UiT
THE ARCTIC UNIVERSITY OF NORWAY

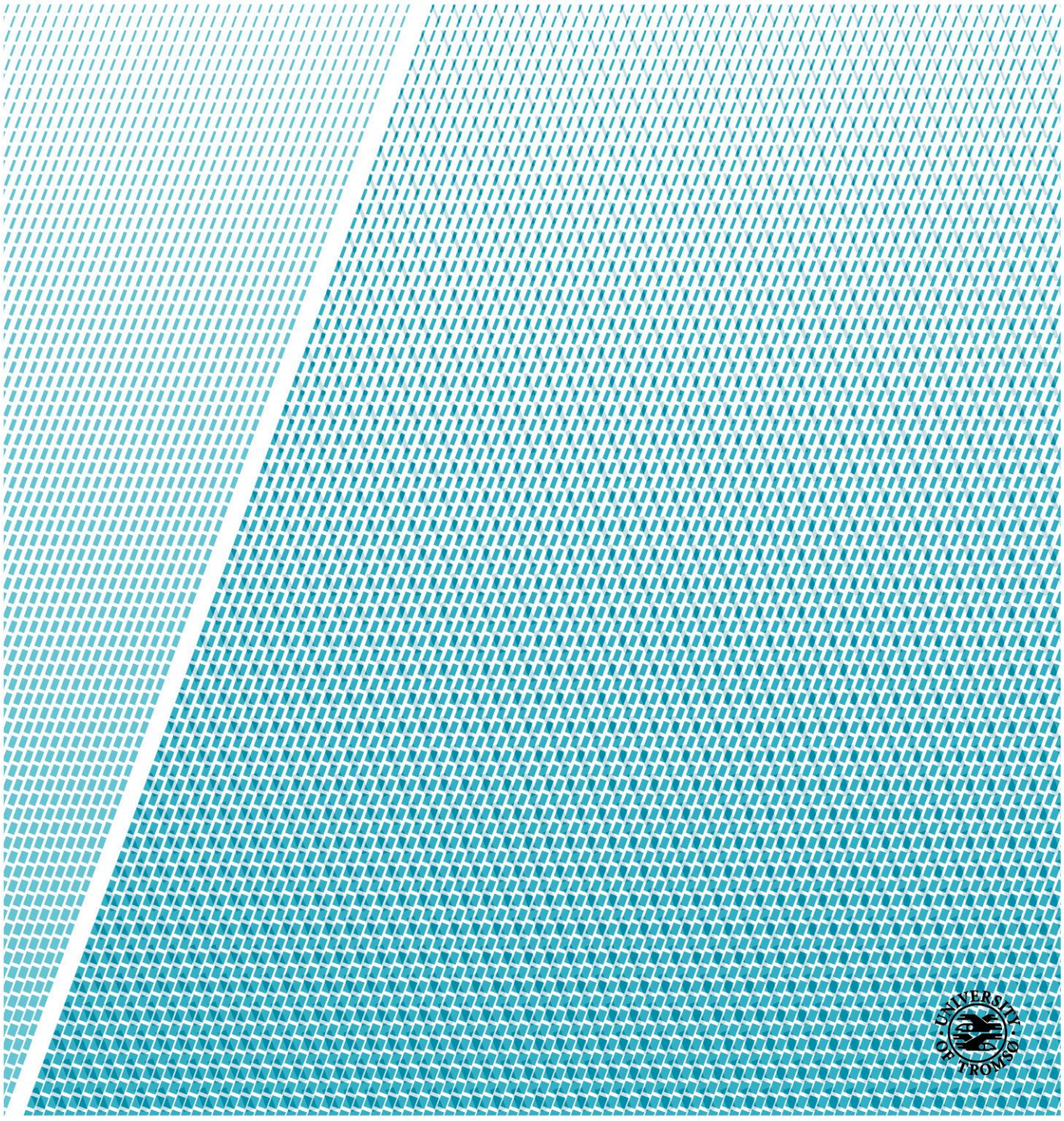
Department of Medical Biology, Faculty of Health Sciences

Analysis of 30 Y-chromosomal STR markers in the Norwegian population

—

Ingrid Støtvig

Master's thesis in Biomedicine MBI-3911. May 2019.



Abstract

Autosomal STR-analysis is the standard method of DNA-typing in casework. However, there are sample types, especially vaginal samples from sexual assaults, that may contain a large fraction of female DNA and a minimal fraction of male DNA. The male fraction may not be successfully amplified using autosomal STR-analysis because it “drowns” in the female fraction. However, the analysis of male-specific YSTR markers may circumvent this obstacle.

The purpose of this study was to analyze YSTR markers in the Norwegian population, including male individuals that are third generation Norwegian on their father’s father side. Two analysis kits based on different technologies were used. The Yfiler Plus PCR Amplification kit (Thermo Fisher Scientific) is based on polymerase chain reaction and capillary electrophoresis and obtains fragment length-based information of the loci in the kit. The ForenSeq DNA Signature Prep is sequence-based and provides both fragment length- and sequence-based information for the loci in the kit. The fragment length-based information is collected to create a foundation for a reference database of YSTR haplotypes that can be used to calculate the statistical weight of the evidence upon a DNA-profile match.

Several loci overlap between the two kits, and the concordance based on fragment length was 99.65%. Forensic parameters and diversity values were calculated for Yfiler Plus, Y-filer (the precursor of Yfiler Plus) and ForenSeq haplotypes in order to evaluate the analytical power of the kits. The ForenSeq loci proved to have the best analytical power, followed by the Yfiler Plus and Yfiler loci. These results also proved that the inclusion of rapidly mutating YSTR markers increase the analytical power. Sequencing leads to an increase in sequence-based allele variants compared to length-based allele variants. Many of the allele variants found here are apparently novel.

The Norwegian dataset proved to comprise of three major genetic substructures belonging to three haplogroups and was significantly different from twelve of thirteen European populations.

Acknowledgments

I would like to thank the Centre of Forensic Genetics at the Arctic University of Norway for this amazing learning opportunity, it has been a wonderful experience and a terrific introduction to laboratory practices used in the field of Forensic Genetics. I had the honor and privilege of being the first master student in the new laboratories and I have learned a lot.

Thank you to my co-supervisors, Gunn-Hege Olsen and Thomas Berg, and an extra special thank you to my primary supervisor Kirstin Janssen. Thank you for teaching me the laboratory work involved in this thesis and thank you for all the time and effort spent on proofreading this thesis, in addition to the literary review the first year. Thank you for pushing me to work harder and inspire me to do better.

Thank you, Richard Kessel, at Verogen, for always being available to assist in any questions regarding the ForenSeq kit.

I would also like to thank the remaining members of the Centre of Forensic Genetics, Marita Olsen, Bente Haldorsen, Nina Mjølnes Salvo, and for a brief time, Sandra Buadu. Thank you for lovely conversations, joint lunches, and cake the first Thursday of every month. I will miss you.

Lastly, thank you to my friends and family for the continuous support over the last five years, and a special thanks to my lovely partner Andreas Larsen for a stream of endless support and motivational speeches.

Ingrid Støtvig

Tromsø, May 2019

Table of contents

List of figures

List of tables

Abbreviations

AMOVA – Analysis of Molecular Variance

CE – Capillary Electrophoresis

CFG – Centre of Forensic Genetics

DC – Discriminatory Capacity

DL – Discrete Laplace

DPMA/B – DNA Primer Mix A/B

FEM – ForenSeq Enzyme Mix

GD – Genetic Diversity

HD – Haplotype Diversity

HP3 - NaOH

HSC – Human Sequencing Control

HT1 – Hybridization Buffer

ISFG – International Society for Forensic Genetics

LNA1 - Library Normalization Additives 1

LNB1 – Library Normalization Beads

LNS2 - Library Normalization Buffer 2

LNW1 - Library Normalization Wash 1

MDS – Multidimensional Scaling

MP – Matching Probability

MPS – Massive Parallel Sequencing

NGS – Next Generation Sequencing

PAR – Pseudoautosomal Regions

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

PFLP – Random Fragment Length Polymorphism

RM – Rapidly Mutating

RSU – Resuspension Buffer

SBS – Sequencing by Synthesis

SNP – Single Nucleotide Polymorphism

SPB – Sample Purification beads

STR – Short Tandem Repeat

TBE buffer – Tris-borate-EDTA Buffer

UAS – Universal Analysis Software

YHRD – Y-chromosomal STR Haplotype Reference Database

YSTR – Y-chromosomal Short Tandem Repeat

Introduction

The most common case of sexual assault is a man (or men) assaulting a woman. In 2017 the Norwegian Police received 1968 reports of assault or attempted assault (1). Of the victims, 1737 were girls or women. In the case of rape with vaginal perpetration, the perpetrator will with high probability leave DNA traces within the woman's vagina. If a vaginal sample is retrieved from the woman for DNA analysis, the vaginal sample usually contains large amounts of female DNA and a small fraction of male DNA. One can analyze this sample in two ways. First, one can perform a traditional short tandem repeat (STR) analysis of the autosomal chromosomes. Second, one can perform a STR-analysis of the Y-chromosome, called YSTR-analysis.

When performing an autosomal DNA analysis on this type of sample, the male contribution may be difficult to interpret or completely lost in the analysis. Vaginal samples usually contain large amounts of DNA, with a large fraction of female DNA and a small fraction of male DNA. For STR-analysis, the sample must be diluted based on the total DNA concentration and the male fraction may be diluted so much that it cannot be detected. However, if a YSTR-analysis was performed the sample would be diluted based only on the DNA concentration of the male fraction, because the female DNA would not be amplified. The small male DNA faction may be enough for a complete, or nearly complete DNA-profile. Figure 1 illustrates the autosomal STR and YSTR-profiles that may be observed in the two analysis for the same sample. YSTR-analysis also makes it easier to identify if there is more than one male contributor in the sample. YSTR-analysis can be carried out regardless of the cell type(s) that make up the male fraction, it can be sperm cells, white blood cells or epithelial cells. The only requirement is that the male DNA concentration is sufficient (2).

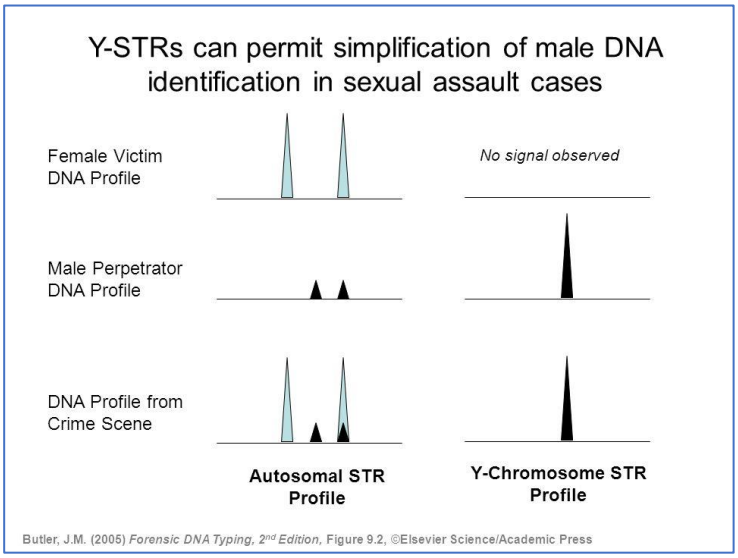


Figure 1 – Illustration of possible DNA-profiles using autosomal STR and YSTR-analysis on a vaginal sample from a sexual assault. The sample contains a large fraction of female DNA and a small fraction male DNA. Figure from (3).

YSTR-analysis can be useful for several other applications. For forensic purposes the most relevant applications are investigations of sexual assaults, testing amelogenin-deficient males, paternity testing, missing person investigations, and bio-geographical ancestry analyses. Other field of interest for YSTR-analysis are migration and evolutionary studies, and genealogical and historical research (3). When running an autosomal STR-analysis, the amelogenin marker is usually used as the sex determining marker (4, 5). The amelogenin marker can have a six bp insertion that prevents correct amplification, or the gene can be

deleted entirely (5). In these cases, YSTR analyses can be used to verify or dismiss if the contributor to the sample is male. However, in some of the newest commercial STR kits, amelogenin is usually not the only sex-determining marker in the kit. If an individual is amelogenin deficient, the analysis results of the other sex markers will assist in correct gender assessment. In motherless paternity cases, YSTR-analysis can be used to identify the father or male relatives. For missing persons investigations, male relatives can be used as reference samples.

In biogeographic ancestry analyses, the geographic origin of a person's paternal ancestors can be determined by predicting the haplogroup to which this individual's YSTR haplotype belongs to (6). A haplogroup is a collection of haplotypes assumed to descend from a shared ancestor. The haplogroups are separated by mutations in the non-recombining regions of the Y-chromosome. Biogeographic ancestry information can be useful in which where autosomal STR-profiles do not match with any reference profiles. The haplogroup information ideally combined with information about the perpetrators physical appearance and age, obtained from witness description or predictions about ancestry and phenotypical traits obtained from single nucleotide polymorphism (SNP) analyses, may guide the investigation in the correct direction (7, 8).

STR markers characteristics

Short tandem repeats (STR) are a type of non-coding repeated DNA sequences that consist of a variable number of repeat units. They are associated with constitutively heterochromatic regions of the chromosomes and can be found on most chromosomes. The analysis of autosomal STR markers is the standard method for DNA typing today. However, as explained earlier the analysis of autosomal STR markers is not always the best option for all sample types, and this thesis focuses therefore on the analysis of YSTR markers. STR markers in general differ in three ways, the length of the repeat unit, the number of repeats, and the overall repeat pattern (9, 10). STR markers are named based on the length of the repeat unit (11, 12); dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, or hexanucleotide. Simple repeats consist of a repeat unit sequence that are repeated several times (10, 12, 13). Compound repeats have two or more adjacent similar repeats. Complex repeats contain several repeat units of varying fragment length and sequence.

Thousands of STR markers in the human genome have been characterized. Ideally, one uses the STR markers that produce the lest noise (stutters), have a suitable mutation rate (not too high and not too low) and are the most polymorphic. These STR markers are usually tetranucleotides, e.g. because tetranucleotides occur more frequently than penta- and hexanucleotides, and because di- and tri-nucleotides often produce more noise (13-17). All these aspects are thoroughly checked before STR markers are selected.

STR markers have become popular in forensic genetics for two reasons. They are extensively polymorphic, and their fragment size is rather small. The extensive polymorphism make STR markers an ideal tool for differentiation of individuals, both within and across

populations and ethnic groups (12, 18). The small size makes them easy to amplify, even if the sample is somewhat degraded, something that is a very common challenge in forensic samples. The polymorphism of STR markers is caused by replication errors in the form of polymerase slippage. The large homologous regions can trigger insertions or deletions of repeat units by aligning incorrectly during DNA replication, as illustrated in Figure 2 (19, 20). The polymerase slippage leads to a large variation of number of repeat units and extensive polymorphism among and within populations making these markers ideal for forensic analysis.

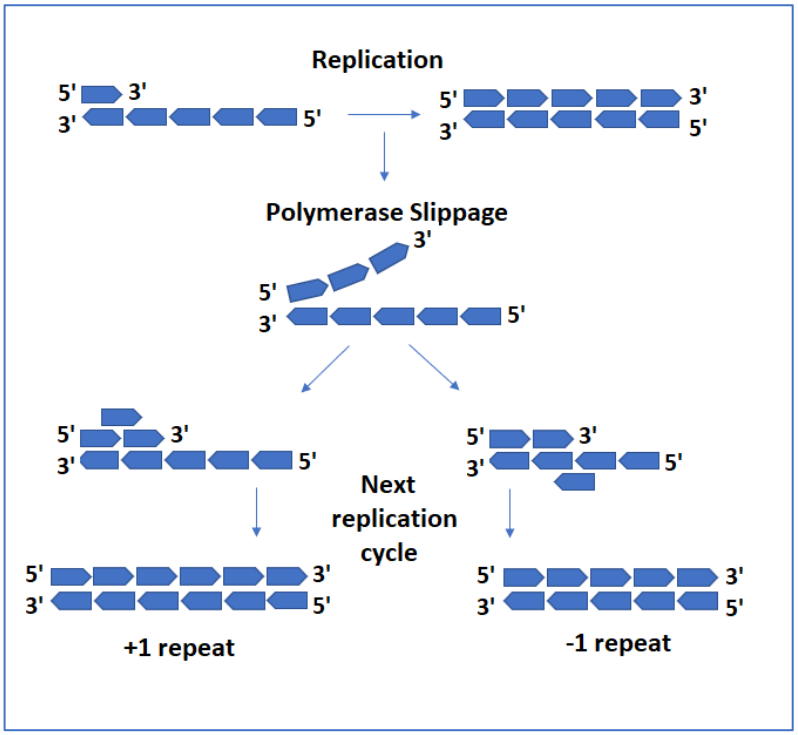


Figure 2 – Mechanism of DNA polymerase slippage. Slippage can lead to addition or subtraction of one repeat unit. Figure modified from (21).

Y-chromosomal analysis

YSTR-analysis is the analysis of a collection of multiple YSTR markers to determine a man's YSTR-profile, also called YSTR haplotype. Due to their polymorphic nature the probability that two unrelated men have the same YSTR haplotype corresponds with the incidence of this haplotype in the population. All paternally related males will have the same YSTR haplotype unless mutations occur, explained in more detail later. For optimal effectiveness across several jurisdictions a set of standardized markers are selected (described in next section). The standard analysis method is polymerase chain reaction (PCR). PCR allows creation of millions of copies of the DNA, where specific primers are used to prime specific YSTR markers than subsequently are amplified by a polymerase. The sample is then injected trough a capillary electrophoresis (CE) where the amplified DNA fragments are separated and measured. A more in dept description is presented later.

YSTR marker history

The first polymorphic YSTR markers discovered were published in 1992 and were, the same year, used to free an imprisoned man from murder and rape charges (22, 23). Up until this point only a few restriction fragment length polymorphisms (RFLP) were identified on the Y-chromosome (22). RFLP is a method that cleaves the DNA at specific restriction sites and then separated the fragments using a gel electrophoresis. Since then, more YSTR markers have been discovered and are used for DNA-typing (24-34). Reasons for the increase in YSTR marker discoveries were improvements in bioinformatical platforms and the increased availability of

sequence information provided by the Human Genome Project (35). In 1997, the European forensic community chose a set of eight standardized YSTR markers, DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393 and DYS385a/b, termed minimal haplotype group (36, 37). The set of YSTR markers were so named to represent the minimal requirement for adequate informative haplotyping in forensic casework (36). In 2003, the U.S. Scientific Working Group on DNA Analysis Methods (SWGDAM) updated the minimal haplotype collection by adding two more YSTR markers, DYS438 and DYS439 (38).

Previously, most YSTR markers used for forensic purposes had mutation rates of approx. 10^{-3} per locus per generation, allowing the separation between groups of closely and distantly related individuals in male lineages (39). However, these YSTR markers do not allow the separation of closely related individuals in the same paternal lineage. In the early 2010, a new, more rapidly mutating YSTR markers were characterized and later implemented in commercial kits (39-41). Rapidly mutating (RM) YSTR markers have a mutation rate of approx. 10^{-2} per locus per generation and have proven successful in discriminating between distantly and closely related individuals (41-43).

Several commercial kits are available to analyze YSTR markers, such as the Yfiler PCR Amplification kit (Thermo Fisher), PowerPlex Y23 System (Promega), Yfiler Plus PCR Amplification kit (Thermo Fisher) and ForenSeq DNA Signature Prep kit (Illumina/Verogen). All four kits contain the minimal haplotype group of YSTR markers described above, except DYS393 that is included in the Yfiler Plus kit only. The kits also contain various additional YSTR markers in order to improve the discriminatory capacity further. Two of these commercial kits are used in this study, the Yfiler Plus PCR Amplification kit and ForenSeq DNA Signature Prep kit. These kits contain 25 and 24 YSTR markers respectively. Table 1 shows the YSTR markers used in each kit and to what degree they overlap.

Table 1 - Overview of the YSTR markers included in Yfiler Plus and ForenSeq. Bold YSTR markers make up the minimal haplotype group, italic YSTR markers are rapidly mutating, and asterisk YSTR markers contain non-nucleotide linkers in the Yfiler Plus kit.

Unique Yfiler Plus PCR Amplification kit STR markers	Overlapping YSTR markers	
DYS393		
*DYS449**	Y-GATA-H4*	**DYS439**
DYS456	**DYS19***	DYS448*
DYS458	**DYS385a/b**	DYS460
*DYS518**	**DYS389I***	DYS481
*DYS627**	**DYS389II***	DYS533
Unique ForenSeq DNA Signature Prep kit YSTR markers	**DYS390***	*DYS570*
DYS505	**DYS391***	*DYS576*
DYS522	**DYS392***	DYS635*
DYS549	DYS437*	*DYF387S1*
DYS612	**DYS438***	
DYS643		

Lineage markers

The genomic material in s cell is divided in two, the nuclear DNA and the mitochondrial DNA. The nuclear DNA is composed of autosomal chromosomes and sex chromosomes. The Y-chromosome makes up a tiny portion of the genomic material. All these DNA types provide varying information and can be used in different ways in forensic genetics. As previously stated, standard DNA-typing is done with autosomal STR markers, because they are the most polymorphic and discriminating containing one allele from each parent as the genetic information is recombined with each offspring. Autosomal markers are termed genotypes as they contain mixed genetic information from both parents. However, the mitochondrial and Y-chromosomal markers are inherited directly from parents to offspring without recombination (3). Thus, these markers are referred to as haplotypes, rather than genotypes as they contain information from one parent and therefore consists of single alleles (in most cases).

The Y-chromosome differs from the autosomal chromosomes and X-chromosome in several ways; number of alleles in each marker, inheritance pattern, recombination and analytical power. Most YSTR markers contain one allele per marker because there is one copy of the loci on the Y-chromosome, except for DYS385a/b and DYF387S1 (among others), which are multicopy loci and exist in more than one copy on the Y-chromosome due to duplication events. As stated, autosomal markers contain one allele from each parent. The YSTR chromosome is inherited in its entirety from father to son without recombination, though mutations can occur. Figure 3 depicts the inheritance pattern of the Y-chromosome and the autosomal chromosomes. The Y-chromosome is inherited as a lineage while the autosomal chromosomes comprises of the shuffled genetic information from the individual's ancestors.

