
	15	[ATT]12 ACT [ATT]2	0,37222
	16	[ATT]13 ACT [ATT]2	0,33426
	17	[ATT]14 ACT [ATT]2	0,08426
	17	ATG [ATT]13 ACT [ATT]2	0,00093
	18	[ATT]15 ACT [ATT]2	0,00463
	19	[ATT]16 ACT [ATT]2	0,00093

D2S1338*	16	[TGCC]6 [TTCC]10	0,04259
	16	[TGCC]5 [TTCC]11	0,00093
	17	[TGCC]6 [TTCC]11	0,18519
	17	[TGCC]5 [TTCC]12	0,00185
	18	[TGCC]7 [TTCC]11	0,03426
	18	[TGCC]6 [TTCC]12	0,06759
	19	[TGCC]7 [TCCC] [TTCC]11	0,00093

	19	[TGCC]8 [TTCC]11	0,00093
	19	[TGCC]7 [TTCC]12	0,11019
	19	[TGCC]6 [TTCC]13	0,02685
	19	[TGCC]6 [TTCC]10 [GTCC] [TTCC]2	0,00093
	20	[TGCC]7 [TCCC] [TTCC]12	0,01944
	20	[TGCC]7 [TTCC]10 [GTCC] [TTCC]2	0,02963
	20	[TGCC]7 [TTCC]13	0,08426
	20	[TGCC]7 [TTCC]2 [TTTC] [TTCC]10	0,00278
	20	[TGCC]6 [TTCC]14	0,01204
	20	[TGCC]7 [TCCC] [TTCC]13	0,00093
	21	[TGCC]7 [TTCC]11 [GTCC] [TTCC]2	0,00556
	21	[TGCC]7 [TTCC]14	0,01574
	21	[TGCC]8 [TTCC]11 [GTCC] [TTCC]2	0,00093
	22	[TGCC]7 [TTCC]12 [GTCC] [TTCC]2	0,02593
	22	[TGCC]7 [TTCC]15	0,00370
	22	[TGCC]6 [TTCC]13 [GTCC] [TTCC]2	0,00278
	22	[TGCC]7 [TTCC]13 [GTCC] [TTCC]2	0,10556
	23	[TGCC]6 [TTCC]14 [GTCC] [TTCC]2	0,00278
	23	[TGCC]8 [TTCC]13 [GTCC] [TTCC]2	0,00556
	24	[TGCC]7 [TTCC]14 [GTCC] [TTCC]2	0,07870
	24	[TGCC]6 [TTCC]15 [GTCC] [TTCC]2	0,00833
	24	[TGCC]5 [TTCC]16 [GTCC] [TTCC]2	0,00093
	25	[TGCC]8 [TTCC]14 [GTCC] [TTCC]2	0,01389
	25	[TGCC]7 [TTCC]15 [GTCC] [TTCC]2	0,08796
	25	[TGCC]6 [TTCC]16 [GTCC] [TTCC]2	0,00093
	26	[TGCC]8 [TTCC]15 [GTCC] [TTCC]2	0,00093
	26	[TGCC]7 [TTCC]16 [GTCC] [TTCC]2	0,01667
	26	[TGCC]6 [TTCC]17 [GTCC] [TTCC]2	0,00093
	27	[TGCC]7 [TTCC]17 [GTCC] [TTCC]2	0,00093

D2S441	9,1	A[TCAT]9	0,00093
	10	[TCAT]10	0,06852
	10	[TCTA]8[TCTG][TCTA]	0,13704
	11	[TCTA]11	0,35741
	11	[TCTA]9[TCTG][TCTA]	0,00926
	11,3	[TCTA]4[TCA][TCTA]7	0,03241
	12	[TCTA]12	0,04259
	12	[TCTA]10[TCTG][TCTA]	0,00185
	12	[TCTA]9[TTTA][TCTA]2	0,00278
	13	[TCTA]13	0,01204
	13	[TCTA]11[TCTG][TCTA]	0,00093
	13	[TCTA]10[TTTA][TCTA]2	0,02222
	14	[TCTA]11[TTTA][TCTA]2	0,27037
	15	[TCTA]12[TTTA][TCTA]2	0,03333
	16	[TCTA]13[TTTA][TCTA]2	0,00833