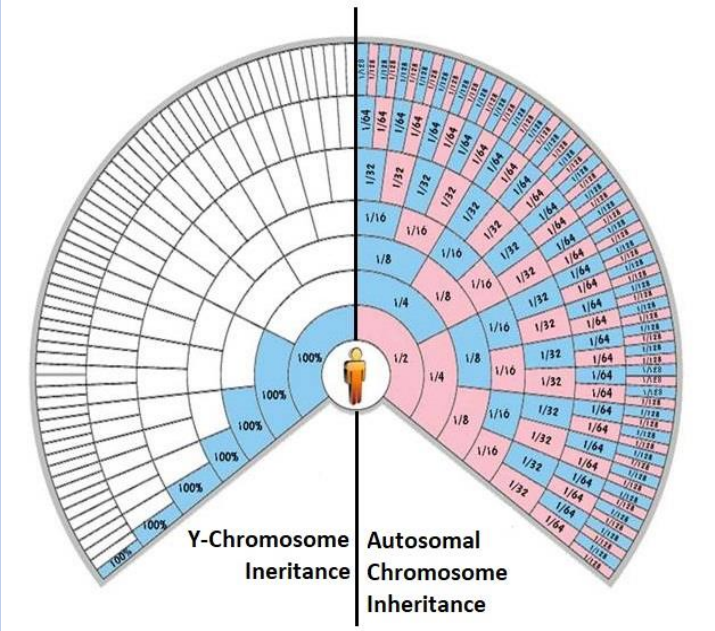


Figure 3 – Illustration of the inheritance pattern for the Y-chromosome (males) and autosomal chromosomes. Figure modified from (44)

During meiosis in males the Y-chromosome does not recombine with the X-chromosomes because these chromosomes are structurally too different and contain different genetic information (45). However, in males the Y and X chromosome do recombine slightly in the pseudoautosomal regions (PAR) at the chromosomal ends of the YSTR to ensure that the sex chromosomes link together during cell division and are correctly divided to the

daughter cells (3, 46). In females the X-chromosomes recombine with each other because they are sister chromatids, hence they are structurally identical and encode the same genetic information. Because all YSTR markers are inherently linked the haplotype will count as one allele, and not a combination of several, as for autosomal STR markers. Due to the reasons described above YSTR-analysis does not have the same analytical powers as autosomal STR-analysis because all paternally related men will have the same haplotype, unless a mutation occurs.

The mitochondrial DNA is inherited in its entirety from mother to offspring and is not actively altered, though mutations may occur. In this way, the mitochondrial DNA is similar to the Y-chromosome which is inherited from father to son without altercation because it does not recombine. Mitochondrial DNA can be utilized in forensic genetics, specifically if a sample heavily degraded (3). Mitochondrial DNA is a small, circular molecule and can exist in several hundred copies in each cell, which makes it ideal to analyze when the nuclear DNA is too degraded for a complete analysis.

Population- and reference databases

In forensic genetics there are not only several types of DNA markers that can be analyzed, there is also a variety of different sample types. Roughly DNA samples can be divided in two, biological trace samples and reference samples. Trace samples are found in and on, among others, items, clothes, and surfaces related to the crime scene and/or the people involved. These samples can for example be a blood, semen or spit stain. These samples are of unknown origin and vary greatly in quality and quantity. Reference samples are collected from known individuals, i.e. the victims, suspects, witnesses, people involved in the police work, or anyone the investigators are interested in.

When a match between the DNA-profiles of a reference sample and a trace sample is obtained calculations are needed to evaluate the chance of a random match (47). In standard autosomal STR typing a population database can be used to calculate the random match probability (RMP). A population database contains allele frequencies from individuals in a given population or ethnic group and exemplifies common and rare alleles in the given population. To calculate RMP for an autosomal STR genotype the allele frequencies for all the alleles in the genotype are multiplied. RMP can be used in standard autosomal STR-typing because the alleles recombine independently. When a match is obtained with a YSTR-profile the random match probability cannot be calculated in the same way due to the uniparental inherence pattern. In this case all alleles are inherited dependently, as a “package”. Therefore, the alleles cannot be summed, and the haplotype needs to be treated as one “allele” with only one frequency. Therefore, there is no need to collect YSTR allele frequencies in a population database. However, it is necessary to collect whole haplotypes in reference databases.

The largest reference database is the Y-chromosome haplotype reference database (YHRD) (www.YHRD.org). YHRD is a free, open access collection of YSTR markers and YSNPs from uploaded population samples. These populations are either national databases or

obtained from smaller geographical areas. The database has three objectives: "*generate reliable YSTR haplotype frequency estimates for YSTR haplotypes to be used in the quantitative assessment of matches in forensic and kinship casework, assessment of male population stratification among world-wide populations as far as reflected by YSTR and YSNP frequency distributors, and provision of advanced tools and further resources concerning YSTRs and YSNPs" (48)*. The database contains YSTR haplotypes from six commercially available analysis kits: Minimal, PowerPlex X, Yfiler, PowerPlex Y23, Yfiler Plus, and Maximal. The kits are ranged in order from most to least entries. The YHRD database provides a haplotype search function and several tools for genetic population analyses.

Length-based versus sequence-based assessment of YSTR markers

One can evaluate YSTR markers in two ways, based on the fragment length or sequence. When typing YSTR markers based on fragment length, the allele values reflect how many times the repeat unit is repeated, e.g. for allele 14 the repeat unit is repeated 14 times. The challenge with the length-based information is that sequence variants that do not alter the overall fragment length of the marker will not be detected. These allele variants are called isoalleles. Typing YSTR markers based on fragment length is the current standard practice and is done using PCR and CE. Next generation sequencing (NGS), or Massive parallel sequencing (MPS), can however evaluate the YSTR markers both based on fragment length and sequence (49, 50). Sequencing will uncover all sequence variants that do not alter the overall length of the YSTR, as opposed to fragment length-based approach. Sequence variants can be located in the repeat unit or in the flanking regions upstream or downstream of the repeat unit. Therefore, sequencing can increase the number of allele variants in YSTR markers, which in turn increases the power of discrimination. In some cases, such as samples containing several donors, this added power can be beneficial.

Polymerase chain reaction and capillary electrophoresis

Today's standard method for DNA-typing is PCR followed by CE. PCR is the process of amplifying specific sequences of DNA to hundreds of millions of copies within a few hours using a thermal cycler (51). Specific primers mark the DNA sequences of interest and these are amplified by DNA polymerases on the forward and backward strand, the remaining DNA is not amplified. The process involves precise cycles of heating and cooling in order to denature the double stranded DNA to anneal the primers and to synthesize the new DNA strand, respectively. For each cycle the DNA is doubled, and the final product is called amplicon.

Since a primer is specific for one region several primers are needed in order to amplify more than one sequence of interest at the time, termed multiplexing. For a successful multiplex all primers used need to be specific, i.e. no cross-over, and compatible, i.e. possess similar requirements for annealing temperature. The addition of each primer requires considerable planning and numerous testing in order to map each primers specificity, thermal

requirements, and possible primer interactions. If the products are relatively different in size, the size is enough to separate the fragments, but if it is desired to amplify products with the same or similar sizes fluorescent labels can be used for separation in CE. There are three ways to introduce fluorescence to PCR products, fluorescent intercalating dyes, fluorescent dNTPs, and fluorescently dye-labeled primers. The latter is most common.

Today, multiplexing of more than 20 STR markers using a 5-or 6-dye detection systems is possible. There are some additional challenges in designing primers for STR marker DNA-typing for forensic purposes. There should not be amplification of non-specific products, meaning the primer cannot bind anywhere else on the DNA, and the primers have to produce results with good peak height, and interlocus balance for loci containing more than one allele. Loci with the same fluorescent dye should also be sufficiently separated by size. If there is not enough space, separation can be obtained using mobility-modifying nonnucleotide linkers (12). These linkers allow reproducible repositioning of the alleles in order to enable interlocus spacing. Each linker unit causes a shift of approx. 2.5 nucleotides. Designing a multiplex PCR is tedious work but the availability of various commercial reagent kits has made the lab work simpler, and most labs do not create their own multiplex PCRs anymore.

After the amplification prosses the DNA fragments need to be separated in order to interpret the results (52). As previously mentioned, the standard method of separation today is CE, and is based on electric charges. The phosphate groups on the DNA molecule make the product negatively charged, and the molecule will therefore want to travel from a negatively charged to a positively charged area. The method is divided into three fully automated steps, injection, separation and detection. The major components in the CE instrument is illustrated in Figure 4. In brief summary, the DNA sample is injected from the sample tray to the inlet buffer, and because this buffer is negatively charged the DNA fragments will travel though the capillary to the positively charged outlet buffer. Before reaching the outlet buffer the DNA fragments pass a detection window where a laser excites the fluorescent dyes. The excitation is registered by a fluorescence detector.

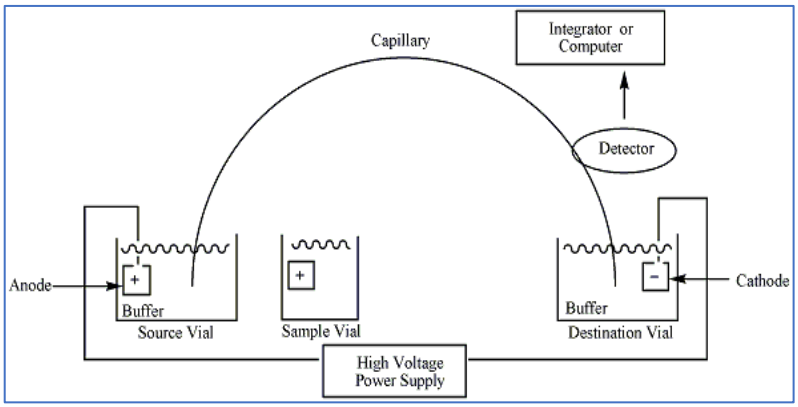


Figure 4 – The components in a capillary electrophoresis instrument. The samples are injected to the anode buffer and travels though the capillary. The DNA fragments in the samples are registered by a detector and this information is transferred to an analytical software. Figure from (53).

Before the DNA samples can be injected into the instrument, they are diluted in formamide. The dilution has two purposes, the formamide denatures the DNA and reduced the amount of salt ions found in the sample. Salt ions compete with the DNA fragments to be injected and therefore need to be diluted. In addition, a size standard is added to the samples

in order to compare the amplicons to fragments of known size. An allelic ladder is also added to one or more of the empty wells in order to compare the DNA fragments to known alleles.

The injection is executed by applying an electric impulse when the capillary is placed in the DNA sample. As mentioned above, salt ions compete with the DNA molecule for loading into the capillary. When the DNA fragments are loaded into the capillary, they start moving towards the positively charged outlet buffer. The capillary contains a viscous polymer solution that acts as a sieving mechanism. Polymer chains in the solution function as obstacles for the DNA fragments passing though the polymer. Smaller DNA fragments will easier bypass the obstacles and reach the detection window faster. Larger DNA fragments will spend longer time on bypassing and arrive the detection window later. Upon arrival in the detection window a laser excites the fluorescent dye. The excitement is captured by a fluorescence detector and is plotted as a function of relative intensity emitted from the various dyes. In order for the instrument to know what fluorescent dye it is detecting a spectral calibration needs to be done on the instrument detector and the software used for data collection. The product of the spectral calibration are matrix files that indicate the degree of overlap one can expect in the dyes used.

Lastly the data needs to be interpreted. Interpretation is done by uploading the data to an interpretive software that visualizes the results. The matrix files from the spectral calibration are used for color separation, the user-defined thresholds (set in the software) are used for peak identification, the size standard is used for peak sizing and the allelic ladder is used to call peaks to the correct allele.

Yfiler Plus PCR Amplification kit

In this thesis the Yfiler Plus PCR Amplification kit (Thermo Fisher Scientific, Waltham MA) was used for DNA-typing with PCR and CE. The multiplex kit amplifies 27 YSTR loci using a 6-dye detection system and non-nucleotide linkers (Table 1) (54). The sample sources are both extracted DNA samples and buccal samples on FTA cards. The Yfiler Plus kit was commercially available in 2014 and is a further development of AmpFLSTR Yfiler PCR Amplification kit that became commercially available in 2006. The AmpFLSTR Yfiler PCR amplification kit amplifies 17 YSTR markers using a 5-dye detection system. The Yfiler Plus kit contains the same YSTR markers as AmpFLSTR Yfiler PCR Amplification kit as well as DYS460, DYS481, DYS533, DYS576, DYS627, DYS518, DYS570, DYS449 and DYF387S1. The last six YSTR markers are rapidly mutating YSTR markers. Therefore, the discriminatory capacity of the Yfiler Plus kit is larger than the AmpFLSTR Yfiler PCR Amplification kit.

Next generation sequencing

In order to obtain sequence-based information of the DNA fragments the DNA sample can be sequenced using NGS. NGS technology enables sequencing of thousands of DNA fragments from multiple sources at once (55). There are three approaches to sequencing, whole-genome sequencing, whole-exome sequencing and targeted gene sequencing. They all

have advantages and disadvantages and are appropriate for various fields of research. For forensic purposes targeted gene sequencing is most relevant and presented further. Targeted gene sequencing is cost-effective and allows sequencing of genes of interest without unnecessary sequencing of irrelevant genes which is timesaving (56). Forensic DNA analysis can be more challenging than DNA analysis in other fields due to the variable nature of the DNA samples. They can be of low quality or quantity, and before analysis it is unknown if the sample contains DNA from more than one person (57). There is also a need for high accuracy and reproducibility.

There are several commercially available NGS platforms today, and the most common kits used for forensic purposes are MiSeq (Illumina/Verogen), Ion Torrent PGM or S5 (Thermo Fisher Scientific). These platforms utilize broadly the same workflow with library preparation, sequencing, imaging and data analysis (57). Library preparation involves also PCR amplification of the DNA sample and preparing the sample for sequencing. The preparation may vary from the NGS platform used. The amplification prior to sequencing can be done using different methods, e.g. amplicon-based method. Amplicon-based methods use specific primers that flank the targeted regions (58). Therefore, the flanking regions of the sequences of interest must be known in order to design appropriate primers (mutations here can cause bias or dropouts). Here index adaptors also can be used in order to separate DNA samples, allowing multiplexing. Amplicon-based enrichment is ideal for multiplexing and allows sequencing of multiple individuals at the same time, unlike CE.

After library preparation the libraries are amplified in order to obtain a measurable amount for sequencing. Illumina/Verogen uses bridge amplification on a flow cell where the amplicons are isothermally amplified to create clusters. The amplicons are evenly distributed on the flow cell and attach to one of the two oligos bound to the surface that are complementary to both ends of the amplicons. The bridge amplification and cluster generation happen as depicted in Figure 5. Upon hybridization between the oligo and the DNA fragment, a polymerase is recruited and synthesizes a new DNA strand from the oligo using the DNA fragment as a template. The now double stranded DNA is denatured, and the template is washed away. The synthesized strand, attached to the flow cell, now hybridizes with the other oligo on the flow cell, and a new strand is synthesized as a bridge. The bridge denatures and the double stranded DNA separates into two separate strands. Now both these will hybridize to the opposite oligo on the flow cell and will be amplified as a bridge as described above. After the amplification is completed all

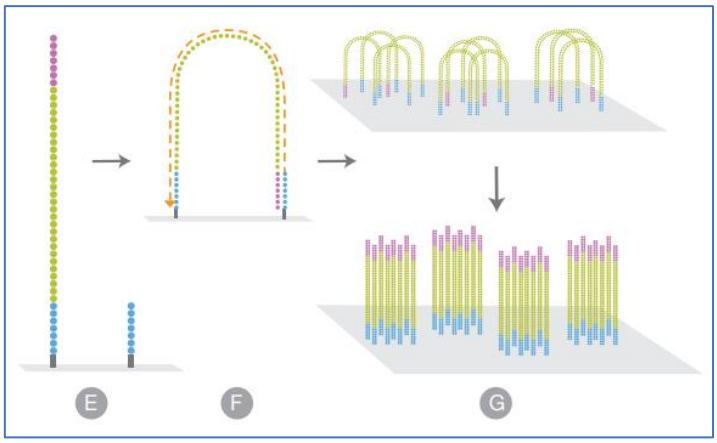


Figure 5 – Bridge amplification. E) DNA attached to flow cell. F) Perform bridge amplification. G) Generate clusters. Figure modified from (59).

reverse strands are removed and the 3' ends of the remaining DNA fragments are blocked to prevent unwanted priming (60). Following bridge amplification, each amplicon will have created a cluster, and these clusters are parallelly sequenced (61).

After the bridge amplification and cluster generation is completed the sequencing is initiated. Illumina/Verogen utilize sequencing by synthesis. All nucleotides are added in each cycle, but they contain fluorescens that allows nucleotide separation (60). The sequencing begins with elongation of the sequencing primer to produce the first read. For each cycle dNTPs tagged with fluorescens will compete for binding to the DNA template. Adenine and thymine, and guanin and cytosine are complementary and bind to each other. When bound the dNTPs are excited and emit a fluorescence signal that is registered by the instrument. The number of cycles determines the length of the fragment synthesized. Synthesis occurs all over the flow cell at the same time, it is therefore important that the clusters do not overlap. Several reads are done on both forward and reverse strands and the produced fragments in all cycles are registered by the instrument.

ForenSeq™ Signature Prep kit

Assessment by sequencing in this thesis done is with the ForenSeq DNA Signature Prep kit on the MiSeq FGx Forensic Genomics System (Illumina/Verogen), which is especially developed for use in forensic genetics (Figure 6). The ForenSeq DNA Signature Prep workflow consists of four steps; library preparation, cluster generation, sequencing and data analysis. The kit includes primers for 24 YSTR markers, in addition to several STR markers and SNPs (62). Two primer mixes are provided in the kit, DNA Primer Mix A (DPMA) and DNA Primer Mix B (DPMB). The latter is used in this study. Both primer mixes contain the 24YSTR, 7 XSTR and 27 autosomal STR markers, and 94 identity SNPs, but DPMB also contains 56 ancestry-informative and 24 phenotypic-informative SNPs. For this study bother primer mixes could have been used but the typing results for the ancestry and phenotype informative SNP information from these samples will be used in different projects carried out by the research group. The kit uses amplicon-based enrichment with two rounds of PCR tagging each amplicon with an index and adaptor in order to analyze multiple DNA samples at once. The cluster generation is done by bridge amplification and the sequencing is done by synthesis.

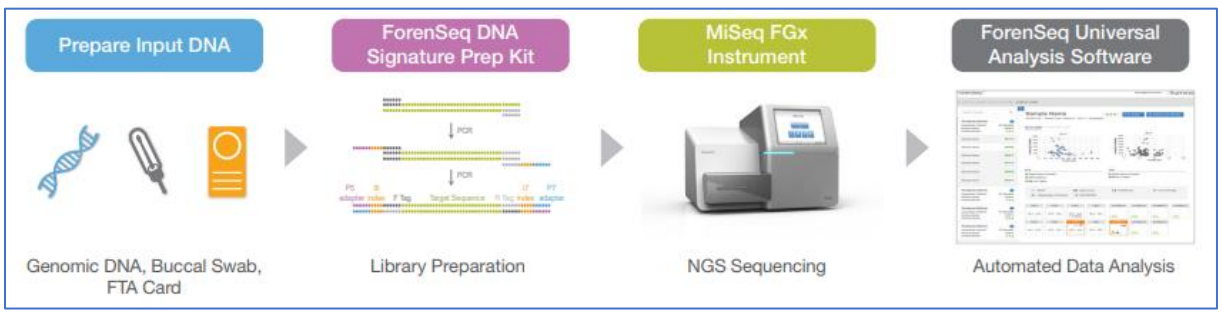


Figure 6 – MiSeq FGx Forensic Genomics System workflow. Before sequencing the input DNA is prepared and libraries are created using the ForenSeq DNA Signature Prep kit. The libraries are injected to the MiSeq FGx Instrument where the sequencing occurs. The final results are uploaded and visualized the ForenSeq Universal Analysis Software (UAS). Figure from (62)

Aims of this study

The Center for forensic genetics (CFG) has implemented methods for standard autosomal STR-typing and wishes to extend the analysis repertoire with YSTR analyses. The aims of this study were as follows:

- Increase the sample size of male contributors with Norwegian ancestry in the research biobank at CFG (the Norwegian population sample).

- Establish analysis protocols for two analysis kits based on different technologies, the capillary electrophoresis-based Yfiler Plus PCR Amplification kit (Thermo Fisher Scientific) and the sequencing-based ForenSeq DNA Signature Prep kit (Illumina/Verogen) for different types of reference samples

- Compare length-based DNA-typing results from both kits to evaluate the concordance in the overlapping YSTR markers.

- Assess sequence variation of YSTR markers in both the repeat unit and the flanking regions (ForenSeq)

- Establish haplotype frequencies in the Norwegian population data set that can be used in reference databases (both kits)

- Compare population genetic parameters for the Norwegian population with other European populations

Materials and methods

Sample collection

126 of the DNA samples used in this study were available in the biobank. These samples were collected in with an informed, written consent in Tromsø and Bodø (63, 64). To increase the sample size more samples were collected to the biobank. The sample set in this study came to a total of 301 individuals. However, for future publication 400 is needed (65). In this study samples were collected from willing donors at the University of Tromsø (UiT) and the collectors' workplace in Tromsø. Totally 217 samples were collected, and of these 175 are third generation Norwegian on their fathers' father side. Each sample consisted of buccal cells that were collected with a sterile foamed tipped applicator (Whatman®, Chicago IL) and deposited onto an FTA card (Whatman® FTA card, Chicago IL/COPAN Nucleic Card™, Brescia Italy). Each FTA card was stored in an evidence bag. A consent sheet and questionnaire were signed and filled out by each donor at time of collection. The questionnaire asked about the donors' birth year and place, and the grandparents' birthplace. It also included other information in line with previous sample collections, e.g. about skin, hair and eye color. However, this information is not relevant for this study but will be used in forthcoming projects.

In total, the sample set included the 175 samples collected in this study and 126 previously collected samples (63, 64). All DNA samples used in this study belong to individuals that have a Norwegian grandfather on their fathers' side. All samples are gathered in a research biobank at the Centre of Forensic Genetics (CFG). The previous collection was done by obtaining blood samples. The donors signed the same informed consent sheet and questionnaire as the ones used in this study. The blood samples were previously purified using the QIAamp DNA Investigator kits (Qiagen, Hilden Germany) and quantified using the Quantifiler™ Trio DNA Quantification kit (Thermo Fisher Scientific, Waltham MA) on the 7500 Real-time PCR system (Thermo Fisher Scientific, Waltham MA) (63, 64, 66). Aliquots of 0.2 ng/µl DNA were available and used directly in the analysis, as described later.

Yfiler Plus PCR Amplification kit

An establishment of the method used according to the manufacturers protocol was needed as this method is new for the laboratory. The manufacturers protocol recommends 1 ng DNA in 25µl reaction volume and testing 26-29 cycles of amplification (54). The laboratories experience using other kits from the same supplier was that 0.5 ng DNA and 29 cycles gave reliable results. Therefore, both 1 and 0.5 ng DNA, using DNA control 007, were amplified using 29 cycles. For this kit 1 ng DNA gave greater results than 0.5 ng DNA. In order to save resources five DNA samples were amplified using half the reagent volume. In this case the amount of DNA used in the samples and controls was 0.5 ng. Amplification was also tested using 29 cycles. The DNA samples used for this testing were previously extracted DNA samples (0.2ng/µl). The samples were successfully amplified and analyzed using these conditions, so

the remaining extracted DNA samples were successfully amplified using these conditions. For the extracted DNA samples, the reaction mix was prepared according to Table 2 and evenly distributed in the PCR tubes. DNA was added subsequently to the tubes before transferring them to the Verity 96 well Thermal Cycler (Thermo Fisher Scientific, Waltham MA) and starting the PCR according to Table 3.