D3S1358	11	TCTA [TCTG]2 [TCTA]8	0,00833
	13	TCTA [TCTG]2 [TCTA]10	0,00278
	14	TCTA TCTG [TCTA]12	0,00093
	14	TCTA [TCTG]2 [TCTA]11	0,08611
	15	TCTA TCTG [TCTA]13	0,02593
	15	TCTA [TCTG]2 [TCTA]12	0,24167
	15	TCTA [TCTG]3 [TCTA]11	0,00648
	16	TCTA TCTG [TCTA]14	0,01481
	16	TCTA [TCTG]2 [TCTA]13	0,17037
	16	TCTA [TCTG]3 [TCTA]12	0,07685
	17	TCTA TCTG [TCTA]15	0,00185
	17	TCTA [TCTG]2 [TCTA]14	0,11296
	17	TCTA [TCTG]3 [TCTA]13	0,08704
	17	TCTA [TCTG]4 [TCTA]12	0,00185
	18	TCTA [TCTG]2 [TCTA]15	0,02222
	18	TCTA [TCTG]3 [TCTA]14	0,12130
	19	TCTA [TCTG]2 [TCTA]16	0,00185
	19	TCTA [TCTG]3 [TCTA]15	0,01296
	20	TCTA [TCTG]3 [TCTA]16	0,00370

D4S2408	8	[ATCT]8	0,25278
	9	[ATCT]9	0,31296
	9	[ATCT] GTCT [ATCT]7	0,02593
	10	[ATCT]10	0,22037
	11	[ATCT]11	0,17315
	12	[ATCT]12	0,01481

D5S818 *	7	[AGAT]7[AGAG]	0,00278
	8	[AGAT]8[AGAG]	0,00093
	8	[AGAT]8[AGAT]	0,00093
	9	[AGAT]9[AGAG]	0,04259
	9	[AGAT]9[AGAT]	0,00093
	10	[AGAT]10[AGAG]	0,04444
	10	[AGAT]10[AGAT]	0,01481
	11	[AGAT]11[AGAG]	0,37500
	11	[AGAT]11[AGAT]	0,02685
	12	[AGAT]12[AGAG]	0,21759
	12	[AGAT]12[AGAT]	0,10278
	13	[AGAT]13[AGAG]	0,12037
	13	[AGAT]13[AGAT]	0,04074
	14	[AGAT]14[AGAG]	0,00648
	15	[AGAT]15[AGAG]	0,00278

D6 S1 04 3*	10	[AGAT]10	0,01111

	11	[AGAT]11	0,33611
	12	[AGAT]12	0,23241
	13	[AGAT]13	0,07037
	14	[AGAT]14	0,05833
	15	[AGAT]15	0,00648
	15	[AGAT]13[ACAT][AGAT]	0,00463
	15	[AGAT]9[ACAT][AGAT]5	0,00093
	16	[AGAT]14[ACAT][AGAT]	0,00278
	17	[AGAT]11[ACAT][AGAT]5	0,04907
	18	[AGAT]12[ACAT][AGAT]5	0,06574
	19	[AGAT]13[ACAT][AGAT]5	0,10370
	20	[AGAT]13[ACAT][AGAT]6	0,00278
	20	[AGAT]14[ACAT][AGAT]5	0,04537
	21	[AGAT]14[ACAT][AGAT]6	0,00556
	21	[AGAT]15[ACAT][AGAT]5	0,00370
	23	[AGAT]11[ACAT][AGAT]4[ACAT][AGAT]6	0,00093

D7S820*	6	[GATA]6 GACA GATT GATA GTTT	0,00093
	7	[GATA]7 GACA GATT GATA GTTT	0,01852
	8	[GATA]8 GACA GATT GATA GTTT	0,14722
	9	[GATA]9 GACA GATT GATA GTTT	0,16667
	10	[GATA]10 GACA GATT GATA GTTT	0,22407
	11	[GATA]11 GACA GATT GATA GTTT	0,22222
	11,1	[GATA]11 A GACA GATT GATA GTTT	0,00741
	12	[GATA]12 GACA GATT GATA GTTT	0,16204
	13	[GATA]13 GACA GATT GATA GTTT	0,04352
	14	[GATA]14 GACA GATT GATA GTTT	0,00648
	15	[GATA]15 GACA GATT GATA GTTT	0,00093

D8S1179	8	[TCTA]8	0,00926
	9	[TCTA]9	0,01204
	10	[TCTA]10	0,08333
	11	[TCTA]11	0,07130
	12	[TCTA]12	0,11852
	12	[TCTA] [TCTG] [TCTA]10	0,00926
	13	[TCTA]13	0,08426
	13	[TCTA]2 [TCTG] [TCTA]10	0,00556
	13	[TCTA] [TCTG] [TCTA]11	0,24630
	13	[TCTA] [TCTG]2 [TCTA]10	0,00093
	14	[TCTA]14	0,03056
	14	[TCTA]2 [TCTG] [TCTA]11	0,05000
	14	[TCTA] [TCTG] [TCTA]12	0,12870
	14	[TCTA] [TCTG] [TGTA] [TCTA]11	0,00833
	15	[TCTA]15	0,00093
	15	[TCTA]2 [TCTG] [TCTA]12	0,06019
	15	[TCTA] [TCTG] [TCTA]13	0,04352

	16	[TCTA]2 [TCTG] [TCTA]13	0,02778
	16	[TCTA]2 [TCTG]2 [TCTA]12	0,00185
	16	[TCTA] [TCTG] [TCTA]14	0,00278
	17	[TCTA]2 [TCTG]2 [TCTA]13	0,00093
	17	[TCTA]2 [TCTG] [TCTA]14	0,00370

D9S1122	9	TAGA TCGA [TAGA]7	0,00093
	10	TAGA TCGA [TAGA]8	0,00926
	10	[TAGA]10	0,02685
	11	[TAGA]11	0,15926
	11	TAGA TCGA [TAGA]9	0,06574
	12	[TAGA]12	0,16667
	12	TAGA TCGA [TAGA]10	0,20185
	13	[TAGA]13	0,05833
	13	TAGA TCGA [TAGA]11	0,25463
	13	TAGA TCGA [TAGA]2 TAGT [TAGA]8	0,00185
	14	[TAGA]14	0,00741
	14	TAGA TCGA [TAGA]12	0,03981
	15	[TAGA]15	0,00185
	15	TAGA TCGA [TAGA]13	0,00185
	16	TAGA TCGA [TAGA]14	0,00370

FGA*	17	[TTTC]3[TTTT][TTCT][CTTT]9[CTCC][TTCC]2	0,00093
	18	[TTTC]3[TTTT][TTCT][CTTT]10[CTCC][TTCC]2	0,02037
	18,2	[TTTC]3[TTTT][TT][CTTT]11[CTCC][TTCC]2	0,00093
	19	[TTTC]3[TTTT][TTCT][CTTT]11[CTCC][TTCC]2	0,05370
	19,2	[TTTC]3[TTTT][TT][CTTT]12[CTCC][TTCC]2	0,00093
	20	[TTTC]3[TTTT][TTCT][CTTT]12[CTCC][TTCC]2	0,14444
	21	[TTTC]3[TTTT][TTCT][CTTT]13[CTCC][TTCC]2	0,19352
	21,2	[TTTC]3[TTTT][TT][CTTT]14[CTCC][TTCC]2	0,00463
	22	[TTTC]3[TTTT][TTCT][CTTT]14[CTCC][TTCC]2	0,18333
	22,2	[TTTC]3[TTTT][TT][CTTT]15[CTCC][TTCC]2	0,00741
	23	[TTTC]3[TTTT][TTCT][CTTT]15[CTCC][TTCC]2	0,14630
	23,2	[TTTC]3[TTTT][TT][CTTT]16[CTCC][TTCC]2	0,00370
	23,3	[TTTC]3[TTTT][TTCT][CTTT]14[CTT][CTTT][CTCC][TTCC]2	0,00093
	24	[TTTC]3[TTTT][TTCT][CTTT]16[CTCC][TTCC]2	0,13611
	24,2	[TTTC]3[TTTT][TT][CTTT]17[CTCC][TTCC]2	0,00093
	25	[TTTC]3[TTTT][TTCT][CTTT]15 GTTT CTTT [CTCC][TTCC]2	0,06759
	25	[TTTC]3[TTTT][TTCT][CTTT]17[CTCC][TTCC]2	0,00278
	26	[TTTC]3[TTTT][TTCT][CTTT]18[CTCC][TTCC]2	0,02407
	27	[TTTC]3[TTTT][TTCT][CTTT]19[CTCC][TTCC]2	0,00556
	28	[TTTC]3[TTTT][TTCT][CTTT]20[CTCC][TTCC]2	0,00185