For the FTA cards (n=175) additional TE buffer was added to the reaction mix in order to provide enough liquid for the sample. These samples were also run using half the reagent volume. A 1.2 mm punch of each FTA card was used per reaction. Because the DNA quantity deposited on FTA cards is very variable, the first PCR-setups were varied out to find the optimal cycle number achieving good analysis results for as many samples as possible. Three setups were analyzed with 26, 27 and 28 cycles, respectively, each with 1.2 mm punches of the five samples. For these initial five samples 27 cycles showed good allele peak alleles in the electropherograms. To confirm that 27 cycles was optimal, the following analysis was carried out with 20 samples (and two controls). The electropherograms showed low allele peal heights and allele dropouts in several YSTR markers for several individuals. Therefore, the remaining runs were carried out using 28 cycles. The increased cycle number gave higher peaks and little to no allele dropouts. In total 93.1% (163/175) of the FTA cards were successfully analyzed using 28 cycles. For the FTA cards the reaction mix was also prepared, according to Table 2, and added to the tubes before the 1.2 mm punches were added. Without adding the liquid first, the punches would sometimes escape from the tubes due to static forces. Although the manufacturer's protocol states that the punches should be added before the reaction mix it is also suggested there to reverse the order if one experiences issues with static forces (p.22). The PCR tubes were then transferred to the Verity 96 well Thermal Cycler and the PCR was started according to Table 3.

The DNA concentration on twelve of the FTA cards were either too low, having low allele peak heights and incomplete DNA-profiles, or too high, having a high number of artefacts and split peaks in the electropherogram that make in difficult to confidently type the alleles. These samples were analyzed using 29 or 27 PCR cycles, respectively. Six of the DNA samples with a low DNA concentration were also analyzed with two 1.2 mm punches in order to increase the DNA input, but for all except one sample none of the YSTR markers were amplified. Probably due to saturation.

I total, adjusting the cycle number resulted in full DNA-profiles for seven samples. For the remaining five samples with insufficient typing result, a small piece of these FTA cards (approx. 25 mm^2) was cut out using sterile scalpels and extracted using the PrepFiler Express extraction kit (Thermo Fisher Scientific, Waltham MA) on the Automate Express DNA Nucleic Acid Extraction System (Thermo Fisher Scientific, Waltham MA). The isolation technology of the kit is based on magnetic beads. DNA is bound to magnetic beads upon cell breakage and sticks there during several washing steps before the DNA is released again in an elution buffer. The extracted DNA was quantitated using the Quantification Trio DNA Quantification kit (Thermo Fisher, Waltham MA) on the 7500 Real Time PCR System (Thermo Fisher Scientific,

Waltham MA) with the HID Real-Time PCR Analysis Software v1.2 (Thermo Fisher Scientific, Waltham MA). The quantification uses real-time PCR, monitoring the amplification as it happens using sequence-specific DNA probes. These probes contain two fluorescent tags, a reporter and quencher, that emit light at different wave lengths. When the probe is intact, and the tags are within close proximity to each other, no fluorescence will be detected. However, when the probe hybridizes to the double stranded DNA, the reporter detaches, and light will emit because the tags are no longer in close proximity. Both the DNA extraction and quantification were done according to manufacturers' protocols (67, 68). The DNA extracts from FTA cards were normalized to a concentration of 0.2 ng/µl and amplified with the Yfiler Plus kit as described for extracted DNA samples previously.

Table 2 – Yfiler Plus reaction mix.

Extracted DNA samples	*µl per reaction*
Master Mix	5
Primer Set	2.5
Tris-EDTA-buffer	2.5
DNA (0.2 ng/µl)	2.5

FTA Cards	*µl per reaction*
Master Mix	5
Primer Set	2.5
Tris-EDTA-buffer	5
FTA punch	1.2 mm

Table 3 – Yfiler Plus PCR conditions on the Veriti Thermal Cycler in 9600 emulation mode.

Initial incubation	Optimal cycle number		Final extension	Final hold
	Denature	Anneal/Extend		
Hold	26-29 cycles*		Hold	Hold
95°C 1 minute	94°C 4 seconds	61.5°C 1 minute	60°C 22 minutes	4°C Up to 24 hours

Before the DNA fragments in the samples could be separated the CE needed to be spectrally calibrated using the DS-36 Matrix Standard kit (J6 Dye Set) (Thermo Fisher Scientific, Waltham MA). After the PCR, the samples were prepared for CE on the 3500XL Genetic Analyzer (Thermo Fisher Scientific, Waltham MA). A mixture of formamide and internal size standard was added to each well on a 96-well plate. Lastly, 1 µl PCR product, independent if it was from extracted DNA or FTA cards, or allelic ladder was added to each well. The CE was done using the 3500xL instrument with 3500 Data Collection Software v3.1 and POP-4 polymer (Thermo Fisher Scientific, Waltham MA) according to the manufacturer's recommendations and settings for this kit.

Upon completing CE, the raw data was transferred into GeneMapper™ ID-X v1.4 (Thermo Fisher Scientific, Waltham MA). Here all data, presented as electropherograms, was reviewed, and some data was manually conferred and assessed. Conferring involves checking the peak heights to evaluate if any allele dropouts have occurred and evaluate the background noise of the samples. Any PCR- or CE-related artefacts such as stutters, pull-up or spikes were removed and characterized as such. Some artefacts also occur if the sample contains too much DNA, i.e. split peaks and/or shoulders for allele peaks. All allele variants not represented in

the allelic ladder, potential null-alleles and duplicated alleles were re-amplified and analyzed to confirm the results. Lastly, the conferred DNA-profiles for all samples were exported from GeneMapper™ ID-X as an Excel spreadsheet.

ForenSeq DNA Signature Prep kit and MiSeq FGx Reagent kit

The ForenSeq DNA Signature Prep kit is used for library preparation. These libraries are pooled and denatured and sequenced using the MiSeq FGx Reagent kit. The library preparation consists of sample amplification, sample tagging using specific indexes, bridge-amplification, magnetic bead-based purification and normalization, before the DNA samples are pooled and denatured (62). YSTR-haplotype data was already available for the extracted DNA samples (n=126) from previous analysis, but YSTR-analysis results had not been evaluated yet (63, 64, 66). At a closer check, nine of these samples had incomplete YSTR-results and were therefore re-sequenced.

A workflow for ForenSeq using extracted DNA samples was already established in the laboratory as it has been used in previous projects (63, 64, 66). The workflow for FTA cards is almost identical, apart from a washing step in the beginning of the procedure. From the amplification and though to the sequencing the procedure is the same as the previously established workflow. The FTA-specific procedure involves a 1xTBE (Tris-borate-EDTA buffer) washing step of the FTA punch before PCR1. The washing step was executed according to the manufacturer's protocol apart from one step. The protocol states that the 1.2 mm FTA punch is to be added before the reaction mix. In this case the reaction mix was added before the 1.2 mm FTA punch. The protocol stated that they should be added in reverse order, but the manufacturer does state this as a solution if one experiences static forces.

A reaction mix (Table 4) for PCR1 was created according to the manufacturers' protocol. The reaction mix was distributed to each well of the 96-reaction plate. Then the nine previously extracted DNA samples (with previous insufficient results) and controls were transferred to the wells. The plate was sealed with adhesive film before placed in the Veriti 96 well thermal cycler for PCR1 (Table 5). The remaining preparations were identical for both extracted DNA samples and FTA cards.

Table 4 – Reaction mix for PCR1 using the ForenSeq amplification kit.

FTA cards	*µl per reaction*
PCR1	4.7
FEM	0.3
DMPB	5.0
dH2O	5.0

Extracted DNA samples	*µl per reaction*
PCR1	4.7
FEM	0.3
DMPB	5.0

Table 5 – PCR1 setting on the Veriti Thermal Cycler. The lid temperature was 100°C.

Initial incubation	PCR1						Final extension	Final hold
Hold	8 cycles				10 cycles		Hold	
98°C 3 min	96°C 45 sec	80°C 30 sec	54°C 2min	98°C 45 sec	96°C 30 min	68°C 3 min	68°C 10 min	10°C

After PCR1, the index 1 (i7), index 2 (i5) and PCR2 reagents were added according to the manufacturers' protocol, and the samples were placed back into the Veriti 96 well thermal cycler for PCR2 (Table 6). The index adaptor combinations are unique for each sample, allowing complete separation of the DNA samples. For subsequent analyses different index-adaptor combinations were used, rotating between four of the twelve available index 1 types. Changing of the index adaptors were done so the samples from two amplifications could be sequenced together, if some of the samples failed to amplify properly, because they did not have overlapping index adaptor combinations.

Table 6 – PCR 2 setting on the Veriti Thermal Cycler. The lid temperature was 100°C

Initial incubation	PCR2			Final extension	Final hold
Hold	15 cycles			Hold	
98°C 30 sec	98°C 20 sec	66°C 30 sec	68°C 90 sec	68°C 10 min	10°C

Library purification

The amplified DNA products are purified by removing other reaction components using magnetic Sample Purification Beads (SPB). Upon completion of PCR2 a 0.8 ml 96 well storage plate was prepared with well-vortexed SPB, and the PCR products were transferred to the plate according to the manufacturer's protocol. The plate was then placed on a magnetic stand and washed twice with 80% ethanol. The DNA binds to the beads that remain in the bottom of the well due to magnetic forces, allowing the supernatant to be removed. Resuspension buffer (RSB) was then added to the wells to release the DNA from the beads. The plate was placed on the magnetic stand and the supernatant, now containing the amplified DNA fragments, was transferred to a new 96-well plate, to be normalized.

Measuring the DNA concentration of the purified libraries

This quality check measuring the DNA concentration of each purified library was added by the research group during troubleshooting in the early stages of implementing this method and is between purification and normalization of the libraries. Measurement of the sample's DNA concentration was done to ensure that the samples contained enough DNA to be sequenced and to adjust the volume pooled library that is used in the sequencing. For this purpose, the dsDNA HS assay kit (Thermo Fisher Scientific, Waltham MA) and the Qubit® 4.0

Fluorometer (Thermo Fisher Scientific, Waltham MA) were used, following the manufacturer's protocol (69).

The Qubit assay tubes (Thermo Fisher Scientific, Waltham MA) were marked according to the 96-well reaction plate layout, two tubes for the Qubit standards were also needed. The standards were used to create a standard curve that each test sample was compared to in order to find the samples DNA concentration. A Qubit® working solution was prepared by mixing HS buffer and reagents, according to Table 7, and transferred to corresponding tubes. Subsequently, DNA from the purified libraries and two standards were transferred to their corresponding assay tubes. The tubes were vortexed and centrifuged before the Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham MA) was used to measure the samples. First, the two standards were measured in order to create the standard curve, then the test samples were measured one by one.

Table 7 – Reagents in the Qubit working solution.

dsDNA HS assay kit reagents	*Volume needed for 32 samples*
Qubit ds DNA HS Buffer	7 164 ml
Qubit dsDNA HS Reagents	36 µl

Normalization and pooling of libraries

The purifies libraries were normalized using Library Normalization Beads (LNB1) according to the manufacturers' protocol. Normalization assures that the DNA samples within each sequence run are equally represented and that each DNA sample provides consistent cluster densities. First a reaction mix containing LNB1 and Library Normalization Additives 1 (LNA1) was prepared and transferred to the wells of a 96-well plate. DNA was then transferred from the purified library plate to the plate containing the LNB1-LNA1 master mix. The plate was then centrifuged and placed back on the magnetic stand, then the samples were washed with Library Normalization Wash 1 (LNW1). Then, the samples were treated with HP3 (NaOH) to denature the double-stranded DNA and release the molecules from the magnetic beads. Finally, the normalized libraries were transferred to a 96 well plate containing Library Normalization Buffer 2 (LNS2) to neutralize the HP3.

After library normalization the libraries were thoroughly mixed and pooled into one library. The Human Sequencing Control (HSC) was then added to the samples, in addition to Hybridization Buffer (HT1, from the MiSeq FGx Reagent kit). HSC, HT1 and a sample of the pooled library was mixed, according to Table 8, and immediately placed on a heating block (96°C) before being placed on ice. The heat and HP3 ensures denaturation and placing it on ice stops the process. The protocol states that 7 µl of the pooled library is to be used in the final dilution. Previously, the research group found that 14 µl was needed in order to successfully genotype the STR markers and achieve SNP typing. This was based on 1 ng DNA extracts. The amount of the pooled normalized library was based on the measurements done with the Qubit® 4.0 Fluorometer. Six sequencing setups were performed in this study and the amount of pooled normalized library added to the final dilution was 12, 9, and 14 (x4) µl.

Table 8 – Final dilution of the pooled and normalized libraries. If less or more pooled library was used the amount of HT1 was altered so the total amount would be 600 µl.

Dilution of libraries	*µl needed*
HT1	589-584
Pooled normalized libraries	9-14
HSC	2

Loading the MiSeq FGx Instrument and initiation of sequencing

The reagent cartridge from the MiSeq FGx Reagent kit was prepared according to the manufacturer's protocol and loaded with the denatured library in position 17 (70). The flow cell was thoroughly washed with dH_2O and dried carefully with lens paper before loaded into the machine. Thereafter, the sequencing by synthesis (SBS) solution and waste bottle were loaded into the machine according to the manufacturer's protocol. The sample information, i.e. the individual sample name and their specific index adaptors combinations, was uploaded to the ForenSeq server. The sequencing was started on the MiSeq. Sequencing runs are monitored and four quality scores, cluster density, cluster passing filter, pre-phasing, and phasing, are obtained.

Data analysis using ForenSeq UAS

The sequencing data was, upon completion, automatically loaded onto the ForenSeq server and analyzed with the ForenSeq Universal Analysis Software (UAS) v1.0.15119 (Illumina/Verogen) using the flanking region setting. The data was quality checked in UAS and manually conferred if necessary. Conferring involves, among other things, evaluating stutters, duplicated alleles and null-alleles. Both the haplotype and sequence information were exported from UAS as Excel spreadsheets.

Statistical calculation of allele and haplotype data

If no other software is mentioned Microsoft Excel was used for analysis of the data. Allele variants and frequencies, obtained from both kits used in this study, were manually counted, estimated and compared using the counting method. The concordance between the haplotypes produced from the two kits was also determined. DYS385a-b and DYF387S1 are multicopy loci and were therefore treated as genotypes. The haplotypes from both kits were counted and compared. Marker genetic diversity (GD) and haplotype diversity (HD) were calculated according to Nei (71). The match probability (MP) and discriminatory capacity (DC) for the kits used were also calculated according to Olofsson et.al. (72).

In addition, YSTR markers of the Yfiler Plus haplotypes were removed to match the YSTR markers included in the kit's predecessor, the Yfiler kit. DYS576, DYS627, DYS460, DYS518, DYS570, DYS449, DYS533, DYS481 and DYF387S1 were removed.

Relative genetic distance and creating of a multidimensional scaling plot

The Y-Chromosome STR Haplotype Reference Database (YHRD) was used to perform an analysis if molecular variance (AMOVA) and create a multidimensional scaling (MDS) plot with the Yfiler Plus haplotypes obtained from this study and other European populations with 150 or more haplotypes reported to YHRD. Pairwise genetic distances between populations (R_{ST}) and corresponding p-values (using 10.000 permutations) were calculated. Standard settings (not relaxed) were used for the MDS. In total, twelve samples were removed from the sample set because the AMOVA and MDS tool does not allow haplotypes containing null-alleles, allele variants, or duplicated alleles. To evaluate the p-values the Bonferroni correction was used:

$$p\ value = \frac{0.05}{\left(1 + \frac{n}{2}\right) * n}$$

If a different significance value than 0.05 is desired simply change the 0.05 in the equation to another number. n is the number of populations used in the analysis, in this case 14.

The Yfiler and Yfiler Plus haplotypes (n=271 and n=290, respectively) in this study were also searched in YHRD to see if there was a match with previously reported haplotypes. There are three calculations that estimate how often one can expect matches, discrete Laplace (DL, only for Yfiler YSTR markers), augmented counting (n+1/N+1), and Kappa. The DL calculation estimates haplotype frequencies within a metapopulation while taking allelic distribution into consideration (73). DL is only calculated for Minimal and Yfiler haplotypes. It also excludes DYS385a/b. The Kappa calculation estimated haplotype frequencies using the frequency of singletons within a population sample (74). The augmented counting (n+1/N+1) calculation is the frequency is obtained by adding the haplotype in question to the database and observations.

Haplogroup predicting

The software Whit Athey's Haplogroup Predictor (http://hprg.com/hapest5/index.html) was used to predict to which haplogroup each haplotype in the Yfiler Plus dataset belongs to. The software has three options: 21-haplogroups, 27-haplogroups, and main 111-markers. The 27-haplogroup was considered non-relevant because it contains some Asian markers. The 21-haplogroups and main 111-markers both contain most of the YSTR markers used by Yfiler Plus kit but predicted different haplogroups. Both prediction models were used to see if there was a difference in haplogroup prediction. The missing YSTR markers are DYS627, DYS518, and DYS387S1. The program includes DYS385a/b, but as separate alleles. Data obtained from the Yfiler Plus analysis does not separate between allele a and b. Because the length of these alleles overlap to some degree it is not possible to distinguish which of the alleles in the genotype belongs to DYS385a and which allele belongs to DYS385b.

The software has a batch program which allows application of a large number of haplotypes. When running the haplotypes though the program a minimum score and minimum probability need to be selected. The suggestion from the software is 40 and 95%, respectively, and used in this study. The program also asks from what European region the haplotypes come from. The program lets you choose between northwest Europe, east Europe, Mediterranean, and equal priors. Two runs were performed, one with northwestern Europe and one with equal priors to see if the haplogroup prediction was differed. The software accepts point mutations and null-alleles, but it does not accept duplicated alleles. Therefore, one individual from the sample size was excluded due to having an allele duplication in DYS635, bringing the sample set to 300.

Principal component analysis

STRAF v1.0.5 was used to create a principal component analysis (PCA) plot. For this analysis the Yfiler Plus haplotype data was used, excluding the multicopy loci DYS385a-b and DYF387S1. The software allows the inclusion of null-alleles and microvariants, but not duplicated alleles. Therefore, one individual was excluded from the sample set due to a duplication in DYS635, bringing the sample set to 300. The image presented in the results is comprised of two PCA axis because this was the software's standard setting.

Results

Establishment of methods

Yfiler Plus PCR Amplification kit

Upon testing the PCR conditions for the Yfiler Plus PCR Amplification kit it was proven that both the extracted DNA samples and the samples deposited on FTA cards could be analyzed using half PCR reagent volume. The ideal cycle number varied for the sample types. The ideal cycle number for the extracted DNA samples was 29 and the ideal cycle number for the FTA cards was 28. All extracted DNA samples (n=126) were successfully analyzed using 29 cycles, meaning they had adequate peak heights, no dropouts and little or no artefacts. Most FTA cards, 166 of 175 (94.9%), were successfully analyzed using 28 cycles. Nine samples were re-amplified using a different cycle number or they were extracted, as described later. For the remaining nine samples full DNA-profiles could be obtained by either amplifying a new punch using a different cycle number or by amplifying extracted DNA obtained from a piece of the FTA card.

In total, 21 amplification setups were done using the Yfiler Plus kit, each with a positive and negative control. Apart from the first analysis that had one negative control and four positive controls in two different concentrations. All positive controls were successfully haplotyped, meaning the DNA-profiles displayed the expected haplotype. All but one of the negative controls were successful and showed no alleles. The control had a contamination in as depicted in Figure 7. As the figure illustrates the typed allele had a relatively high RFU value. The remaining YSTR markers in this control showed no signs of contamination.

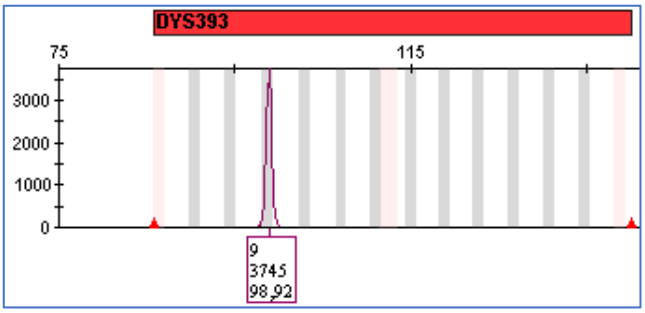


Figure 7 - Contamination in one YSTR marker in one of the negative controls. The contamination is an allele (9) in DYS393.

ForenSeq DNA Signature Prep kit and MiSeq Reagent kit

The sequencing data used in this study has partly been obtained during previous projects and partly in this study. Seven amplification setups and six sequencing setups were performed in this study using the ForenSeq DNA Signature prep kit. Quality scores for all six sequencing setups were within the manufacturers recommendation and are listed in Table 9. The maximal and minimal sample intensities are also listed in the table.

Table 9 – Quality scores for the six sequencing setups performed in this study using the ForenSeq kit. All scores are within the manufacturer's recommendations.

Quality scores	*1*	*2*	*3*	*4*	*5*	*6*
Cluster density (K/mm2)	1323	1118	1362	1575	1421	1389
Cluster passing filter (%)	93,44	95,95	93,52	92,09	93,52	93,34

Pre-phasing (%)	0,142	0,151	0,141	0,141	0,144	0,138
Phasing (%)	0,058	0,059	0,094	0,073	0,049	0,043
Max. # reads per sample	610 956	443 341	591 513	810 182	766 166	435 136
Min. # reads per sample	89 170	148 724	219 454	304 114	233 393	150 933

Of the seven positive controls, only three had complete typing results for the YSTR markers, as illustrated in Table 10. The fist sequencing runs contained two positive and negative controls as it contained both extracted DNA samples and FTA cards. The number of reads of the controls is also included in the table. The signals are either not detected, below the analytical threshold, or between the analytical and interpretational threshold. Even though the positive controls had multiple dropouts this was not the case for the sequenced DNA samples. Therefore, the samples were used in this study and not re-sequenced.

Table 10 – Positive controls for the seven amplifications in this study using the ForenSeq amplification kit. Total number of dropouts include dropouts in all the included STR markers and SNPs.

Positive control	*YSTR dropout*	*Intensity*	*Total # dropouts*
1	1	275 349	1
2	-	245 702	10
3	6	193 485	14
4	13	159 811	38
5	4	318 488	8
6	-	306 459	-
7	-	254 568	-

Unique allele variants

The results obtained from the analysis of 301 samples using the Yfiler Plus kit contained a few unique findings. Usually, one allele is expected per YSTR marker except for two alleles in DYF387S1 and DYS385a/b. Two samples had null-alleles (DYF387S1, DYS438), one individual had two alleles in DYS635, four individuals had three alleles in DYF387S1, and five individuals had microvariants that were not represented in the allelic ladder (DYS458, DYS449, DYS627, DYS385a/b). There were also unique findings obtained using the ForenSeq. Of the 288 samples that were sequenced results of 286 were used in this study. The two samples that were excluded had either dropouts in the YSTR markers or were a mixture and therefore not possible to type with confidence. Of the remaining 286 samples, three were di-allelic in YSTR markers that are expected to have one allele (DYS612, DYS392, DYS635), one sample contained a null-allele in DYF387S1, one contained a microvariant in DYS385a/b and five were tri-allelic in DYF387S1. These unique findings are listed in Table 11. Note there are two inconsistent findings between the two kits, DYS392 and the last DYF387S1. DYS392 is typed 17 using Yfiler Plus and 11,17 using ForenSeq. One of the DYF387S1 markers is typed 36,37 using Yfiler Plus and 36,37,37 by ForenSeq.

Table 11 – Unique findings obtained using the Yfiler Plus and ForenSeq kit.