Penta D	2,2	GAAAAGA[AAAGA]2	0,00093
	6	AAAAG [AAAGA]6	0,01111
	7	AAAAG [AAAGA]7	0,00833

	8	AAAAG [AAAGA]8	0,01759
	9	AAAAG [AAAGA]9	0,20648
	10	AAAAG [AAAGA]10	0,13426
	11	AAAAG [AAAGA]11	0,11389
	12	AAAAG [AAAGA]12	0,24815
	13	AAAAG [AAAGA]13	0,18241
	14	AAAAG [AAAGA]14	0,06019
	14,1	AAAAG [AAAGA]3 A[AAAGA]11	0,00185
	15	AAAAG [AAAGA]15	0,01296
	16	AAAAG [AAAGA]16	0,00093
	17	AAAAG [AAAGA]17	0,00093

Penta E*	5	[AAAGA]5	0,07315
	7	[AAAGA]7	0,15370
	8	[AAAGA]8	0,01204
	9	[AAAGA]9	0,02037
	10	[AAAGA]10	0,09630
	11	[AAAGA]11	0,10093
	12	[AAAGA]12	0,16481
	13	[AAAGA]13	0,06481
	14	[AAAGA]14	0,06019
	15	[AAAGA]15	0,06019
	16	[AAAGA]16	0,07407
	17	[AAAGA]17	0,05926
	18	[AAAGA]18	0,02593
	19	[AAAGA]19	0,01574
	20	[AAAGA]20	0,00463
	21	[AAAGA]21	0,01389

TH01	5	[AATG]5	0,00093
	6	[AATG]6	0,21667
	7	[AATG]7	0,20185
	8	[AATG]8	0,09630
	9	[AATG]9	0,14352
	9,3	[AATG]6[ATG][AATG]3	0,33426
	10	[AATG]10	0,00648

TPOX	6	[AATG]6	0,00093
	7	[AATG]7	0,00093
	8	[AATG]8	0,51111
	9	[AATG]9	0,09630
	10	[AATG]10	0,06019
	11	[AATG]11	0,29074
	12	[AATG]12	0,03981

vWA*	13	[TCTA]2 [TCTG]4 [TCTA]3 [TCCA] [TCTA]3 TCCA TCCA	0,00093
	14	[TCTA] [TCTG] [TCTA] [TCTG]4[TCTA]3[TCCA][TCTA]3 TCCA TCCA	0,06296
	14	[TCTA] [TCTG]3[TCTA]10 TCCA TCTA	0,01667
	14	[TCTA] [TCTG]4[TCTA]9 TCCA TCTA	0,00278
	15	[TCTA] [TCTG] [TCTA] [TCTG]4[TCTA]3[TCCA] [TCTA]3[TCCA] TCCA TCCA	0,00185
	15	[TCTA] [TCTG]3[TCTA]11 TCCA TCTA	0,05926
	15	[TCTA] [TCTG]4[TCTA]10 TCCA TCTA	0,01944
	16	[TCTA] [TCTG]3[TCTA]12 TCCA TCTA	0,03241
	16	[TCTA] [TCTG]4[TCTA]11 TCCA TCTA	0,15556
	16	[TCTG]4[TCTA]12 TCCA TCTA	0,00093
	17	[ACTA] [TCTG]4[TCTA]12 TCCA TCTA	0,00093
	17	[TCTA] [TCTG]3[TCTA]13 TCCA TCTA	0,02130
	17	[TCTA] [TCTG]4[TCTA]12 TCCA TCTA	0,27963
	18	[TCTA] [TCTG]3[TCTA]14 TCCA TCTA	0,00556
	18	[TCTA] [TCTG]4[TCTA]13 TCCA TCTA	0,19722
	18	[TCTA] [TCTG]5[TCTA]12 TCCA TCTA	0,00648
	19	[TCTA] [TCTG]3[TCTA]15 TCCA TCTA	0,00185
	19	[TCTA] [TCTG]4[TCTA]14 TCCA TCTA	0,10000
	19	[TCTA] [TCTG]5[TCTA]13 TCCA TCTA	0,00185
	19	[TCTA] [TCTG]6[TCTA]12 TCCA TCTA	0,00093
	20	[TCTA] [TCTG]4[TCTA]15 TCCA TCTA	0,02685
	21	[TCTA] [TCTG]4[TCTA]16 TCCA TCTA	0,00463

Table 1: List of individual sequence-based alleles observed in the population studied. Novel variants are highlighted in red and in bold, whilst allele designations that do not follow ISFG recommendations are marked with an asterisk. Flanking region sequences reported by the software are highlighted in light grey to avoid confusion with repeat region sequences.