YSTR Marker	*Finding*	*Haplo-/Genotype*	*Yfiler Plus*	*ForenSeq*
DYS392	Di-allelic	17/11,17	X	X
DYS635	Di-allelic	23,24	X	X
DYS438	Null-allele	0	X	NS
DYS385a/b	Microvariant	13,14.2	X	X
DYF387S1	Null-allele	0	X	X
	Tri-allelic	36,39,40	X	X
		35,37,38	X	X
		34,35,36	X	X
		34,35,36	X	X
		36,37/36,37,37	X	X
DYS458	Microvariant	17.2	X	MNI
DYS627	Microvariant	19.1	X	MNI
		21.2	X	MNI
DYS449	Microvariant	30.2	X	MNI
DYS612	Di-allelic	30,31	MNI	X

NS – Not sequenced

MNI – Marker not included

Concordance and allele frequencies

Analysis using both Yfiler Plus and ForenSeq provide length-based information for the YSTR markers in the respective kits. The concordance between the two kits could be evaluated for the 19 overlapping YSTR markers. Due to limited time and a shipping delay from the manufacturer, not all samples were sequenced. Therefore, the sample set obtained with the ForenSeq kit contains fewer results than the sample set obtained for the Yfiler Plus kit (n=286 and n=301, respectively). Therefore, the evaluation of concordance is based on analysis results from 286 individuals. All allele calls are concordant between the two kits expect for DYS392 in one of the samples. Making the concordance is 99.65% between the Yfiler Plus and the ForenSeq kits. Thus, the concordance between the Yfiler Plus and the ForenSeq kit is 99.65%. The single sample with discordant typing results, DYS392 was typed as allele 17 with Yfiler Plus and as alleles 11 and 17 with the ForenSeq kit. The intensity of allele 11 was 43% higher than for allele 17.

Figure 8 shows the length-based allele frequencies of the YSTR markers analyzed with both the Yfiler Plus and ForenSeq kit for overlapping markers (A), and markers that are unique for each kit (B and C, respectively). The allele frequencies for overlapping YSTR markers are slightly different due to the difference in sample sizes (Figure 8A). As the figures reveal there is a large difference in allele variance. The YSTR markers with the lowest allele variance are DYS393, DYS392, DYS391, DYS448 and DYS460. Some of the YSTR markers with the highest allele variance are DYS481, DYS627, DYS449, and DYS518. Of the two multi-copy loci, DYS385a/b and DYF387S1, the latter has the most allele variance.

Figure 8 – Length-based allele frequencies of the YSTR markers included in the Yfiler Plus kit and the ForenSeq kit. The blue bars represent allele frequencies obtained by analysis using the Yfiler Plus kit, and the orange bars represent those obtained using the ForenSeq kit. A) Shows the allele frequencies of YSTR markers that exists in both kits. B) Shows the allele frequencies of YSTR markers unique for the Yfiler Plus kit. C) Shows the allele frequencies of YSTR markers unique for the ForenSeq kit. YSTR markers with an asterisk are rapidly mutating.

The allele frequencies represented in Figure 8 are used to calculate the marker genetic diversities (GD) presented in Figure 9. As Figure 9 shows, the YSTR markers with the lowest degree of allele variation also have the lowest GD values, and the markers with the highest degree of allele variation have higher GD values. There lowest genetic diversity is observed in DYS293. All other markers have GD-values above 0.5. The highest genetic diversity, 0.901, is observed for DYF387S1.

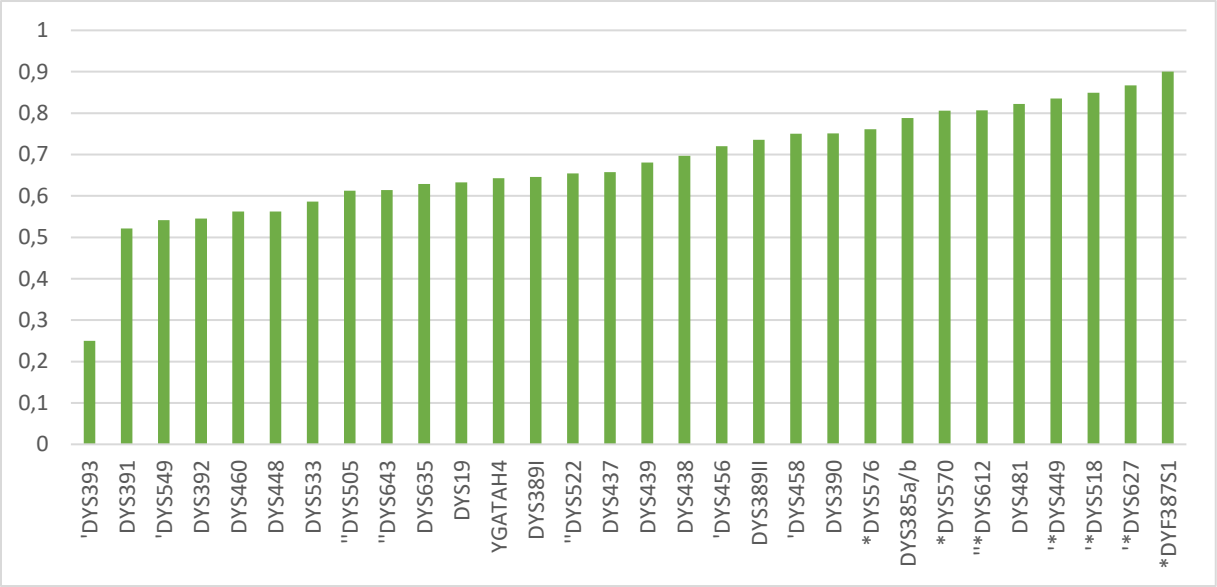


Figure 9 – Marker genetic diversities (GD) for all YSTR markers used in both kits. For overlapping YSTR markers the average GD value is used. The asterisk YSTR markers are rapidly mutating. YSTR markers with one apostrophe are unique for the Yfiler Plus kit, and the YSTR markers with two apostrophes are unique for the ForenSeq kit.

Sequence-based allele variants

A total of 286 samples were successfully sequenced using the ForenSeq amplification kit. In addition to length-based allele typing the ForenSeq kit also provides sequence information of the repeat unit and the flanking regions. Out of the 24 YSTR markers included in the ForenSeq kit, 15 markers have extra allele variants based on sequence information (Table 12 and Figure 10). Table 12 shows a summary of the amount of sequence variations for each of the YSTR markers in the ForenSeq kit. The single copy locus with the most sequence variants was DYS389II, with 14 new allele variants, and the multicopy locus with the most sequence variants was DYF387S1, with 22 new allele variants. Figure 10 illustrates how many allele variants are obtained evaluating length-based and sequence-based information in both the repeat units and the flanking regions. The bar graph is divided in four groups based on type of allele variant, and each group is sorted from YSTR markers with the largest increase of sequence-based allele variants to YSTR markers with the lowest increase.

It is worth noting that four of the YSTR markers have sequence information for one of the flanking regions only due to the positioning of the primer (Verogen, personal communication). DYS389II is missing the flanking region downstream of the repeat unit, while DYS570, DYS19 and DYS392 are missing the flanking region upstream of the repeat unit. All but two YSTR markers have equally long flanking region sequences. For DYS460 and DYS488 the flanking region downstream of the repeat unit is cut unevenly, dependent on the fragment length of the repeat unit. The downstream flanking regions in DYS612 and DYD387S1 were trimmed as they occurred downstream of a poly-adenine stretch, causing lower sequencing quality (Verogen, personal contact).

As Figure 10 and Table 12 highlight, there is a large variation of number of alleles across the 24 YSTR markers included in the ForenSeq amplification kit. Sequencing of the repeat unit

leads to an increased number of allele variants in eleven YSTR markers, and sequencing of the flanking region leads to an increased number of allele variants in six YSTR markers. Two YSTR alleles have sequence variants in both the repeat unit and the flanking region. The YSTR with the largest increase in allele variants are DYF387S1, DYS389II, DYS635, and DYS390. The number of allele variants increased times 3.1, 1.7, 1.5, and 1.4, respectively. For the remaining YSTR markers that had an increase in allele variants the increase varied from 50% to 14%. DYS576, DYS533, DYS505, DYS392, DYS391, DYS19, DYS439, DYS438, and DYS643 had no increase in allele variants when sequenced.

Table 12 – Number of sequence-based allele variants obtained for each YSTR markers using the ForenSeq kit.

# extra sequence-based allele variants	*YSTR markers*
None	DYS505, DYS576, DYS19, DYS391, DYS439, DYS438, DYS643, DYS533, DYS392
1	DYS522, DYS389I, Y-GATA-H4, DYS549, DYS437, DYS448, DYS460, DYS570
2	DYS481
4	DYS635, DYS385a/b, DYS612
7	DYS390
14	DYS389II
22	DYS387S1

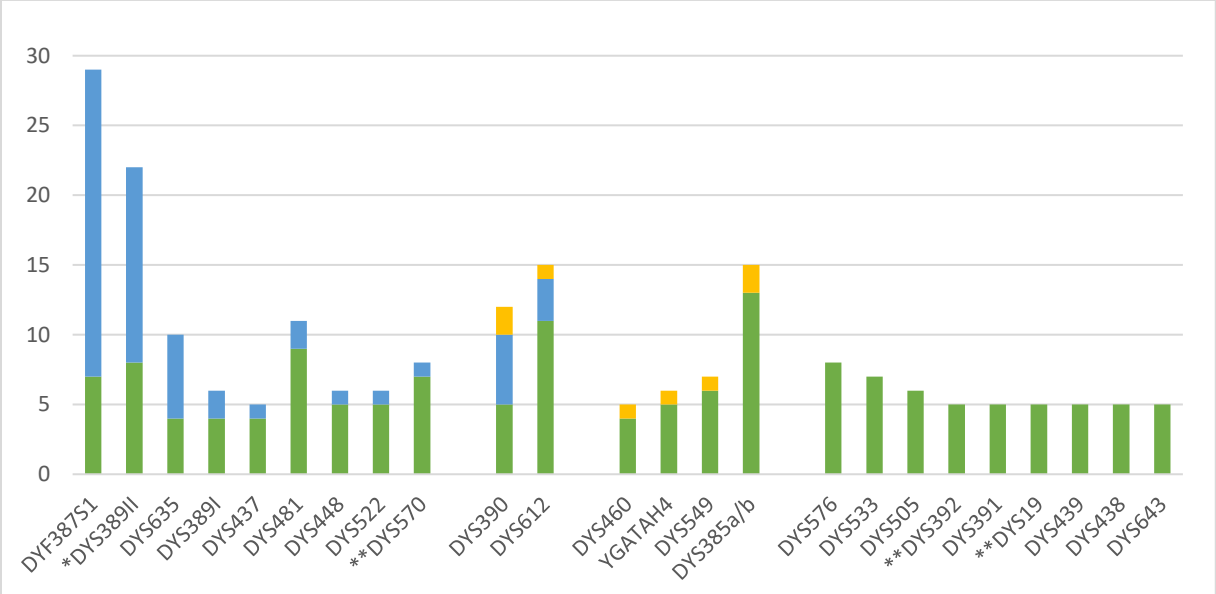


Figure 10 – Length-based vs. sequence-based allele variants obtained from analysis using the ForenSeq amplification kit. The green represents the number of observed length-based allele variants, blue represents the number of sequence-based variants located in the repeat unit, and yellow represents the number of sequence-based allele variants located in the flanking region. YSTR markers with one asterisk is missing the downstream flanking region, and YSTR markers with two asterisks are missing the upstream flanking region.

The following tables are examples of sequence-based allele variants in YSTR markers with the largest increase in allele variants when sequenced. Table 13 show all sequence-based allele variants with fragment length 37 obtained in the YSTR marker DYF387S1. Table 14 show

all sequence-based allele variants with fragment length 29 obtained in the YSTR marker DYS389II. The allele variants in both tables show variation in the repeat sequence only. Full sequence information for all YSTR markers are found in Appendix 2 Table 1-24.

Table 13 – All sequence-based allele variants with fragment length 37 in the DYF387S1 marker obtained using the ForenSeq kit.

Occurrences	*Repeat Unit*
113	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)15
56	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)14
13	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)13
1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)16
1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG AAAG (GAAG)12 (AAAG)13

Upstream flanking region: GAAGAAAGAGAAAA.

Downstream flanking region: AAAATAAAAAAAA

Table 14 – All sequence-based allele variants with fragment length 29 in the DYS389II marker obtained using the ForenSeq kit.

Occurrences	*Repeat Unit*
81	(TCTG)5 (TCTA)11 N48 (TCTG)3 (TCTA)10
13	(TCTG)5 (TCTA)12 N48 (TCTG)3 (TCTA)9
4	(TCTG)4 (TCTA)12 N48 (TCTG)3 (TCTA)10
1	(TCTG)5 (TCTA)11 N48 (TCTG)2 (TCTA)11
1	(TCTG)5 (TCTA)12 N48 (TCTG)2 (TCTA)10

Upstream flanking region: ATCTGTATTATCTATGTGTGTG

N48: TCATTATACCTACTTCTGTATCCAACTCTCATCTGTATTATCTATGTA

Forensic parameters and diversity values

Three sets of length-based haplotypes were assessed in this study. The haplotypes obtained from using the Yfiler Plus and the ForenSeq kit, and the haplotypes obtained using the YSTR markers in the Yfiler kit. Describing calculations were performed for all haplotypes in order to compare the kits. As Table 15 illustrated the YSTR markers included in the three kits provided a different number of unique haplotypes. The YSTR markers in the Yfiler kit produced 271 unique haplotypes, where 18, three, and two haplotypes were observed in two, three and four individuals, respectively. The Yfiler Plus kit produced 290 unique haplotypes, where nine and one haplotypes were observed in two and three individuals, respectively. Lastly, the ForenSeq kit produced 276 unique haplotypes, where only five haplotypes were observed in two individuals. The frequency of unique haplotypes was different in all kits, the ForenSeq kit provides the highest frequency, followed by Yfiler Plus, and Yfiler. The haplotype diversity (HD) and discriminatory capacity (DC) was also highest in the ForenSeq haplotypes, followed by the Yfiler Plus haplotypes, and lastly, the Yfiler haplotypes. The matching probability (MP) was lowest in the Yfiler Plus haplotypes, followed by the ForenSeq haplotypes and lastly, the Yfiler haplotypes.

Table 15 – The estimated forensic parameters and diversity values for Yfiler, Yfiler Plus, and ForenSeq loci in the Norwegian population.

	Yfiler	*Yfiler Plus*	*ForenSeq*
# a haplotype was observed	(n = 301)	(n = 301)	(n = 286)
1	248	280	276
2	18	9	5
3	3	1	-
4	2	-	-
% unique haplotypes	82.39%	93.02%	96.50%
HD	0.9997	0.9998	0.9999
MP	$4.18x10^{-3}$	$3.59x10^{-3}$	$3.62x10^{-3}$
DC	91.51 %	96.55 %	98.22 %

Pairwise genetic distance and multidimensional scaling plot

Pairwise genetic distances (R_{ST}) were calculated between the Norwegian population and thirteen other European populations, and a MDS plot was created for visualization, using the AMOVA tool in the YHRD database (75, 76). In Table 16 the pairwise genetic distance (R_{ST}) with accompanying p-values between the Norwegian and other European populations are presented. Figure 12 shows the MDS. For pairwise genetic distances between the other European populations see Appendix 1 Table 1. Due to multiple comparisons, the Bonferroni correction was used for p-values, lowering the significance level from the standard 0.05 to 0.0004. When using 0.0004 as the significance level, all but the Danish population are significantly different from the Norwegian population. However, the second closest to the Norwegian population is the Russian population. The populations that are genetically most different from the Norwegians are Spain and Greenland, with R_{ST} values of 0.2091 and 0.1857, respectively.

Considering the other populations there are three groups of populations that are relatively close and contain some insignificant p-values (Appendix 1 Table 1). The Norwegian, Danish and Russian, as described above, the Hungarian and Slovenian population, and the Italian, Austrian, German, and Swiss population. The most isolated population is Greenland.

Table 16 – Pairwise genetic distances (R_{ST}) between the Norwegian and other European populations, the values are obtained using the AMOVA tool in the YHRD reference database.

Population	*R_{ST}*	*Norway*	*p-value*
Russian Federation	0.0149		0.0002
Denmark	0.0177		0.0004
Germany	0.0367		0.0000
Austria	0.0392		0.0000
Lithuania	0.0393		0.0000
Slovenia	0.0455		0.0000
Poland	0.0620		0.0000
Hungary	0.0699		0.0000
Switzerland	0.0778		0.0000
Belgium	0.0841		0.0000

Italy	0.0953	0.0000
Greenland	0.1857	0.0000
Spain	0.2091	0.0000

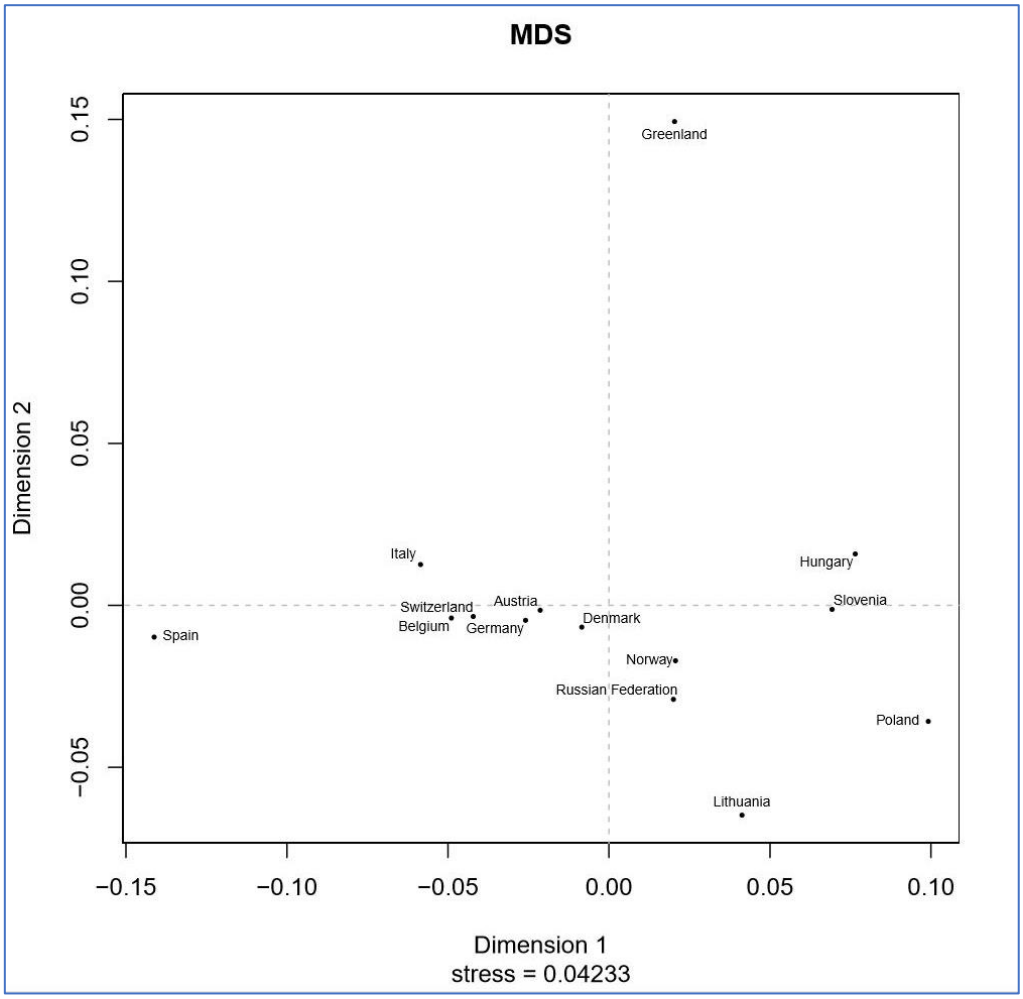


Figure 11 - MDS created using the AMOVA tool in the YHRD reference database comparing 14 European populations.

Haplotype search in the Y-chromosome STR haplotype reference database

YHRD was used to search for potential matches for the haplotypes obtained in this study. Haplotypes obtained from both the Yfiler Plus and Yfiler YSTR markers were used in a search. No matches were found for the 290 haplotypes obtained using the Yfiler Plus kit among the 45,892 Yfiler Plus haplotypes in the database. However, the database provides three calculations that estimate how often you can expect a match. The augmented counting (n+1/N+1) and Kappa calculation is identical for all 290 haplotypes.

The 271 haplotypes including Yfiler YSTR markers only, gave several matches when searched among the 209,111 Yfiler haplotypes in the database. Of the 271 haplotypes 138 haplotypes had one or more matches in the database. 112 haplotypes were observed between one and ten times, 21 were observed between eleven and 46 times, and 5 haplotypes were observed between 50 and 144 times. The remaining 133 haplotypes had no match (Table 17).

Here calculations on how often one can expect a match were also presented. As the table illustrates, the more times a haplotype is observed the smaller the n+1/N+1 and Kappa calculation is.

The DL was also calculated for all searches using the Yfiler Plus and Yfiler haplotypes. However, these values are not included because they are unique for the 561 haplotypes.

Table 17 – Matches obtained using the Yfiler and Yfiler Plus haplotypes to perform a search in YHRD. The table also shows the calculations n+1/N+1 (confidence interval 95%) and Kappa.

# YHRD matches	# Haplotypes	n+1/N+1	Kappa
Yfiler Plus 0	290	45 893	541 208
Yfiler 0	133	209 112	482 825
1	32	104 556	241 412
2	33	69 704	160 941
3	9	52 278	120 706
4	7	41 822	96 565
5	13	34 852	80 470
6	4	29 873	68 975
7	2	26 139	60 353
8	4	23 235	53 647
9	3	20 911	48 282
10	5	19 010	43 893
11	1	17 426	40 235
12	2	16 086	37 140
13	2	14 937	34 487
14	1	13 941	32 188
15	1	13 070	30 176
16	1	12 301	28 401
17	3	11 617	26 823
18	1	11 006	25 411
19	1	10 456	24 141
25	2	8 043	18 570
28	1	7 211	16 649
29	1	6 970	16 094
31	1	6 535	15 088
39	1	5 228	12 070
42	1	4 863	11 228
46	1	4 449	10 272
50	1	4 100	9 467
61	1	3 373	7 787
90	1	2 298	5 305
96	1	2 156	4 977
144	1	1 442	3 329

Haplogroup prediction

YSTR markers can be used to determine an individual's haplogroup, and this was done using Whit Athey's Haplogroup Predictor. Using either the 21-marker or main 111-marker module did not make a difference for the haplogroup prediction. There was only one individual for which the prediction was different, when using "northwest Europe" or "equal priors" in the analysis settings. When selecting "northwest European", this individual predicted to belong to haplogroup I1, and when run with "equal priors", the individual's haplogroup could not be determined (blank). The collection of predicted haplogroups is illustrated as counts and frequencies in Table 18, in addition to the possible place of origin of the haplogroups. Of the 20 haplogroups present in the 111-marker module ten of them are present in the Norwegian population. The majority of the individuals are divided in three haplogroups, R1a, I1, and R1b. The possible place of origin for these haplogroups is Eurasia, northern Europe, and Western Asia. 12% of the individuals are unpredicted.

Table 18 – Results of haplogroup predication using Whit Athey's Haplogroup Predictor (n=300), using both "Northwest Europe" and "Equal Priors" in the analysis settings. Predictions were different for one individual only, affecting the counts for I1 and Blank. The possible place of origin for each haplogroup is also listed.

Haplogroup	*Count*	*Frequency (%)*	*Possible place of origin*
R1a	89	29.67	Eurasia (77, 78)
I1	*83 (84)*	*27.67 (28.00)*	Northern Europe (79)
R1b	76	25.33	Western Asia (80)
Blank	*36 (35)*	*12.00 (11.67)*	-
Q	5	1.67	South Central Siberia (81)
I2b1	4	1.33	Europe (82)
G2a	3	1.00	Western Asia (83)
E1b1b	1	0.33	Horn of Africa (84)
I2a	1	0.33	Europe (82)
J2b	1	0.33	Western Asia (85)
E1b1a1	1	0.33	Horn of Africa (86)

Principal Component Analysis

The first two axes of the PCA projection clearly shows that the observed haplotypes in the individuals in the sample set are divided into three distinct genetic groups, with some individuals being placed in between the two top groups (Figure 12).

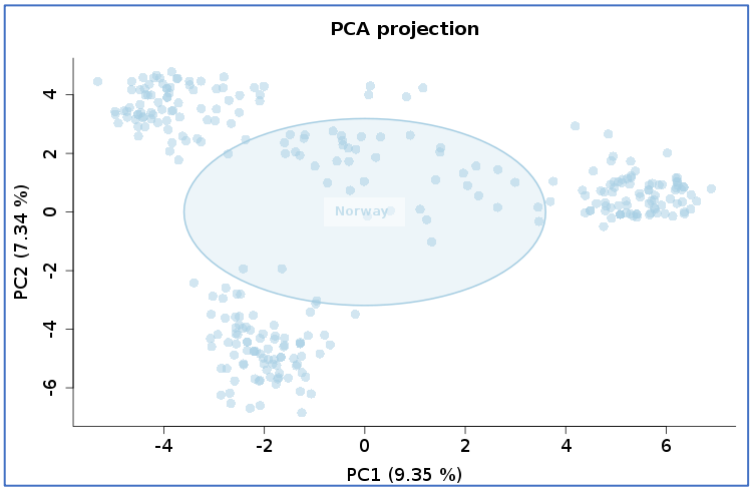


Figure 12 – PCA projection of the Norwegian sample set (n=300) using STRAF v.1.0.5. Each dot represents one sample.

Discussion

Establishment of methods

DNA analysis using the Yfiler Plus PCR Amplification kit was performed on both extracted DNA and punches of FTA cards storing buccal cell samples. Both sample types were successfully amplified and typed using the kit. All reactions were run on half reaction volume. As done in several other studies analyzing high-quality samples, mainly to save costs (72, 87). However, upon analysis of biological stain samples in criminal investigation the full reagent volume should be used in order to increase the chance of successful analysis as the quality and quantity of these sample likely are not as good as those analyzed in this study. It is also proven that choosing an ideal cycle number for buccal cell samples deposited on FTA cards is more difficult than choosing one for the extracted DNA samples due to the variable amount of DNA deposited on the FTA card. When extracting and quantifying DNA samples the DNA concentration is known and can easily by diluted so all analyzed DNA samples contain the same amount of DNA. One of the negative controls contained a contamination in the DYS393 marker. The contamination was not found in any other DNA samples analyzed simultaneously with this control and were therefore not re-analyzed.

The extracted DNA samples and FTA cards were also analyzed using the ForenSeq DNA Signature Prep kit. The analysis was done according to the manufacturer's protocol, with the only deviation that the master mix was added to the wells before the FTA punches. Upon sequencing most of the positive controls showed un-complete analysis results for YSTR markers, and in some cases other STR markers and SNPs. However, the intensity does not guarantee successful DNA-typing or allele dropouts. The DNA sample (not control) with the lowest intensity, 89 170 reads, had typed alleles in all YSTR markers. However, positive control four had an intensity of 159 811 reads and had four dropouts and two signals under the analytical threshold in the YSTR markers. The positive controls were variable and achieved a lower total number of reads than expected, some even having uncomplete profiles. The reason is unclear. Only one lot of the positive control was used, and the controls library was always prepared in the same position. In order to verify if these are the reasons for the low intensity another lot should be tested, and the sample library should be prepared in a different position. Regardless, the analysis results for the DNA samples were adequate, and were therefore used in this study.

Unique allele variants

Several unique allele variants where observed when using the Yfiler Plus and ForenSeq kits. The findings belong to one of three categories, null-allele, microvariant or allele duplication. A null-allele, a type of allele dropout, may occur when the specific region on the DNA template fails to be amplified (88). Null-alleles occur due to hybridization problems between the primer and the DNA template caused by mutations in or near the 3' end of the template leading to little or no synthesis of the DNA template. The type of anomaly was

detected when the same sample was analyzed with two different primer sets (89). Different primers comprise of different oligonucleotides and therefore a specific mutation may cause hybridization issues with one primer and not another (89, 90). In this study, one of the samples containing a null-allele (DYF387S1) was typed as such by both kits, so neither primer could compensate for the mutation. The second sample containing a null-allele (DYS438) was typed using the Yfiler Plus kit but was not sequenced. According to the Y-chromosomal STR haplotype reference database (YHRD) 22 observations of null-alleles are logged so far for DYF387S1, and 18 null-allele observations are logged for DYS438 (91, 92).

The alleles in the allelic ladder have a ±0.5 bp window, and if an allele falls outside this window and does not align with the ladder it becomes an allele variant or a microvariant (88). All microvariants observed in this study were between-ladder alleles, meaning they fall between alleles in the allelic ladder. A few individuals had microvariants in DYS449, DYS627 and DYS385a/b. All microvariants observed in this study are among the microvariants that have been reported to YHRD so far (93-96).

Most of the YSTR markers on the Y-chromosome are single-copy loci, meaning they only exist in one copy and should therefore be represented by one allele. Some YSTR markers are multi-copy markers, and two of them, DYS385a/b and DYF387S1, are included in the two kits used in this study. Four cases of allele-duplications were observed in which two alleles were present in three single-copy YSTR markers, DYS635, DYS392 and DYS612. Regarding the duplication in DYS392, the ForenSeq kit typed 11,17 while the Yfiler Plus kit typed 17. This discrepancy is further discussed in a later paragraph. There were also nine cases of three alleles present in DYF387S1, which is a multi-copy locus and should contain maximum two alleles. Four of the cases are produced by both kits used in this study, while one case is produced by analysis using the ForenSeq kit only. In that case, DYF387S1 is typed 36, 37, 37 by ForenSeq and 36, 37 by Yfiler Plus. Meaning allele 37 is comprised to two alleles with the same fragment length but with a different sequence. Considering the fragment length-based information only, the two kits provide the same results, one 36 allele and one 37 allele.

The presence of more than the expected number of alleles is thought to be caused by duplication events. There are a couple of reasons why mutations generally accumulate faster on the Y-chromosome compared to the other chromosomes (97, 98). The Y-chromosomes contains few genes, so the chromosome is generally under less maintenance pressure. Due to the minimized pressure mutations are more easily adopted. Furthermore, the Y-chromosome does not recombine, allowing the mutations to accumulate. Therefore, the Y-chromosome contains a much higher proportion of segmental duplications compared to the remaining genome average, 35 vs. 5%, respectively (99, 100). A mutation can be a single-base insertion/deletion and lead to a new primer binding site, causing amplification of non-specific amplicons, or it can be a large-scale mutation, duplicating a whole YSTR marker, followed by one of the copies undergoing an independent mutation causing two alleles. Duplication events are caused by non-allelic, homologous recombination (101). In that case it is common that the alleles are one repeat unit apart as single set mutations are common for STR markers, meaning

the allele mutates from fragment length 24 to 25, then 26 and so on (97). As the results show, most of the duplications have happened in RM YSTR markers (DYF387S1, DYS612). Duplications have proven to be more common in YSTR markers so the results are expected (39).

According to YHRD, at the time of citation, several types of a di-allelic results have been reported for DYS635, DYS392, and DYF387S1 (91, 102, 103). DYS612 is not included YHRD. The duplication in DYS635 found in this study is already reported in the YHRD database. However, none of the remaining duplications have been reported to YHRD. That means the duplications are novel, or more likely, these results have been obtained before but removed from the published sample set. Publishers may choose to exclude these results if they suspect contamination or do not want to analyze this data as the "regular" samples (104-106).

As presented, one individual obtained different allele-typing in DYS392 by the two kits. The Yfiler Plus kit typed 17, and the ForenSeq kit typed 11,17. The alleles typed by ForenSeq are unbalanced and have a low number of reads compared to the other YSTR markers in the sample. The sample's overall intensity is 611 964 reads and alleles were typed in all the included YSTR markers. As previously mentioned, duplications are usually one repeat apart, this is six repeat units apart. According to YHRD the allele combination 11,17 has not been observed before (102). However, there has been one observation of 11,14. To verify the typing results obtained in this study, the sample should be re-sequenced. Re-sequencing was not done due to limited time.

Allele frequencies

The allele frequencies obtained from using both kits are based on a different sized sample set, therefore they are not identical. However, they display the same frequency trends. It is clear that there is a large difference in allele polymorphism as some YSTR markers display as little as four allele variants, while others contain many more. Most of the YSTR markers are included in one or more of the kits used in the YHRD database. DYS505, DYS612, and DYS522 are not included. The remaining YSTR markers were assessed to see if the allele frequencies were similar between the database and those obtained in this study. The majority of YSTR markers showed the same trends with some exceptions. Some of the YSTR markers had a different top peak, e.g. DYS448, in this study 20 occurs the most, while in the YHRD database 19 occurs the most (107). These markers displayed trend, though the graph was slightly skewed. Some YSTR markers had "holes" in the frequency, e.g. DYS481 had a negatively skewed bell shape, while the curve obtained in this study looks more like an M, as this study does not contain many 24 alleles. The obtained allele frequencies were further used to calculate the marker genetic diversities of the YSTR alleles.

The YSTR marker with the definitively lowest GD value is DYS392, followed by DYS391, DYS549, and DYS392. The rapidly mutation YSTR markers show the highest GD values, in addition to DYS481 and DYS3851/b. The marker genetic diversities vary between the YSTR markers used in this study, but they also vary between populations. DYS393 was the YSTR

marker with the lowest GD value in this study. However, in other population studies this YSTR marker does not always have the lowest GD value. A summary of seven population studies analyzing 15 populations with the Yfiler Plus kit shows that the following nine YSTR markers were among the three YSTR markers with the lowest GD values, DYS437, DYS391, DYS393, DYS438, DYS389, DYS392, DYS460, DYS389 and DYS456, where the most frequent were the first three (72, 87, 108-112).

The two kits used in this study contain several RM YSTR markers. The Yfiler Plus kit contains six RM YSTR markers and the ForenSeq kit contains four RM YSTR markers. All the RM YSTR markers have GD values above 0.7, ranging from 0.76 to 0.90. Initially these RM YSTR markers were estimated for European populations, but it was in a later study concluded that the RM YSTR markers increase GD values, as well as HD and DC, for populations worldwide (39, 40). Therefore, it is expected that RM YSTR markers have high GD values, and they are frequently observed with a GD value above 0.7 (72, 87, 110, 111). However, one study found that four RM YSTR markers had a lower GD value when analyzing the Algerian population, the GD value for DYS627 was estimated as low as 0.24 (109).

Sequence-based allele variants

Several YSTR markers showed an increase in sequence-based allele variant compared to fragment length-based allele variants. Eleven YSTR markers had an increase when sequencing the repeat unit, while six had an increase when sequencing the flanking region. Sequencing of the repeat unit obtained 58 new allele variants considering all YSTR markers in the kit, while sequencing of the flanking region obtained eight new allele variants.

Many of the YSTR markers that have a large increase in allele-variants when assessing sequence-based information in this study are frequently the YSTR markers with the largest allele variant increase when sequenced in other populations, e.g. DYS387S1, DYS389II, DYS635, DYS448, and DYS390 (113-117). This is expected as some loci are proven to have many sequence variants, while others have few or none (118). The trend is apparently present across propulsions.

Three of the allele variants obtained by sequencing of the flanking region are mutations in the far end of the downstream flanking region and may be a result of sequencing errors (DYS460 allele 10, DYS612 2x allele 33). It is common for sequencing errors to occur more frequently the longer the sequenced DNA strand is. Possibly due to a combination of a weakened polymerase activity and thus the proofreading becomes poorer, as well as few available reagents as most of them have already been used. In two of the cases the mutation leads to the formation of a poly-adenine tail, which promotes polymerase detachment. These samples should be re-sequenced to validate the results. Though sequencing of the flanking region increases the number of allele variants one can question if the effort is worth the results. In this case the flanking region gave an approx. 6% increase in allele variants while the repeat unit gave an approx. 40% increase.

Other studies have also found that the sequence variation in the flanking regions leads to few new alleles compared to the repeat unit (114, 116). Another challenge regarding flanking regions is that different commercially available kits may produce flanking regions of variable length for the same YSTR marker because different primers are used. Making it difficult to compare results obtained using various kits. However, the flanking regions do add some discriminatory power and if a standardization for the length of the flanking regions is determined it would be easier to compare various kits.

To the authors knowledge, 19 of the sequence-based allele variants are novel, while the remaining 47 are presented in previous literature (115, 117, 119). However, as there is no database for sequenced YSTR markers it is cumbersome tracking down previously reported sequence-based variants. In addition, not all studies follow the same nomenclature, making comparison difficult, especially if the provided sequence information is from the opposite strand.

Recently a database for sequenced autosomal STR markers has been create, NOMAUT (www.nomaut.org). Hopefully, a database for sequenced YSTR markers will be created soon. A database will also require that nomenclature is standardized. The apparent amount of novel sequence-based allele variants emphasizes the importance of creating a database and agreeing on a nomenclature, as well as emphasizing the importance of sequence-based population studies. The International Society for Forensic Genetics (ISFG) proposed, in 2016, considerations on minimal nomenclature requirements (120). Examples of considerations are that the forward strand should always be presented, choice of reference sequence, and updated allele frequency databases.

The main reasons for utilizing MPS is an increase in allele variants for several loci. For the purpose of YSTR markers, this means that DNA samples contained more than one male donor can be easier distinguished from one another. MPS makes easier in separating the major and minor component, as well as separate the minor component from the stutters of the major component. It is more sensitive, as all amplified data is sequenced individually, and it gives better YSTR-typing results from samples that are degraded or have a low DNA concentration (118). In addition, MPS can allow separation of related individuals as mutations separating them can be visualized.

Haplotypes and calculations of analytical power

The haplotypes obtained from the Yfiler Plus, Yfiler, and ForenSeq YSTR markers were assessed by the counting method. The Yfiler markers produce the smallest frequency of unique haplotypes and the lowest discriminatory capacity. The Yfiler Plus markers produce a higher frequency of unique haplotypes and an increased discriminatory capacity, in line with several studies (72, 87, 108-112, 121, 122). However, analysis using the ForenSeq markers provide the highest frequency of unique haplotypes and discriminatory capacity. Note that the sample set for ForenSeq is slightly smaller than the Yfiler/Yfiler Plus sample set (286 vs. 301). An increase in unique haplotype frequency was associated with both an increase in HD

and DC, and a decrease in match probability. However, the DC is dependent on the size of the sample set. Therefore, the Yfiler Plus markers have a higher DC than the ForenSeq markers even though they produce a smaller unique haplotype frequency than the ForenSeq markers. This reasoning is supported by previous studies (72, 111, 112).

At the time of sample collection all participants filled out a questionnaire asking about the donor's birth year and place, as well as morphological traits. For the individuals that have identical haplotypes these questionnaires were reviewed to exclude that the same individual was sampled several times. All 301 samples were collected from different individuals. However, it is not possible to determine how closely the individuals are related, or if they in fact are related at all.

A selection of studies from 2015 to 2019 done in various countries have analyzed and compared numerous populations using the Yfiler Plus kit (72, 87, 108-112, 121-123). The following comparison of these studies is based on the frequency of unique haplotypes for each population as the sample size varies considerably. The Danish, Caucasian (USA), African American (USA), Nantong Han (China) and European (Australia) all had a unique haplotype frequency of 1, meaning none of the individuals shared haplotypes. The following populations had a unique frequency between 0.99 and 0.90, Japanese, Austrian, Asian (USA), Mongolian, Libyan, Lithuanian, Hispanic (USA), east Asia (Australia), Duar (inner Mongolia), Somalis, Egyptians, Saudi Arabian, and the Norwegian sample set analyzed in this study. The Algerian and Moroccan populations had a unique haplotype frequency below 0.9, and the lowest frequencies, below 0.6, came from the Greenlandic and Australian Aboriginal population. There is no other population data from Scandinavian countries available for comparison.

Analysis of molecular variance and multidimensional scaling plot

The pairwise genetic distances (R_{ST}) between the Norwegian sample set and thirteen other European populations were calculated with the AMOVA & MDS tool provided with the YHRD database. When applying the Bonferroni correction there is one non-significant p-value, in the comparison between the Norwegian and Danish population. The remaining populations were significantly different from the Norwegian sample set. The second closest population is the Danish population. These results are realistic considering the fact that these are the two populations geographically closest to Norway. Drawing from than conclusion one can postulate that a comparison to a Swedish population lead to values similar to the comparison of Denmark and the Russian Federation due to its geographical position.

The YHRD tool that allows AMOVA also creates an MDS plot containing the same populations as used in the AMOVA. The Danish and Russian Federation populations were closest to the Norwegian sample set. Other geographically close populations, such as Switzerland, Austria, Germany, Italy, and Belgium or the eastern European countries were also close to on the MDS plot. Furthermore, the most secluded populations were Greenland and Spain. The latter is geographically far west of any other populations included in the analysis,

and the Greenlandic populations has a different population history compared to the rest of the European populations.

Haplotype search in the Y-chromosome STR haplotype reference database

The 290 and 271 haplotypes obtained from the Yfiler Plus and Yfiler YSTR markers, respectively, were used to perform a search in the YHRD database. The haplotypes obtained using the ForenSeq kit were not used as not all of the YSTR markers are included in the database. No matches were found for the Norwegian Yfiler Plus haplotypes. However, using the Yfiler haplotypes, a match was found for 138 of the haplotypes. It is obvious that the additional YSTR markers included in the Yfiler Plus kit add more discriminatory power. Of the nine added YSTR markers, six of them are RM and one is highly discriminatory. The increase in discriminatory power may be the result of including more YSTR markers to the haplotype combined with the fact that some of these markers are rapidly mutating and may increase the discriminating power even further. However, it should be noted that the size of the Yfiler Plus database is much lower than the Yfiler database, 45 892 and 209 111, respectively. The table also illustrates that if a haplotype is observed in several times the augmented counting (n+1/N+1) and Kappa calculations are lower than if the haplotype is not previously observed. Future expansion of the Yfiler Plus database may lead to matches also for the Yfiler Plus haplotypes. In addition, no northern European populations, except Denmark and the Russian Federation are included in the Yfiler Plus database. However, all northern European population are included in the Yfiler database. The AMOVA tool showed that geographically close countries have a shorter relative genetic distance. Therefore, it is likely that a large proportion of the Yfiler haplotype matches are to neighboring countries.

The Yfiler Plus and Yfiler haplotypes will be uploaded and registered in YHRD. However, before this can be done the laboratory must perform a laboratory-comparison test in order to upload the haplotypes to the database.

Haplogroup prediction

The individuals in this study were divided in three major haplogroups, R1a, I1 and R1b. The haplogroups were predicted using Whit Athey's Haplogroup Predictor. Haplogroup R1a is also named R-M420 and possibly originates from Eurasia. The R1a haplogroup is one of the widest spread haplogroups, though the substructure within the large geographic area is rather poorly characterized (78). In this study approx. 30% of the individuals were predicted to belong in R1a, somewhat larger than shown in previous studies. One study found the R1a haplogroup in 20% of individuals in Norway (77), though the sample set was rather low (n=74). Another study carried out a few years later with a larger sample set (n=118) concluded the same, that the presence in Norway peaks at approx. 20% (78).

Approx. 28% of the individuals in this study were predicted to belong in haplogroup I1. Haplogroup I1 is also known as I-M253 and most likely originates from Northern Europe (79). One study suggested an I1 frequency of approx. 40% in the Norwegian population, using a

sample set of 72 (79). The haplogroup was also found in high frequencies (≥25%) in Sweden, Germany and in the Sami population. Another study suggested the frequency to be approx. 30% in a Norwegian sample set (n=20) (124).

Lastly, approx. 25% of the individuals were predicted to belong in haplogroup R1bHaplogroup R1b is known as R-M343, Hg1 and Eu18 and possibly originates from Western and central Asia (80). The haplogroup exists in rather large frequencies in Northwest Europe (80). One study has found a frequency of 28% in a sample set of 52 Norwegian males (125), a value consistent with the frequency obtained in this study.

There is also a rather large frequency, approx. 12%, of the individuals for which no haplogroup could be predicted. Depending on the European region selected, northwest Europe or equal priors, there are 35 or 36 samples, respectively, that do not qualify for the haplogroups included in the analysis. The only setting altered was the European region. As the method stated a minimum score and probability was chosen to be 40 and 95%, as recommended by Whit Athey's Haplogroup Predictor. Using other settings may have led to changed results that may have included fewer blank individuals. Other haplogroups that were observed in this study include Q, I2b1, G2a, E1b1b, I2a, J2a and E1b1a1. These haplogroups possibly originate from south central Siberia, Europe, western Asia, horn of Africa, Europe, western Asia and the horn of Africa, respectively (81-86). These haplogroups will not be discussed in further detail due to their small frequencies.

Principal component analysis

Since the haplogroup analysis clearly illustrated three major haplogroups a principal component analysis (PCA) was performed to see if the haplogroup division could be supported by a PCA plot. A PCA is used for quality control and population substructure detection (126, 127). PCA plots can be created based on SNPs or YSTR markers, and the plot can be used to compare populations or individual samples within a population (128). The PCA plot clearly illustrates a threefold division in the sample set, in compliance with the haplogroup results. Each dot on the PCA plot represents a sample, and each dot could be clicked to see what sample it represented. In total 15 dots from each of the three major groups were inspected and compared to the haplogroup they were assigned to. The comparison lead to the assumption that the top left cluster represents the R1b haplogroup, the bottom left cluster represents R1a and the cluster to the right represents I1. The division of the clusters also indicates that the R1a and R1b haplogroup are closer related than to I1. To the authors knowledge, there is no previous study where a PCA plot has been created based on a Norwegian sample set comprising of YSTR markers. Nevertheless, PCA plots have been extensively used to compare populations based on YSTR markers (129-134).

Conclusion and future perspectives

The collection of biological material in this study has made a significant contribution to the research biobank at CFG. All males in the current research biobank that were third generation Norwegian on their farther father's side (n=301) were successfully typed using the Yfiler Plus PCR Amplification kit, and 286 samples were successfully sequenced using the ForenSeq DNA Signature Prep and MiSeq FGx Reagent kit, with minimal changes to the manufacturers' protocol. Further work is needed to sequence the remaining samples. Furthermore, is has to be established if biological material from more male contributors has to be collected and analyzed to for fill requirements for publishing the Norwegian sample data set to the Y-chromosomal STR haplotype reference database (YHRD). Any future collection and analysis will further improve the result and likely strengthen the findings of this study. The type of samples examined in this study are reference samples. Before being able to use YSTR analyses for casework, the laboratory needs to perform a thorough validation study with relevant sample types, including low DNA-concentrations, degraded DNA and mixtures.

The concordance between the Yfiler Plus and the ForenSeq kit was very high considering typing results based on allele-length. The number of allele variants increased in 15 of the YSTR markers when including sequence information of both the repeat units and flanking regions obtained using the ForenSeq kit. Of the two regions, the variation in the repeat unit provided the largest increase in new allele variants. The implementation of sequencing-based analysis methods will become more and more common in forensic genetics in the near future. Sequence-based databases are needed to gather as much population data as possible, allowing easier search and matching of sequence variants, as YHRD provides for fragment length-based alleles. Databases start to appear, but it is not possible to report population data for YSTR-markers yet. This will also require establishment of a standard nomenclature. Several apparent novel sequence variants were discovered in this study, underlining the importance of sequence-based populations studies, as YSTR markers in a Norwegian sample set have not previously been sequenced

Based on the forensic parameters and diversity values it is clear that the inclusion of rapidly mutating YSTR markers increased the discriminatory capacity and decreases the matching probability. The ForenSeq kit had a higher frequency of unique haplotypes and discriminatory capacity compared to the Yfiler Plus and Yfiler loci. Analysis with the ForenSeq is quite expensive so the Yfiler Plus kit is a cheaper option as it is also highly discriminating. The ForenSeq kit also collects information on other STR markers as well as SNPs in the same analysis. If the goal is to gather more information than the Y-chromosomal haplotype, analysis using ForenSeq may be more convenient that running more analysis with different kits.

Comparing the genetic distances between the Norwegian sample set to other populations showed that the Norwegian population is significantly different from almost all other European populations included in the analysis. However, the comparison only included one other Scandinavian population. Future inclusion of other Scandinavian populations will give a clearer image of the genetic distances in northern Europe.

References

1. Voldtektssituasjonen i Norge 2017 www.politiet.no2018 [Available from: https://www.politiet.no/globalassets/04-aktuelt-tall-og-fakta/voldtekt-og-seksuallovbrudd/voldtektssituasjonen-i-norge-2017.
2. Soares-Vieira JA, Billerbeck AE, Iwamura ES, Zampieri RA, Gattas GJ, Munoz DR, et al. Y-STRs in forensic medicine: DNA analysis in semen samples of azoospermic individuals. J Forensic Sci. 2007;52(3):664-70.
3. Butler JM. Chapter 16 - Lineage Markers: Y Chromosome and mtDNA Testing. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 363-96.
4. Dutta P, Bhosale S, Singh R, Gubrellay P, Patil J, Sehdev B, et al. Amelogenin Gene - The Pioneer in Gender Determination from Forensic Dental Samples. J Clin Diagn Res. 2017;11(2):ZC56-ZC9.
5. Kashyap VK, Sahoo S, Sitalaximi T, Trivedi R. Deletions in the Y-derived amelogenin gene fragment in the Indian population. BMC Med Genet. 2006;7:37.
6. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136(5):621-35.
7. Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet. 2015;18:49-65.
8. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015;18:33-48.
9. Urquhart A, Kimpton CP, Downes TJ, Gill P. Variation in Short Tandem Repeat sequences —a survey of twelve microsatellite loci for use as forensic identification markers. Int J Legal Med. 1994;107(1):13-20.
10. Tautz D. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. EXS. 1993;67:21-8.
11. Kwong M, Pemberton TJ. Sequence differences at orthologous microsatellites inflate estimates of human-chimpanzee differentiation. BMC Genomics. 2014;15:990.
12. Butler JM. Chapter 8 - Short Tandem Repeat Markers. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 147-73.
13. Urquhart A, Oldroyd NJ, Kimpton CP, Gill P. Highly discriminating heptaplex short tandem repeat PCR system for forensic identification. Biotechniques. 1995;18(1):116-8, 20-1.
14. Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A. 1997;94(3):1041-6.
15. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A. 1998;95(18):10774-8.
16. Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics. 1992;12(2):241-53.
17. Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. PCR Methods Appl. 1993;3(1):13-22.
18. Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, et al. Current trends in microsatellite genotyping. Mol Ecol Resour. 2011;11(4):591-611.
19. Harding RM, Boyce AJ, Clegg JB. The evolution of tandemly repetitive DNA: recombination rules. Genetics. 1992;132(3):847-59.
20. Krebs JE, Goldstein ES, Kilpatrick ST. Clusters and Repeats. Lewin's Genes XII: Jones & Bertlett Learning 2018.
21. Goldstein DB, Schlotter C. Microsatellites: Evolution and Applications. New York: Oxford University Press; 1999. p. 343.
22. Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet. 1992;89(4):389-94.

23. Roewer L, Epplen JT. Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. Forensic Sci Int. 1992;53(2):163-71.
24. Gopinath S, Zhong C, Nguyen V, Ge J, Lagace RE, Short ML, et al. Developmental validation of the Yfiler((R)) Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications. Forensic Sci Int Genet. 2016;24:164-75.
25. Hall A, Ballantyne J. The development of an 18-locus Y-STR system for forensic casework. Anal Bioanal Chem. 2003;376(8):1234-46.
26. Hanson EK, Ballantyne J. A highly discriminating 21 locus Y-STR "megaplex" system designed to augment the minimal haplotype loci for forensic casework. J Forensic Sci. 2004;49(1):40-51.
27. Hanson EK, Ballantyne J. An ultra-high discrimination Y chromosome short tandem repeat multiplex DNA typing system. PLoS One. 2007;2(8):e688.
28. Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med. 1997;110(3):125-33, 41-9.
29. Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, et al. Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. Forensic Sci Int. 2005;151(1):111-24.
30. Lim SK, Xue Y, Parkin EJ, Tyler-Smith C. Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals. Int J Legal Med. 2007;121(2):124-7.
31. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, et al. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci. 2006;51(1):64-75.
32. Rodig H, Roewer L, Gross A, Richter T, de Knijff P, Kayser M, et al. Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. Forensic Sci Int. 2008;174(2-3):182-8.
33. Thompson JM, Ewing MM, Frank WE, Pogemiller JJ, Nolde CA, Koehler DJ, et al. Developmental validation of the PowerPlex(R) Y23 System: a single multiplex Y-STR analysis system for casework and database samples. Forensic Sci Int Genet. 2013;7(2):240-50.
34. Vermeulen M, Wollstein A, van der Gaag K, Lao O, Xue Y, Wang Q, et al. Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. Forensic Sci Int Genet. 2009;3(4):205-13.
35. Ayub Q, Mohyuddin A, Qamar R, Mazhar K, Zerjal T, Mehdi SQ, et al. Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. Nucleic Acids Res. 2000;28(2):e8.
36. Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, et al. Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int. 2001;118(2):106-13.
37. Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med. 1997;110(3):125-33.
38. Butler JM. Recent Developments in Y-Short Tandem Repeat and Y-Single Nucleotide Polymorphism Analysis. Forensic Sci Rev. 2003;15(2):91-111.
39. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet. 2010;87(3):341-53.
40. Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, et al. A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Sci Int Genet. 2012;6(2):208-18.
41. Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, et al. Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. Hum Mutat. 2014;35(8):1021-32.
42. Adnan A, Ralf A, Rakha A, Kousouri N, Kayser M. Improving empirical evidence on differentiating closely related men with RM Y-STRs: A comprehensive pedigree study from Pakistan. Forensic Sci Int Genet. 2016;25:45-51.

43. Niederstatter H, Berger B, Kayser M, Parson W. Differences in urbanization degree and consequences on the diversity of conventional vs. rapidly mutating Y-STRs in five municipalities from a small region of the Tyrolean Alps in Austria. Forensic Sci Int Genet. 2016;24:180-93.
44. Morse SP. Genealogy Beyond the Y Chromosome Autosomes Exposed www.stevemorse.org2012 [cited 2019 04.05]. Available from: https://stevemorse.org/genetealogy/beyond.htm.
45. Strachan T, Read AP, Strachan T. Chromosome Structure and Function Human molecular genetics. New York: Garland Science; 2011. p. 31-6.
46. Helena Mangs A, Morris BJ. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Curr Genomics. 2007;8(2):129-36.
47. Butler JM. Chapter 11 - Statistical Interpretation: Evaluating the Strength of Forensic DNA Evidence. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 229-58.
48. Aims and Objectives: YHRD; [Available from: https://yhrd.org/.
49. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic Sci Int Genet. 2015;18:118-30.
50. Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet. 2015;18:78-89.
51. Butler JM. Chapter 7 - DNA Amplification (The Polymerase Chain Reaction). In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 125-46.
52. Butler JM. Chapter 9 - Fundamentals of DNA Separation and Detection. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 175-203.
53. Apblum. Capillary electrophoresis. In: Capillaryelectrophoresis, editor. www.wikipedia.org2004.
54. User Guide: Yfiler Plus PCR Amplification Kit: Thermo Fisher Scientific; [cited 2019 16.03]. Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2F4485610_YfilerPlus_UG.pdf&title=VXNlciBHdWlkZTogWWZpbGVyIFBsdXMgUENSIEFtcGxpZmljYXRpb24gS2l0.
55. 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 Sci Int Genet. 2015;17:110-21.
56. Targeted Sequencing www.thermofisher.com: Thermo Fisher Scientific; [Available from: https://www.thermofisher.com/no/en/home/life-science/sequencing/dna-sequencing/targeted-sequencing.html.
57. Yang Y, Xie B, Yan J. Application of Next-generation Sequencing Technology in Forensic Science. Genomics, Proteomics Bioinformatics. 2014;12(5):190-7.
58. Selecting the best NGS enrichment assay for your needs Oxford Gene Technology - A Sysmex Group Company 2016 [Available from: https://www.ogt.com/resources/literature/1357_selecting_the_best_ngs_enrichment_assay_for_your_needs.
59. Targeted Next-Generation Sequencing for Forensic Genomics Illumina; [cited 2019 15.05]. Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/appspotlights/app_spotlight_forensics.pdf.
60. Illummina Sequencing by Synthesis. www.youtube.com: Illumina.
61. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnol. 2008;26:1135.
62. ForenSeq DNA Signature Prep Reference Guide: Verogen; [cited 2019 18.03]. Available from: https://verogen.com/wp-content/uploads/2018/08/ForenSeq-DNA-Prep-Guide-VD2018005-A.pdf.
63. Kirsebom MK. Forensic DNA phenotyping SNP-based prediction of eye and hair colour in the Norwegian population.: Arctic University of Tromsø; 2016.
64. Zakariassen M. Predicting visible traits in a Norwegian population A prototype SNP panel for massive parallel sequencing.: Arctic University of Tromsø; 2016.

65. Gusmão L, Butler JM, Linacre A, Parson W, Roewer L, Schneider PM, et al. Revised guidelines for the publication of genetic population data. Forensic Sci Int Genet. 2017;30:160-3.
66. Buadu S. Forensic DNA genotyping by means of next generation sequencing.: Arctic University of Tromsø; 2018.
67. Quantifiler HP and Trio DNA Quantification Kits: Thermo Fisher Scientific; [cited 2019 20.03]. Available from: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485356_Quantifiler_HP_Trio_DNA_QR.pdf.
68. PrepFiler Express and PrepFiler Express Forensic DNA Extraction Kits: Thermo Fisher Scientific; [cited 2019 20.03]. Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2F4443104_PrepFilrExPFExBTA_QR.pdf&title=UXVpY2sgUmVmZXJlbmNlOiBQcmVwRmlsZXIgRXhwcmVzcyBhbmQgUHJlcEZpbGVyIEV4cHJlc3MgRm9yZW5zaWMgRE5BIEV4dHJhY3Rpb24gS2l0cw==.
69. Qubit dsDNA HS Assay Kits: Thermo Fisher Scientific; [cited 2019 19.03]. Available from: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FQubit_dsDNA_HS_Assay_UG.pdf&title=VXNlciBHdWlkZTogUXViaXQgZHNETkEgSFMgQXNzYXkgS2l0cw==.
70. MiSeq FGx Instrument Reference Guide: Verogen; [cited 2019 18.03]. Available from: https://verogen.com/wp-content/uploads/2018/08/MiSeq-FGx-Reference-Guide-VD2018006-A.pdf.
71. Nei M, Tajima F. DNA Polymorphism Detectable by Restriction Endonucleases. Genetics. 1981;97(1):145.
72. Olofsson JK, Mogensen HS, Buchard A, Børsting C, Morling N. Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit. Forensic Sci Int Genet. 2015;16:232-6.
73. Andersen MM, Eriksen PS, Morling N. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. J Theor Biol. 2013;329:39-51.
74. Brenner CH. Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. Forensic Sci Int Genet. 2010;4(5):281-91.
75. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992;131(2):479-91.
76. Kruskal JB. Nonmetric multidimensional scaling: A numerical method. Psychometrika. 1964;29(2):115-29.
77. Underhill PA, Myres NM, Rootsi S, Metspalu M, Zhivotovsky LA, King RJ, et al. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet. 2010;18(4):479-84.
78. Underhill PA, Poznik GD, Rootsi S, Jarve M, Lin AA, Wang J, et al. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. Eur J Hum Genet. 2015;23(1):124-31.
79. Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, et al. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am J Hum Genet. 2004;75(1):128-37.
80. Myres NM, Rootsi S, Lin AA, Järve M, King RJ, Kutuev I, et al. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet. 2010;19:95.
81. Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. Mol Biol Evol. 2004;21(1):164-75.
82. H. M. Haplogroup I2 www.eupedia.com2018 [Available from: https://www.eupedia.com/europe/Haplogroup_I2_Y-DNA.shtml#I2a2a.
83. Genealogy ISoG. Y-DNA Haplogroup G and its Subclades - 2018 Internetional Society of Genetic Genealogy 2018 [Available from:

https://docs.google.com/spreadsheets/d/1hW2SxSLFSJS3r_MIdw9zxsS8NXo_4PG0_fFFeXGwEyc/edit#gid=0.
84. Vilain R. [Treatment of steatomery in the female: theory and practice]. Ann Chir Plast. 1975;20(2):135-46.
85. Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, et al. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet. 2004;115(5):357-71.
86. Trombetta B, Cruciani F, Sellitto D, Scozzari R. A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. PLoS One. 2011;6(1):e16073.
87. Pickrahn I, Müller E, Zahrer W, Dunkelmann B, Cemper-Kiesslich J, Kreindl G, et al. Yfiler® Plus amplification kit validation and calculation of forensic parameters for two Austrian populations. Forensic Sci Int Genet. 2016;21:90-4.
88. Butler JM. Chapter 10 - STR Genotyping and Data Interpretation. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 205-27.
89. Walsh S. Non-amplification of a vWA allele [2]1998. 1103-4 p.
90. Westen AA, Kraaijenbrink T, Robles de Medina EA, Harteveld J, Willemse P, Zuniga SB, et al. Comparing six commercial autosomal STR kits in a large Dutch population sample. Forensic Sci Int Genet. 2014;10:55-63.
91. Locus Information on DYF387S1 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYF387S1.
92. Locus Information on DYS438 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYS438.
93. Locus Information on DYS449 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYS449.
94. Locus Information on DYS627 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYS627.
95. Locus Information on DYS385 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYS385.
96. Locus Information on DYS458 www.YHRD.org: YHRD; [cited 2019 02.05]. Available from: https://yhrd.org/tools/marker/DYS458.
97. Butler J, Decker A, Kline M, Vallone P. Chromosomal Duplications Along the Y-Chromosome and Their Potential Impact on Y-STR Interpretation. 2005;50(4):JFS2004481-7.
98. Kuan L-C, Su M-T, Kuo P-L, Kuo T-C. Direct duplication of the Y chromosome with normal phenotype – incidental finding in two cases. Andrologia. 2013;45(2):140-4.
99. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423(6942):825-37.
100. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297(5583):1003-7.
101. Jobling MA. Copy number variation on the human Y chromosome. Cytogenet Genome Res. 2008;123(1-4):253-62.
102. Locus Information on DYS392 www.YHRD.org: YHRD; [cited 2019 03.05]. Available from: https://yhrd.org/tools/marker/DYS392.
103. Locus Information on DYS635 www.YHRD.org: YHRD; [cited 2019 03.05]. Available from: https://yhrd.org/tools/marker/DYS635.
104. Coyle HM, Budowle B, Bourke MT, Carita E, Hintz JL, Ladd C, et al. Population data for seven Y-chromosome STR loci from three different population groups residing in Connecticut. J Forensic Sci. 2003;48(2):435-7.
105. Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. Am J Hum Genet. 2002;71(3):466-82.

106. Ciavaglia S, Linacre A. An optimised forensic STR multiplex assay set for the Australasian carpet python. Forensic Sci Int Genet. 2018;34:231-48.
107. Locus Information on DYS448 www.YHRD.org: YHRD; [cited 2019 03.05]. Available from: https://yhrd.org/tools/marker/DYS448.
108. Jankauskiene J, Kukiene J, Ivanova V, Aleknaviciute G. Population data and forensic genetic evaluation with the Yfiler™ Plus PCR Amplification kit in the Lithuanian population. Forensic Sci Int Genet. 2017;6:e606-e7.
109. D'Atanasio E, Iacovacci G, Pistillo R, Bonito M, Dugoujon J, Moral P, et al. Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa. Forensic Sci Int Genet. 2019;38:185-94.
110. Henry J, Dao H, Scandrett L, Taylor D. Population genetic analysis of Yfiler® Plus haplotype data for three South Australian populations. Forensic Sci Int Genet. 2019.
111. Tao R, Wang S, Zhang J, Zhang J, Yang Z, Zhang S, et al. Genetic characterization of 27 Y-STR loci analyzed in the Nantong Han population residing along the Yangtze Basin. Forensic Sci Int Genet. 2019;39:e10-e3.
112. Watahiki H, Fujii K, Fukagawa T, Mita Y, Kitayama T, Mizuno N. Polymorphisms and microvariant sequences in the Japanese population for 25 Y-STR markers and their relationships to Y-chromosome haplogroups. Forensic Sci Int Genet. 2019.
113. Wendt FR, Churchill JD, Novroski NMM, King JL, Ng J, Oldt RF, et al. Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx™ forensic genomics system. Forensic Sci Int Genet. 2016;24:18-23.
114. 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 Sci Int Genet. 2016;25:214-26.
115. 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 Sci Int Genet. 2017;28:1-9.
116. Wendt FR, King JL, Novroski NMM, Churchill JD, Ng J, Oldt RF, et al. Flanking region variation of ForenSeq™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans. Forensic Sci Int Genet. 2017;28:146-54.
117. Casals F, Anglada R, Bonet N, Rasal R, van der Gaag KJ, Hoogenboom J, et al. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. Forensic Sci Int Genet. 2017;30:66-70.
118. de Knijff P. Smoothing the Tranistion to MPS for Forensic Laboratories [Internet]. ForensicConnect; 2018 29.03.2018. Podcast
119. Staadig A, Tillmar A. An overall limited effect on the weight-of-evidence when taking STR DNA sequence polymorphism into account in kinship analysis. Forensic Sci Int Genet. 2019;39:44-9.
120. 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 Sci Int Genet. 2016;22:54-63.
121. Gopinath S, Zhong C, Nguyen V, Ge J, Lagacé RE, Short ML, et al. Developmental validation of the Yfiler® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications. Forensic Sci Int Genet. 2016;24:164-75.
122. Khubrani YM, Wetton JH, Jobling MA. Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. Forensic Sci Int Genet. 2018;33:98-105.
123. Wang C, Su M, Li Y, Chen L, Jin X, Wen S, et al. Genetic polymorphisms of 27 Yfiler® Plus loci in the Daur and Mongolian ethnic minorities from Hulunbuir of Inner Mongolia Autonomous Region, China. Forensic Sci Int Genet. 2019;40:e252-e5.
124. Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, et al. Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun. 2015;6:7152.

125. Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, et al. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet. 2000;67(6):1526-43.
126. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nature Genet. 2008;40:646.
127. Gouy A, Zieger M. STRAF—A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci Int Genet. 2017;30:148-51.
128. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
129. Lovell A, Moreau C, Yotova V, Xiao F, Bourgeois S, Gehl D, et al. Ethiopia: between Sub-Saharan Africa and Western Eurasia. Ann Hum Genet. 2005;69(3):275-87.
130. Capredon M, Brucato N, Tonasso L, Choesmel-Cadamuro V, Ricaut FX, Razafindrazaka H, et al. Tracing Arab-Islamic inheritance in Madagascar: study of the Y-chromosome and mitochondrial DNA in the Antemoro. PLoS One. 2013;8(11):e80932.
131. Capelli C, Redhead N, Romano V, Calì F, Lefranc G, Delague V, et al. Population Structure in the Mediterranean Basin: A Y Chromosome Perspective. Ann Hum Genet. 2006;70(2):207-25.
132. Al-Zahery N, Pala M, Battaglia V, Grugni V, Hamod MA, Hooshiar Kashani B, et al. In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. BMC Evol Biol. 2011;11:288.
133. Di Cristofaro J, Pennarun E, Mazieres S, Myres NM, Lin AA, Temori SA, et al. Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS One. 2013;8(10):e76748.
134. Voskarides K, Mazieres S, Hadjipanagi D, Di Cristofaro J, Ignatiou A, Stefanou C, et al. Y-chromosome phylogeographic analysis of the Greek-Cypriot population reveals elements consistent with Neolithic and Bronze Age settlements. Investig Genet. 2016;7:1.

Appendix 1

Table 1 – Pairwise genetic distance (R_{ST}) and the corresponding p-values comparing 14 European populations. The marginally significant p-values are bold and green, and the non-significant p-values bold and red.

Population	*Norway*	*Austria*	*Belgium*	*Denmark*	*Germany*	*Greenland*	*Hungary*	*Italy*	*Lithuania*	*Poland*	*Russian F.*	*Slovenia*	*Spain*	*Switzerland*
Norway	-	0.0000	0.0000	**0.0004**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000
Austria	0.0392	-	0.0001	**0.0027**	**0.0076**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0004**
Belgium	0.0841	0.0197	-	0.0000	**0.0025**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0393**
Denmark	0.0177	0.0099	0.0412	-	**0.0016**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
Germany	0.0367	0.0045	0.0104	0.0102	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0020
Greenland	0.1857	0.1260	0.1703	0.1577	0.1431	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Hungary	0.0699	0.0757	0.1484	0.0879	0.0947	0.1470	-	0.0000	0.0000	0.0000	0.0000	**0.0007**	0.0000	0.0000
Italy	0.0953	0.0267	0.0196	0.0593	0.0307	0.1701	0.1322	-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Lithuania	0.0393	0.1028	0.1330	0.0847	0.0922	0.2284	0.1106	0.1439	-	0.0000	0.0000	0.0000	0.0000	0.0000
Poland	0.0620	0.1294	0.1886	0.1076	0.1225	0.2449	0.0687	0.2002	0.0684	-	0.0000	0.0000	0.0000	0.0000
Russian F.	0.0149	0.0381	0.0644	0.0325	0.0336	0.1487	0.0621	0.0759	0.0218	0.0601	-	0.0000	0.0000	0.0000
Slovenia	0.0455	0.0671	0.1355	0.0615	0.0781	0.1584	0.0142	0.1377	0.0932	0.0310	0.0471	-	0.0000	0.0000
Spain	0.2091	0.1183	0.0490	0.1695	0.0887	0.2822	0.2931	0.0825	0.2378	0.2990	0.1617	0.2795	-	0.0000
Switzerland	0.0778	0.0161	0.0076	0.0271	0.0117	0.1662	0.1515	0.0312	0.1398	0.1916	0.0701	0.1321	0.0818	-

The pairwise genetic distances (R_{ST}) were calculated based on population data from the following studies:

- 392 Haplotypes from Austria
  - Erhart D., Berger B., Niederstätter H., Gassner C., Schennach H., Parson W. (2012), 'Frequency data for 17 Y-chromosomal STRs and 19 Y-chromosomal SNPs in the Tyrolean district of Reutte, Austria.', *International Journal of Legal Medicine* 126(6), 977-8.
  - Pickrahn I., Müller E., Zahrer W., Dunkelmann B., Cemper Kiesslich J., Kreindl G., Neuhuber F. (2016), 'Yfiler (®) Plus amplification kit validation and calculation of forensic parameters for two Austrian populations.', *Forensic Science International Genetics* 21, 90-94
  - Niederstätter H., Berger B., Kayser M., Parson W. (2016), 'Differences in urbanization degree and consequences on the diversity of conventional vs. rapidly mutating Y-STRs in five municipalities from a small region of the Tyrolean Alps in Austria.', *Forensic Science International: Genetics* 24, 180-93
  - Roewer L., Croucher PJ., Willuweit S., Lu TT., Kayser M., Lessig R., Knijff D., Jobling MA., Tyler Smith C., Krawczak M. (2005), 'Signature of recent historical events in the European Y-chromosomal STR haplotype distribution.', *Hum Genet* 116(4), 279-91

- Berger B., Lindinger A., Niederstätter H., Grubwieser P., Parson W. (2005), 'Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay.', *Int J Legal Med*119(4), 241-6

- 160 Haplotypes from Belgium
  - Roewer L., Krawczak M., Willuweit S., Nagy M., Alves C., Amorim A., Anslinger K., Augustin C., Betz A., Bosch E., Cagliá A., Carracedo A., Corach D., Dekairelle AF., Dobosz T., Dupuy BM., Füredi S., Gehrig C., Gusmaõ L., Henke J., Henke L., Hidding M., Hohoff C., Hoste B., Jobling MA., Kärgel HJ., Knijff D., Lessig R., Liebeherr E., Lorente M., Martínez Jarreta B., Nievas P., Nowak M., Parson W., Pascali VL., Penacino G., Ploski R., Rolf B., Sala A., Schmidt U., Schmitt C., Schneider PM., Szibor R., Teifel Greding J., Kayser M. (2001), 'Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes.', *Forensic Sci Int* 118(2-3), 106-13
  - Maesschalck D., Vanhoutte E., Knaepen K., Vanderheyden N., Cassiman JJ., Decorte R. (2005), 'Y-chromosomal STR haplotypes in a Belgian population sample and identification of a micro-variant with a flanking site mutation at DYS19.', *Forensic Sci Int*152(1), 89-94
  - Mertens G., Jehaes E., Leijnen G., Rand S., Jacobs W., Marck V. (2007), 'Twelve-locus Y-STR haplotypes in the Flemish population.', *J Forensic Sci* 52(3), 755-7
- 177 Haplotypes from Denmark (no published study)
- 495 Haplotypes from Germany
  - Purps J., Siegert S., Willuweit S., Nagy M., Alves C., Salazar R., Angustia SM., Santos LH., Anslinger K., Bayer B., Ayub Q., Wei W., Xue Y., Tyler Smith C., Bafalluy MB., Martínez Jarreta B., Egyed B., Balitzki B., Tschumi S., Ballard D., Court DS., Barrantes X., Bäßler G., Wiest T., Berger B., Niederstätter H., Parson W., Davis C., Budowle B., Burri H., Borer U., Koller C., Carvalho EF., Domingues PM., Chamoun WT., Coble MD., Hill CR., Corach D., Caputo M., D'amato ME., Davison S., Decorte R., Larmuseau MH., Ottoni C., Rickards O., Lu D., Jiang C., Dobosz T., Jonkisz A., Frank WE., Furac I., Gehrig C., Castella V., Grskovic B., Haas C., Wobst J., Hadzic G., Drobnic K., Honda K., Hou Y., Zhou D., Li Y., Hu S., Chen S., Immel UD., Lessig R., Jakovski Z., Ilievska T., Klann AE., García CC., Knijff D., Kraaijenbrink T., Kondili A., Miniati P., Vouropoulou M., Kovacevic L., Marjanovic D., Lindner I., Mansour I., Al Azem M., Andari AE., Marino M., Furfuro S., Locarno L., Martín P., Luque GM., Alonso A., Miranda LS., Moreira H., Mizuno N., Iwashima Y., Neto RS., Nogueira TL., Silva R., Nastainczyk Wulf M., Edelmann J., Kohl M., Nie S., Wang X., Cheng B., Núñez C., Pancorbo MM., Olofsson JK., Morling N., Onofri V., Tagliabracci A., Pamjav H., Volgyi A., Barany G., Pawlowski R., Maciejewska A., Pelotti S., Pepinski W., Abreu Glowacka M., Phillips C., Cárdenas J., Rey Gonzalez D., Salas A., Brisighelli F., Capelli C., Toscanini U., Piccinini A., Piglionica M., Baldassarra SL., Ploski R., Konarzewska M., Jastrzebska E., Robino C., Sajantila A., Palo JU., Guevara E., Salvador J., Ungria MC., Rodriguez JJ., Schmidt U., Schlauderer N., Saukko P., Schneider PM., Sirker M., Shin KJ., Oh YN., Skitsa I., Ampati A., Smith TG., Calvit LS., Stenzl V., Capal T., Tillmar A., Nilsson H., Turrina S., Leo D., Verzeletti A., Cortellini V., Wetton JH., Gwynne GM., Jobling MA., Whittle MR., Sumita DR., Wolańska Nowak P., Yong RY., Krawczak M., Nothnagel M., Roewer L. (2014), 'A global analysis of Y-chromosomal haplotype diversity for 23 STR loci.', *Forensic Sci Int Genet* 12, 12-23
  - Anslinger K., Keil W., Weichhold G., Eisenmenger W. (2000), 'Y-chromosomal STR haplotypes in a population sample from Bavaria.', *Int J Legal Med* 113(3), 189-92
  - Schmidt U., Meier N., Lutz S. (2003), 'Y-chromosomal STR haplotypes in a population sample from southwest Germany (Freiburg area).', *Int J Legal Med* 117(4), 211-7

- Schneider PM., Meuser S., Waiyawuth W., Seo Y., Rittner C. (1998), 'Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations.', *Forensic Sci Int* 97(1), 61-70
- Hidding M., Schmitt C. (2000), 'Haplotype frequencies and population data of nine Y-chromosomal STR polymorphisms in a German and a Chinese population.', *Forensic Sci Int* 113(1-3), 47-53
- Henke J., Henke L., Chatthopadhyay P., Kayser M., Dülmer M., Cleef S., Pöche H., Felske Zech H. (2001), 'Application of Y-chromosomal STR haplotypes to forensic genetics.', *Croat Med J* 42(3), 292-7
- Immel UD., Kleiber M., Klintschar M. (2005), 'Y chromosome polymorphisms and haplotypes in South Saxony-Anhalt (Germany).', *Forensic Sci Int* 155(2-3), 211-5
- Rodig H., Grum M., Grimmecke HD. (2007), 'Population study and evaluation of 20 Y-chromosome STR loci in Germans.', *Int J Legal Med* 121(1), 24-7
- Junge A., Madea B. (1999), 'Population studies of the Y-chromosome specific polymorphisms DYS19, DYS389 I + II, DYS390 and DYS393 in a western German population (Bonn area).', *Forensic Sci Int* 101(3), 195-201
- Hohoff C., Dewa K., Sibbing U., Hoppe K., Forster P., Brinkmann B. (2007), 'Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany.', *Int J Legal Med* 121(5), 359-63
- Lessig R., Edelmann J. (1998), 'Y chromosome polymorphisms and haplotypes in west Saxony (Germany).', *Int J Legal Med* 111(4), 215-8
- Kayser M., Lao O., Anslinger K., Augustin C., Bargel G., Edelmann J., Elias S., Heinrich M., Henke J., Henke L., Hohoff C., Illing A., Jonkisz A., Kuzniar P., Lebioda A., Lessig R., Lewicki S., Maciejewska A., Monies DM., Pawłowski R., Poetsch M., Schmid D., Schmidt U., Schneider PM., Stradmann Bellinghausen B., Szibor R., Wegener R., Wozniak M., Zoledziewska M., Roewer L., Dobosz T., Ploski R. (2005), 'Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis.', *Hum Genet* 117(5), 428-43

- 186 Haplotypes from Greenland
  - Bosch E., Rosser ZH., Nørby S., Lynnerup N., Jobling MA. (2003), 'Y-chromosomal STR haplotypes in Inuit and Danish population samples.', *Forensic Sci Int* 132(3), 228-32
- 218 Haplotypes from Hungary
  - Pamjav H., Á F., Fehér T., Fóthi E. (2017), 'A study of the Bodrogköz population in north-eastern Hungary by Y chromosomal haplotypes and haplogroups.', *Mol Genet Genomics* 292(4), 883-894
  - Nagy M., Henke L., Henke J., Chatthopadhyay PK., Völgyi A., Zalán A., Peterman O., Bernasovská J., Pamjav H. (2007), 'Searching for the origin of Romanies: Slovakian Romani, Jats of Haryana and Jat Sikhs Y-STR data in comparison with different Romani populations.', *Forensic Sci Int* 169(1), 19-26
  - Füredi S., Woller J., Pádár Z., Angyal M. (1999), 'Y-STR haplotyping in two Hungarian populations.', *Int J Legal Med* 113(1), 38-42
  - Beer Z., Csete K., Varga T. (2004), 'Y-chromosome STR haplotype in Szekely population.', *Forensic Sci Int* 139(2-3), 155-8
  - Egyed B., Füredi S., Angyal M., Boutrand L., Vandenberghe A., Woller J., Pádár Z. (2000), 'Analysis of eight STR loci in two Hungarian populations.', *Int J Legal Med* 113(5), 272-5

- Völgyi A., Zalán A., Szvetnik E., Pamjav H. (2009), 'Hungarian population data for 11 Y-STR and 49 Y-SNP markers.', *Forensic Sci Int Genet* 3(2), e27-8

- 627 Haplotypes from Italy
  - Robino C., Ralf A., Pasino S., Marchi D., Ballantyne KN., Barbaro A., Bini C., Carnevali E., Casarino L., Gaetano D., Fabbri M., Ferri G., Giardina E., Gonzalez A., Matullo G., Nutini AL., Onofri V., Piccinini A., Piglionica M., Ponzano E., Previderè C., Resta N., Scarnicci F., Seidita G., Sorçaburu Cigliero S., Turrina S., Verzeletti A., Kayser M. (2015), 'Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise.', *Forensic Sci Int Genet* 15, 56-63
  - Rapone C., D'atanasio E., Agostino A., Mariano M., Papaluca MT., Cruciani F., Berti A. (2016), 'Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler(®) Plus kit) in an Italian population sample.', *Forensic Sci Int Genet* 21, e1-5
  - Lacerenza D., Aneli S., Di C., Critelli R., Piazza A., Matullo G., Culigioni C., Robledo R., Robino C., Calò C. (2017), 'Investigation of extended Y chromosome STR haplotypes in Sardinia', *Forensic Science International: Genetics* Epub ahead of Print, In press
  - Sarno S., Tofanelli S., De S., Quagliariello A., Bortolini E., Ferri G., Anagnostou P., Brisighelli F., Capelli C., Tagarelli G., Sineo L., Luiselli D., Boattini A., Pettener D. (2016), 'Shared language, diverging genetic histories: high-resolution analysis of Y-chromosome variability in Calabrian and Sicilian Arbereshe', *European Journal of Human Genetics* 24, 600-6
  - Grignani P., Peloso G., Fattorini P., Previderè C. (2000), 'Highly informative Y-chromosomal haplotypes by the addition of three new STRs DYS437, DYS438 and DYS439.', *Int J Legal Med* 114(1-2), 125-9
  - Presciuttini S., Caglià A., Alù M., Asmundo A., Buscemi L., Caenazzo L., Carnevali E., Carra E., Battisti D., Stefano D., Domenici R., Piccinini A., Resta N., Ricci U., Pascali VL. (2001), 'Y-chromosome haplotypes in Italy: the GEFI collaborative database.', *Forensic Sci Int*122(2-3), 184-8
  - Robino C., Inturri S., Gino S., Torre C., Gaetano D., Crobu F., Romano V., Matullo G., Piazza A. (2006), 'Y-chromosomal STR haplotypes in Sicily.', *Forensic Sci Int* 159(2-3), 235-40
  - Ferri G., Ceccardi S., Lugaresi F., Bini C., Ingravallo F., Cicognani A., Falconi M., Pelotti S. (2008), 'Male haplotypes and haplogroups differences between urban (Rimini) and rural area (Valmarecchia) in Romagna region (North Italy).', *Forensic Sci Int* 175(2-3), 250-5
  - Cerri N., Verzeletti A., Bandera B., Ferrari D. (2005), 'Population data for 12 Y-chromosome STRs in a sample from Brescia (northern Italy).', *Forensic Sci Int* 152(1), 83-7
  - Turrina S., Atzei R., Leo D. (2006), 'Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay.', *Int J Legal Med* 120(1), 56-9
  - Ghiani ME., Vona G. (2002), 'Y-chromosome-specific microsatellite variation in a population sample from Sardinia (Italy).', *Coll Antropol* 26(2), 387-401
  - Ferri G., Alù M., Corradini B., Radheshi E., Beduschi G. (2009), 'Slow and fast evolving markers typing in Modena males (North Italy).', *Forensic Sci Int Genet* 3(2), e31-3
  - Onofri V., Alessandrini F., Turchi C., Fraternale B., Buscemi L., Pesaresi M., Tagliabracci A. (2007), 'Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis.', *Int J Legal Med* 121(3), 234-7

- Verzeletti A., Cerri N., Gasparini F., Poglio A., Mazzeo E., Ferrari D. (2009), 'Population data for 15 autosomal STRs loci and 12 Y chromosome STRs loci in a population sample from the Sardinia island (Italy).', *Leg Med (Tokyo)* 11(1), 37-40
- Rodríguez V., Tomàs C., Sánchez JJ., Castro JA., Ramon MM., Barbaro A., Morling N., Picornell A. (2009), 'Genetic sub-structure in western Mediterranean populations revealed by 12 Y-chromosome STR loci.', *Int J Legal Med* 123(2), 137-41
- Brisighelli F., Blanco Verea A., Boschi I., Garagnani P., Pascali VL., Carracedo A., Capelli C., Salas A. (2012), 'Patterns of Y-STR variation in Italy.', *Forensic Sci Int Genet* 6(6), 834-9
- Piglionica M., Baldassarra SL., Giardina E., Stella A., D'ovidio FD., Frati P., Lenato GM., Resta N., Dell'erba A. (2013), 'Population data for 17 Y-chromosome STRs in a sample from Apulia (Southern Italy).', *Forensic Sci Int Genet* 7(1), e3-4

- 251 Haplotypes from Lithuania
  - Giedrė Ruzgaitė, Marija čaplinskienė , Rima Baranovienė , Jūratė Jankauskienė , Jolanta Kukienė , Savanevskytė K., Bunokienė (2015), 'Forensic application of Y-chromosomal STR analysis in Lithuanian population', *Biologija 61(2)* 61(2), 60-72
  - Jankauskiene J., Kukiene J., Ivanova V., Aleknaviciute G. (2017), 'Population data and forensic genetic evaluation with the Yfiler™ Plus PCR Amplification kit in the Lithuanian population', *Forensic Science International Genetics Supplement Series* 6, e606-607
  - Lessig R., Edelmann J. (2001), 'Population data of Y-chromosomal STRs in Lithuanian, Latvian and Estonian males.', *Forensic Sci Int* 120(3), 223-5
- 612 Haplotypes from Poland
  - Spólnicka M., Dąbrowska J., Szabłowska Gnap E., Pałeczka A., Jabłońska M., Zbieć Piekarska R., Pięta A., Boroń M., Konarzewska M., Kostrzewa G., Płoski R., Rogalla U., Woźniak M., Grzybowski T. (2017), 'Intra- and inter-population analysis of haplotype diversity in Yfiler® Plus system using a wide set of representative data from Polish population', *For Sci Int Genet* 28, e22–e25
  - Ploski R., Wozniak M., Pawlowski R., Monies DM., Branicki W., Kupiec T., Kloosterman A., Dobosz T., Bosch E., Nowak M., Lessig R., Jobling MA., Roewer L., Kayser M. (2002), 'Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis.', *Hum Genet* 110(6), 592-600
  - Pawłowski R., Dettlaff Kakol A. (2003), 'Population data of nine Y-chromosomal STR loci in northern Poland.', *Forensic Sci Int* 131(2-3), 209-13
  - Pepinski W., Niemcunowicz Janica A., Skawronska M., Koc Zorawska E., Janica J., Soltyszewski I. (2004), 'Y-chromosome STR haplotypes in a population sample of the Byelorussian minority living in the northeastern Poland.', *Forensic Sci Int* 140(1), 117-21
  - Pepinski W., Niemcunowicz Janica A., Skawronska M., Koc Zorawska E., Janica J., Soltyszewski I. (2004), 'Y-chromosome STR haplotypes and alleles in the population sample of Old Believers residing in the Northeastern Poland.', *Forensic Sci Int* 143(1), 65-8
  - Pepinski W., Skawronska M., Niemcunowicz Janica A., Ptaszynska Sarosiek I., Koc Zorawska E., Janica J., Berent JA. (2005), 'Population genetics of Y-chromosome STRs in a population sample of the Lithuanian minority residing in the northeastern Poland.', *Forensic Sci Int* 153(2-3), 264-8
  - Pepinski W., Niemcunowicz Janica A., Ptaszynska Sarosiek I., Skawronska M., Koc Zorawska E., Janica J., Soltyszewski I. (2004), 'Population genetics of Y-chromosome STRs in a population of Podlasie, Northeastern Poland.', *Forensic Sci Int* 144(1), 77-82

- Janica J., Pepinski W., Niemcunowicz Janica A., Skawronska M., Aleksandrowicz Bukin M., Ptaszynska Sarosiek I., Koc Zorawska E. (2005), 'Y-chromosome STR haplotypes and alleles in the ethnic group of Polish Tatars residing in the Northeastern Poland.', *Forensic Sci Int* 150(1), 91-5
- Rebała K., Szczerkowska Z. (2005), 'Polish population study on Y chromosome haplotypes defined by 18 STR loci.', *Int J Legal Med* 119(5), 303-5
- Wolańska Nowak P., Branicki W., Parys Proszek A., Kupiec T. (2009), 'A population data for 17 Y-chromosome STR loci in South Poland population sample--some DYS458.2 variants uncovered and sequenced.', *Forensic Sci Int Genet* 4(1), e43-4

- 314 Haplotypes from Russian Federation
  - Dudás E., Vágó Zalán A., Vándor A., Saypasheva A., Pomozi P., Pamjav H. (2019), 'Genetic history of Bashkirian Mari and Southern Mansi ethnic groups in the Ural region', *Molecular Genetics and Genomics* Epub ahead of print, In press
  - Lessig R., Edelmann J., Kleemann WJ., Kozhemyako V. (2006), 'Population data of Y-chromosomal STRs in Russian males of the Primorye region population.', *Forensic Sci Int* 159(1), 71-6
  - Roewer L., Krüger C., Willuweit S., Nagy M., Rodig H., Kokshunova L., Rothämel T., Kravchenko S., Jobling MA., Stoneking M., Nasidze I. (2007), 'Y-chromosomal STR haplotypes in Kalmyk population samples.', *Forensic Sci Int* 173(2-3), 204-9
  - Roewer L., Willuweit S., Krüger C., Nagy M., Rychkov S., Morozowa I., Naumova O., Schneider Y., Zhukova O., Stoneking M., Nasidze I. (2008), 'Analysis of Y chromosome STR haplotypes in the European part of Russia reveals high diversities but non-significant genetic distances between populations.', *Int J Legal Med* 122(3), 219-23
  - Pakendorf B., Novgorodov IN., Osakovskij VL., Danilova AP., Protod'jakonov AP., Stoneking M. (2006), 'Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts.', *Hum Genet* 120(3), 334-53
  - Woźniak M., Derenko M., Malyarchuk B., Dambueva I., Grzybowski T., Mišcicka Sliwka D. (2006), 'Allelic and haplotypic frequencies at 11 Y-STR loci in Buryats from South-East Siberia.', *Forensic Sci Int* 164(2-3), 271-5
  - Pakendorf B., Novgorodov IN., Osakovskij VL., Stoneking M. (2007), 'Mating patterns amongst Siberian reindeer herders: inferences from mtDNA and Y-chromosomal analyses.', *Am J Phys Anthropol* 133(3), 1013-27
  - Nasidze I., Schädlich H., Stoneking M. (2003), 'Haplotypes from the Caucasus, Turkey and Iran for nine Y-STR loci.', *Forensic Sci Int* 137(1), 85-93
  - Rosser ZH., Zerjal T., Hurles ME., Adojaan M., Alavantic D., Amorim A., Amos W., Armenteros M., Arroyo E., Barbujani G., Beckman G., Beckman L., Bertranpetit J., Bosch E., Bradley DG., Brede G., Cooper G., Côrte Real HB., Knijff D., Decorte R., Dubrova YE., Evgrafov O., Gilissen A., Glisic S., Gölge M., Hill EW., Jeziorowska A., Kalaydjieva L., Kayser M., Kivisild T., Kravchenko SA., Krumina A., Kucinskas V., Lavinha J., Livshits LA., Malaspina P., Maria S., Mc Elreavey K., Meitinger TA., Mikelsaar AV., Mitchell RJ., Nafa K., Nicholson J., Nørby S., Pandya A., Parik J., Patsalis PC., Pereira L., Peterlin B., Pielberg G., Prata MJ., Previderé C., Roewer L., Rootsi S., Rubinsztein DC., Saillard J., Santos FR., Stefanescu G., Sykes BC., Tolun A., Villems R., Tyler Smith C., Jobling MA. (2000), 'Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language.', *Am J Hum Genet* 67(6), 1526-43

- Thèves C., Balaresque P., Evdokimova LE., Timofeev IV., Alekseev AN., Sevin A., Crubézy E., Gibert M. (2010), 'Population genetics of 17 Y-chromosomal STR loci in Yakutia.', *Forensic Sci Int Genet* 4(5), e129-30
- Pimenoff VN., Comas D., Palo JU., Vershubsky G., Kozlov A., Sajantila A. (2008), 'Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers.', *Eur J Hum Genet* 16(10), 1254-64
- Trynova EG., Tsitovich TN., Vylegzhanina EY., Bandurenko NA., Parson W. (2011), 'Presentation of 17 Y-chromosomal STRs in the population of the Sverdlovsk region.', *Forensic Sci Int Genet* 5(3), e101-4

- 194 Haplotypes from Slovenia
  - Sterlinko H., Pajnic IZ., Balazic J., Komel R. (2001), 'Human Y-specific STR haplotypes in a Slovenian population sample.', *Forensic Sci Int* 120(3), 226-8
- 316 Haplotypes from Spain
  - García O., Yurrebaso I., Mancisidor ID., López S., Alonso S., Gusmão L. (2015), 'Data for 27 Y-chromosome STR loci in the Basque Country autochthonous population', *Forensic Science International: Genetics* Epub ahead of print, in press
  - Martinez Cadenas C., Blanco Verea A., Hernando B., George BJ., Brion M., Carracedo A., Salas A., Capelli C. (2016), 'The relationship between surname frequency and Y chromosome variation in Spain', *European Journal of Human Genetics* 24, 120-8
  - Saiz M., Jesus M., Antonio J., Carlos J., Javier L. (2019), 'Genetic structure in the paternal lineages of South East Spain revealed by the analysis of 17 Y-STRs', *Scientific Reports*9, 5234
  - Kayser M., Krawczak M., Excoffier L., Dieltjes P., Corach D., Pascali V., Gehrig C., Bernini LF., Jespersen J., Bakker E., Roewer L., Knijff D. (2001), 'An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations.', *Am J Hum Genet* 68(4), 990-1018
  - Martinez Jarreta B., Nievas P., Abecia E., Hinojal R., Budowle B. (2003), 'Haplotype distribution of nine Y-chromosome STR-loci in two northern Spanish populations (Asturias and Aragon).', *J Forensic Sci* 48(1), 204-5
  - Gené M., Borrego N., Xifró A., Piqué E., Moreno P., Huguet E. (1999), 'Haplotype frequencies of eight Y-chromosome STR loci in Barcelona (North-East Spain).', *Int J Legal Med* 112(6), 403-5
  - Martín P., García Hirschfeld J., García O., Gusmão L., García P., Albarrán C., Sancho M., Alonso A. (2004), 'A Spanish population study of 17 Y-chromosome STR loci.', *Forensic Sci Int* 139(2-3), 231-5
  - Pestoni C., Cal ML., Lareu MV., Rodríguez Calvo MS., Carracedo A. (1999), 'Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain).', *Int J Legal Med* 112(1), 15-21
  - Zarrabeitia MT., Riancho JA., Sánchez Diz P., Sánchez Velasco P. (2001), '7-Locus Y chromosome haplotype profiling in a northern Spain population.', *Forensic Sci Int*123(1), 78-80
  - Gamero JJ., Romero JL., González JL., Carvalho M., Anjos MJ., Real FC., Vide MC. (2002), 'Y-chromosome STR haplotypes in a southwest Spain population sample.', *Forensic Sci Int* 125(1), 86-9

- Hurles ME., Veitia R., Arroyo E., Armenteros M., Bertranpetit J., Pérez Lezaun A., Bosch E., Shlumukova M., Cambon Thomsen A., Mc Elreavey K., De L., Röhl A., Wilson IJ., Singh L., Pandya A., Santos FR., Tyler Smith C., Jobling MA. (1999), 'Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism.', *Am J Hum Genet* 65(5), 1437-48
- Tomàs C., Jiménez G., Picornell A., Castro JA., Ramon MM. (2006), 'Differential maternal and paternal contributions to the genetic pool of Ibiza Island, Balearic Archipelago.', *Am J Phys Anthropol* 129(2), 268-78
- Ambrosio B., Novelletto A., Hernandez C., Dugoujon JM., Fortes Lima C., Rodriguez JN., Calderon R. (2012), 'Y-STR genetic diversity in autochthonous Andalusians from Huelva and Granada provinces (Spain).', *Forensic Sci Int Genet* 6(2), e66-71
- Valverde L., Rosique M., Köhnemann S., Cardoso S., García A., Odriozola A., Aznar JM., Celorrio D., Schuerenkamp M., Zubizarreta J., Davis MC., Hampikian G., Pfeiffer H., Pancorbo D. (2012), 'Y-STR variation in the Basque diaspora in the Western USA: evolutionary and forensic perspectives.', *Int J Legal Med* 126(2), 293-8
- Gaibar M., Esteban E., Moral P., Gómez Gallego F., Santiago C., Bandrés F., Luna F., Fernández Santander A. (2010), 'STR genetic diversity in a Mediterranean population from the south of the Iberian Peninsula.', *Ann Hum Biol* 37(2), 253-66
- Valverde L., Köhnemann S., Rosique M., Cardoso S., Zarrabeitia M., Pfeiffer H., Pancorbo D. (2012), '17 Y-STR haplotype data for a population sample of Residents in the Basque Country.', *Forensic Sci Int Genet* 6(4), e109-11
- Aler M., Salas A., Sánchez Diz P., Murcia E., Carracedo A., Gisbert M. (2001), 'Y-chromosome STR haplotypes from a Western Mediterranean population sample.', *Forensic Sci Int* 119(2), 254-7

- 143 Haplotypes from Switzerland
  - Haas C., Wangensteen T., Giezendanner N., Kratzer A., Bär W. (2006), 'Y-chromosome STR haplotypes in a population sample from Switzerland (Zurich area).', *Forensic Sci Int*158(2-3), 213-8
  - Gehrig C., Hochmeister M., Budowle B. (2000), 'Swiss allele frequencies and haplotypes of 7 Y-specific STRs.', *J Forensic Sci* 45(2), 436-9

Appendix 2

Table 1-24 – Sequence-based allele variants obtained using the ForenSeq kit to sequence the Norwegian sample set obtained in this study (n=286). If no sequence variants are found in the flanking regions, or non-repeat fractions of the repeat unit the sequences are listed under the table. Totally, 15 YSTR markers contain additional sequence-based allele variants, nine in the repeat unit only, four in the flanking region only, and two in both the repeat unit and flanking region. DYS460 and DYS488 cut the downstream flanking region unevenly. DYS389II is missing the downstream flanking region, and DYS570, DYS219, and DYS392 is missing the upstream flanking region. Novel sequence variants are emphasized using a bold asterisk. Singe base mutations in the repeat unit and flanking regions are red and underlined.

DYS392 Allele variants	n	Repeat Unit
11	179	(TAT)11
12	12	(TAT)12
13	72	(TAT)13
14	23	(TAT)14
17	1	(TAT)17

Downstream flanking region: TTACTAAGGAATGGGATTGGTAGGTTTAA

DYS391 Allele variants	n	Repeat Unit
9	1	(TCTA)9
10	159	(TCTA)10
11	118	(TCTA)11
12	7	(TCTA)12
13	1	(TCTA)13

Upstream flanking region: ATATCTGTCTGTCTG

Downstream flanking region: TCTGCCTATCTGCCTGCCTACCTATCCCTCTAT

DYS448 Allele variants	n	Repeat Unit	Flanking Region
17	2	(AGAGAT)11 N42 (AGAGAT)6	AGAGAGGTAAAGATAGAGATAAATTTCCAGACCGGC
18	4	(AGAGAT)11 N42 (AGAGAT)7	AGAGAGGTAAAGATAGAGATAAATTTCCAG
19	109	(AGAGAT)11 N42 (AGAGAT)8	AGAGAGGTAAAGATAGAGATAAAT
	1	(AGAGAT)12 N42 (AGAGAT)7	AGAGAGGTAAAGATAGAGATAAAT
20	156	(AGAGAT)12 N42 (AGAGAT)8	AGAGAGGTAAAGATAGAG
21	14	(AGAGAT)13 N42 (AGAGAT)8	AGAGAGGTAAAG

Upstream flanking region: GAGATAGAGACATGGATAA

N42: ATAGAGATAGAGAGATAGAGATAGAGATAGATAGATAGAGAA

DYS533 Allele variants	n	Repeat Unit
8	2	(TATC)8
9	2	(TATC)9
10	5	(TATC)10
11	119	(TATC)11
12	140	(TATC)12
13	17	(TATC)13
14	1	(TATC)14

Upstream flanking region: TAACTATATAACTATGTATTATCTATCAATCTTCTACCTATCATCTTTCTAGCTAGCTATCATC

Downstream flanking region: ATCTATCATCTTCTATTGTTT

DYS460 Allele variants	n	Repeat Unit	Flanking Region
9	1	(ATAG)9	ATAATAGACAAATACATAATAAATGATAGGCAGAGGATAGATGATATGGATAGACAGATATATCTAATAGGTAGATGATA GATAATAGGTAGATAGAAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA
10	109	(ATAG)10	ATAATAGACAAATACATAATAAATGATAGGCAGAGGATAGATGATATGGATAGACAGATATATCTAATAGGTAGATGATA GATAATAGGTAGATAGAAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA
	1*	(ATAG)10	ATAATAGACAAATACATAATAAATGATAGGCAGAGGATAGATGATATGGATAGACAGATATATCTAATAGGTAGATGATA GATAATAGGTAGATAGAAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACA
11	153	(ATAG)11	ATAATAGACAAATACATAATAAATGATAGGCAGAGGATAGATGATATGGATAGACAGATATATCTAATAGGTAGATGATA GATAATAGGTAGATAGAAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA
12	22	(ATAG)12	ATAATAGACAAATACATAATAAATGATAGGCAGAGGATAGATGATATGGATAGACAGATATATCTAATAGGTAGATGATA GATAATAGGTAGATAGAAGATAGGTAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA

Upstream flanking region: GTCAAGACAGTAGCAAGCACAAGAATACCAGAGGAATCTGACACCTCTGAC

DYS635 Allele variants	n	Repeat Unit
20	2	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)10
21	58	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)11
22	42	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)12
	6	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)8
23	148	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)9
	9	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)13
24	14	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)10

	5	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)14
25	2	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)11
	1	(TCTA)4 (TGTA)2 (TCTA)2 (TGTA)2 (TCTA)15

Upstream flanking region: TGGCTTCTCACTTTGCATAGAATC

Downstream flanking region: TCACATTTTCTTTATCCATTCATTGATTGATGGATATTTGGGCTGGTTCCACATTTTTCCAACTGTGAATTTTGCTGCTATAAACAATGTATGTGCAAGTGTCTT

DYS437 Allele variants	n	Repeat Unit
14	123	(TCTA)8 (TCTG)2 (TCTA)4
15	76	(TCTA)9 (TCTG)2 (TCTA)4
16	84	(TCTA)10 (TCTG)2 (TCTA)4
	1*	(TCTA)9 (TCTG)3 (TCTA)4
17	2	(TCTA)11 (TCTG)2 (TCTA)4

Upstream flanking region: ATGCCCATCCGG

Downstream flanking region: TCATCTATCATCTGTGAATGATGTCTATCTACTTATCTATGAATGATATTTATCTGTGGTTATCTATCTATCTATATCATCTGTGAATGACAGGGTCTTCCTCTG

DYS389I Allele variants	n	Repeat Unit
12	87	(TCTG)3 (TCTA)9
	3	(TCTG)2 (TCTA)10
13	129	(TCTG)3 (TCTA)10
	1*	(TCTG)2 (TCTA)11
14	65	(TCTG)3 (TCTA)11
15	1	(TCTG)3 (TCTA)12

Upstream flanking region: ATCTGTATTATCTATGTA

Downstream flanking region: TCCCTCCCTCTATCAATCTATCTATTTATCTAGCAGTCCATCATCTATCTATGACATTCTT

DYS19 Allele variants	n	Repeat Unit
13	8	(TAGA)2 TAGG (TAGA)10
14	144	(TAGA)2 TAGG (TAGA)11
15	85	(TAGA)2 TAGG (TAGA)12
16	45	(TAGA)2 TAGG (TAGA)13
17	4	(TAGA)2 TAGG (TAGA)14

Downstream flanking region:
TATAGTGACACTCTCCTTAACCCAGATGGACTCCTTGTCCTCACTACATGGCCATGGCCCGAAGTATTACTCCTGGTGCCCCAGCCACTATTTCCAGGTGCAGAGATTGACCA

Y-GATA-H4 Allele variants	n	Repeat Unit	Flanking Region
10	5	(TAGA)10	ATGGATAGATTAGATGGATGAATAGATAGATAGATAGATACATAGATAG
11	96	(TAGA)11	ATGGATAGATTAGATGGATGAATAGATAGATAGATAGATACATAGATAG
12	134	(TAGA)12	ATGGATAGATTAGATGGATGAATAGATAGATAGATAGATACATAGATAG
	1*	(TAGA)12	ATGGATAGATTAGATGGAAGAATAGATAGATAGATAGATACATAGATAG
13	44	(TAGA)13	ATGGATAGATTAGATGGATGAATAGATAGATAGATAGATACATAGATAG
14	6	(TAGA)14	ATGGATAGATTAGATGGATGAATAGATAGATAGATAGATACATAGATAG

Upstream flanking region: AGATAGATAGATAGATCTATAGATAGATAGGTAGGTAGG

DYS439 Allele variants	n	Repeat Unit
10	80	(GATA)10
11	126	(GATA)11
12	62	(GATA)12
13	17	(GATA)13
14	1	(GATA)14

Upstream flanking region: GATAGATATACAGATAGATAGATACATAGGTGGAGACAGATAGATGATAAATAGAA
Downstream flanking region: GAAAGTATAAGTAAAGAGATGAT

DYS438 Allele variants	n	Repeat Unit
9	11	(TTTTC)9
10	104	(TTTTC)10
11	91	(TTTTC)11
12	76	(TTTTC)12
13	4	(TTTTC)13

Upstream flanking region: GGTAAACAGTATA
Downstream flanking region: TATTTGAAATGGAGTTTCACTCTTGTTGCCCAGGCT

DYS389II Allele variants	n	Repeat Unit
25	1	(TCTG)5 (TCTA)8 N48 (TCTG)3 (TCTA)9

27	1	(TCTG)5 (TCTA)10 N48 (TCTG)3 (TCTA)9
28	72	(TCTG)5 (TCTA)11 N48 (TCTG)3 (TCTA)9
	5	(TCTG)5 (TCTA)10 N48 (TCTG)3 (TCTA)10
	2*	(TCTG)5 (TCTA)11 N48 (TCTG)2 (TCTA)10
	1	(TCTG)4 (TCTA)11 N48 (TCTG)3 (TCTA)10
29	81	(TCTG)5 (TCTA)11 N48 (TCTG)3 (TCTA)10
	13	(TCTG)5 (TCTA)12 N48 (TCTG)3 (TCTA)9
	4	(TCTG)4 (TCTA)12 N48 (TCTG)3 (TCTA)10
	1*	(TCTG)5 (TCTA)11 N48 (TCTG)2 (TCTA)11
	1*	(TCTG)5 (TCTA)12 N48 (TCTG)2 (TCTA)10
30	35	(TCTG)5 (TCTA)12 N48 (TCTG)3 (TCTA)10
	13	(TCTG)4 (TCTA)12 N48 (TCTG)3 (TCTA)11
	19	(TCTG)5 (TCTA)11 N48 (TCTG)3 (TCTA)11
31	25	(TCTG)5 (TCTA)12 N48 (TCTG)3 (TCTA)11
	2	(TCTG)5 (TCTA)13 N48 (TCTG)3 (TCTA)10
	1	(TCTG)6 (TCTA)12 N48 (TCTG)3 (TCTA)10
	1	(TCTG)4 (TCTA)13 N48 (TCTG)3 (TCTA)11
	1*	(TCTG)4 (TCTA)12 N48 (TCTG)3 (TCTA)12
32	4	(TCTG)5 (TCTA)13 N48 (TCTG)3 (TCTA)11
	1	(TCTG)6 (TCTA)12 N48 (TCTG)3 (TCTA)11
33	2	(TCTG)5 (TCTA)14 N48 (TCTG)3 (TCTA)11

Upstream flanking region: ATCTGTATTATCTATGTGTGTG

N48: TCATTATACCTACTTCTGTATCCAACTCTCATCTGTATTATCTATGTA

DYS576 Allele variants	n	Repeat Unit
14	3	(AAAG)14
15	10	(AAAG)15
16	60	(AAAG)16
17	100	(AAAG)17
18	70	(AAAG)18
19	32	(AAAG)19
20	10	(AAAG)20

21	1	(AAAG)21

Upstream flanking region: TCAGCCAAGCAACATAGCAAGACCTCATCTCTGAATA

Downstream flanking region: AAAAAGCCAAGACAAATACGCTTATTACTCCCATCTCCT

DYS390 Allele variants	n	Flanking Region	Repeat Unit
22	46	TCTATCTA	(TCTG)8 (TCTA)9 TCTG (TCTA)4
	2*	TCTATCTA	(TCTG)8 (TCTA)10 TCTG (TCTA)3
23	76	TCTATCTA	(TCTG)8 (TCTA)10 TCTG (TCTA)4
	3	TCTATCTA	(TCTG)9 (TCTA)9 TCTG (TCTA)4
	2	TCTATCTA	(TCTG)7 (TCTA)11 TCTG (TCTA)4
23	67	TCTATCTA	(TCTG)8 (TCTA)11 TCTG (TCTA)4
	3	TCTCTCTA	(TCTG)8 (TCTA)11 TCTG (TCTA)4
	1	TCTATCTA	(TCTG)7 (TCTA)12 TCTG (TCTA)4
25	75	TCTATCTA	(TCTG)8 (TCTA)12 TCTG (TCTA)4
	3*	TCTCTCTA	(TCTG)8 (TCTA)12 TCTG (TCTA)4
	4	TCTATCTA	(TCTG)9 (TCTA)11 TCTG (TCTA)4
26	4	TCTATCTA	(TCTG)8 (TCTA)13 TCTG (TCTA)4

Downstream flanking region: TCATCTATCTATCTTTCCTTGTTTCTGAGTATACACATTGCAATGTTTTCATTTTACTGTCATCCATTTCT

DYS570 Allele variants	n	Repeat Unit
16	27	(TTTC)16
17	67	(TTTC)17
18	79	(TTTC)18
19	55	(TTTC)19
20	35	(TTTC)20
	1	TTCC (TTTC)20
21	14	(TTTC)21
22	8	(TTTC)22

Downstream flanking region: TTTTTGTAGATAGG

DYF387S1 Allele variants	n	Repeat Unit
34	3	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)12
	2	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)13
	1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)7 (AAAG)14
	1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)11
35	41	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)12
	19	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)13
	1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)14
	1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)7 (AAAG)15
	1*	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)11
36	60	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)13
	43	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)14
	5	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)12
	2	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)7 (AAAG)16
37	113	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)15
	56	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)14
	13	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)13
	1	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)8 (AAAG)16
	1*	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG AAAG (GAAG)12 (AAAG)13
38	35	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)15
	48	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)14
	37	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)16
	2	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)13
39	9	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)15
	9	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)10 (AAAG)16
	19	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)14
	14	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)17
40	5	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)12 (AAAG)15
	2	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)9 (AAAG)18
	2	(AAAG)3 GTGA (GAAG)4 (AAAG)2 GAAG (AAAG)2 (GAAG)11 (AAAG)16

Upstream flanking region: GAAGAAAGAGAAAA

Downstream flanking region: AAAATAAAAAAA

DYS385a/b Allele variant	n	Flanking Region	Repeat Unit	Flanking Region
10	7	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)10	GAGAAAAAGAAAGGA
	1*	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGGAAGGAAGGAAGGAAGGAAGGGAAA	(GAAA)10	GAGAAAAAGAAAGGA
11	169	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)11	GAGAAAAAGAAAGGA
12	8	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)12	GAGAAAAAGAAAGGA
13	67	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)13	GAGAAAAAGAAAGGA
	2*	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGGAAA	(GAAA)13	GAGAAAAAGAAAGGA
14	211	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)14	GAGAAAAAGAAAGGA
	2*	AAAGAAAAGAAATAAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)14	GAGAAAAAGAAAGGA
14.2	1*	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)14 GA	AAGAAAAAGAAAGGA
15	54	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)15	GAGAAAAAGAAAGGA
16	10	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)16	GAGAAAAAGAAAGGA
17	3	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)17	GAGAAAAAGAAAGGA
18	2	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)18	GAGAAAAAGAAAGGA
19	2	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)19	GAGAAAAAGAAAGGA
20	1	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)20	GAGAAAAAGAAAGGA
21	1	AAAGAAAAGAAATGAAATTCAGAAAGGAAGGAAGGAAGGAGAAAGAAAGTAAAAAAGAAAGAAAGAGAAAAA GAGAAAAAGAAAGAAAGAGAAGAAAGAGAAAGAGGAAAGAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGG	(GAAA)21	GAGAAAAAGAAAGGA

DYS481 Allele variants	n	Repeat Unit
20	19	(CTT)20
21	10	(CTT)21
	2	CTG (CTT)20
22	69	(CTT)22
	1	CTG (CTT)21
23	58	(CTT)23
24	19	(CTT)24
25	71	(CTT)25
26	22	(CTT)26
27	10	(CTT)27
28	5	(CTT)28

Upstream flanking region: TGGCTAACGCTGTTCAGCATGCTG

Downstream flanking region: TTTTGA

DYS549 Allele variants	n	Flanking Region	Repeat Unit
10	1	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)10
11	29	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)11
12	180	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)12
	1*	TAATAAGGTAGACATAGCAATTAGGTAGGTGAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)12
13	63	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)13
14	10	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)14
15	2	TAATAAGGTAGACATAGCAATTAGGTAGGTAAAGAGGAAGATGATAGATGATTAGAAAGAT	(GATA)15

Downstream flanking region: GAAAAAATCTACATAAACAAAATCACAAATGGAAAAGGGGACATTACCA

DYS505 Allele variants	n	Repeat Unit
9	1	(TCCT)9
10	2	(TCCT)10
11	110	(TCCT)11
12	136	(TCCT)12
13	35	(TCCT)13
14	2	(TCCT)14

Upstream flanking region: GTTTACTTTCTTTCTCTTTTTCTCTTTTTTTCTTTATTTTTCTTTCTCTGTTCTTTTTCTC
Downstream flanking region: TCTTTCCCTCCTTCTTTCTCTTTA

DYS643 Allele variants	n	Repeat Unit
9	2	(CTTTT)9
10	152	(CTTTT)10
11	36	(CTTTT)11
12	85	(CTTTT)12
13	11	(CTTTT)13

Upstream flanking region: TGATTTTTGCAGGTGTTCACTGCAAGCCATGCCTGGTTAAACTACTGTGC
Downstream flanking region: CTTTCTTTTTAAAACTT

DYS522 Allele variants	n	Repeat Unit
10	139	(GATA)10
11	86	(GATA)11
12	37	(GATA)12
	1	(GATA)11 GACA
13	12	(GATA)13
14	11	(GATA)14

Upstream flanking region: ATAGAT
Downstream flanking region: GACAGATGTCCACCATGAGGTTC

DYS612 Allele variants	n = 286	Repeat Unit	Flanking Region
24	1	(TCT)4 CCT (TCT)19	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
26	2	(TCT)4 CCT (TCT)21	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
27	12	(TCT)4 CCT (TCT)22	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
28	17	(TCT)4 CCT (TCT)23	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
29	36	(TCT)4 CCT (TCT)24	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC

X

	1*	(TCT)4 CCT TCT CCT (TCT)22	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
30	83	(TCT)4 CCT (TCT)25	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
	1*	(TCT)4 CCT TCT CCT (TCT)23	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
31	79	(TCT)4 CCT (TCT)26	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
32	31	(TCT)4 CCT (TCT)27	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
33	11	(TCT)4 CCT (TCT)28	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
	4*	(TCT)4 CCT (TCT)28	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAAAAAAA
	1	(TCT)4 CCT (TCT)15 CCT (TCT)12	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAAAAAAA
34	7	(TCT)4 CCT (TCT)29	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC
35	1	(TCT)4 CCT (TCT)30	GTCACTTTTCCAAATTATTTTCTTTTGCCTTCCCTCAGTTCCCTTTTTGGCTCTAGATACCCATGG CAAGTGTCAAGCAATGCACAAGAATGAACAAAAGGC

Upstream flanking region: TTTCACACAGGTTCAGAGGTTTGCCTCCTCCTCCTCCTCTT