Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS

Faculty of Chemistry, Biotechnology and Food Science (KBM)

Massively parallel sequencing of extended SNP panels with applications in relationship inference

Kristine Haugland Larsen

Biotechnology - Genetics

Acknowledgment

This thesis represents the completion of my master's degree in Biotechnology at The Norwegian University of Life Sciences (NMBU). The project was conducted at Oslo University Hospital (OUS), Department of Forensic Sciences – Section of Forensic Genetic Kinship and Identity (REFA) in the period August 2018 to May 2019.

First and foremost, I would like to thank my engaged supervisors at REFA, Daniel Kling and Peter Jørgen Tønnessen Haddeland, for invaluable counselling and thorough feedback. I am incredibly grateful for all their support and advices, and for always being available throughout the entire project. Furthermore, I would like to thank the rest of the team at REFA for being very inclusive and helpful with answering questions.

I would also like to thank my supervisor at NMBU, Professor Thore Egeland, for first introducing me to the forensic genetic field, and for valuable guidance during the project.

Finally, I would like to thank my dear family, friends and boyfriend for support and encouragement, it has meant a lot to me. Especially, I would like to thank Sofie for helpful discussions, support and the best company throughout the entire education.

Oslo, May 2019

Kristine Haugland Larsen

Abstract

Short tandem repeat (STR) markers are the current golden standard in forensic genetics, whereas single nucleotide polymorphisms (SNPs) have emerged as an alternative. In this thesis it was investigated if using SNP panels as a supplement to STR panels could lead to more conclusive results in complex kinship cases. 16 samples from eight complex kinship cases with inconclusive results were analysed with two STR panels and two supplementary SNP panels. The results were evaluated by comparing the likelihood ratio (LR) calculated based on the two STR panels, with the LR calculated based on the combination of both STR- and SNP panels. The STR analyses were performed with traditional capillary electrophoresis, while the SNPs were analysed with massively parallel sequencing using the Ion Torrent™ Personal Genome Machine™ (PGM™) System (Thermo Fisher Scientific). Sequencing on the PGM™ is not a part of the accredited routine at REFA, and this method was thoroughly evaluated and reviewed in this project.

Several software and tools were used in the evaluations in the project. For instance, the statistical software Familias was used to calculate LR for the real cases and the simulations, and to perform a blind search. Results from both the real cases and the simulations showed a notable decrease in the number of inconclusive cases when SNPs were included in the analyses. However, throughout this project it has been revealed that there are several important aspects that can affect the final conclusion in complex kinship cases, especially when a large number of markers are applied. These have been reviewed in terms of further work with constructing a SNP panel that can be used in routine work, and it was suggested that the markers should be ancestry-insensitive, not in linkage disequilibrium with each other, and that linkage should be calculated and included in the analyses. It was also suggested that larger- or better suited panels were needed to solve cases where the proposed relationship was half sibling of parent or equivalent. Additionally, it was shown in this project that the use of the correct allele frequency databases in the calculations was crucial, especially if ancestry-sensitive markers were applied.

Finally, it was concluded that SNPs are well suited as supplement to STRs in complex kinship cases, but that further investigations should be performed in respect to construct a panel and a procedure that is suitable for routine work.

Sammendrag

I dag benyttes mikrosatellitter (STRer) som standardmarkører i rettsmedisinske genetikkanalyser. Enkeltnukleotidpolymorfismer (SNPer) har imidlertid kommet inn som alternative markører. I denne studien har det blitt undersøkt om flere konkluderende resultater kan oppnås i komplekse slektskapssaker ved å benytte SNP-paneler som supplement til standard STR-paneler. 16 prøver fra åtte komplekse slektskapsaker ble benyttet i undersøkelsene, og ble analysert med to STR paneler og to supplementerende SNP paneler. Resultatene ble vurdert ved at bevisvekt (LR) basert på de to STR-panelene ble sammenlignet med LR basert på både STR- og SNP-paneler. STR-analysene ble utført med tradisjonell kapillærelektroforese, mens SNPene ble analysert med massiv parallell sekvensering på Personal Genome Machine™ (PGM™) (Thermo Fisher Scientific). Sekvensering på PGM™ er ikke en del av den akkrediterte rutinen på REFA, og metoden ble derfor grundig gjennomgått og evaluert i dette prosjektet.

Flere verktøy og programvarer ble benyttet i evalueringene i dette prosjektet. Blant annet ble den statistiske programvaren Familias benyttet til å utføre LR-beregninger for de reelle sakene og for simuleringer, samt til å utføre et blindsøk. Resultatene fra både de reelle sakene og simuleringene viste at når SNPer ble inkludert i analysen så man en merkbar nedgang i saker hvor det ikke kunne konkluderes. I løpet av prosjektet har det imidlertid blitt avdekket flere viktige aspekter som kan påvirke den endelige konklusjonen i komplekse slektskapssaker, spesielt når det benyttes et stort antall markører. Disse aspektene har blitt gjennomgått med tanke på videre utarbeiding av et SNP-panel som kan brukes i rutinearbeid. Det ble foreslått at inkluderte markører ikke bør være sensitive for geografisk avstamning, ikke i koblingsulikevekt med hverandre, samt at genetisk kobling bør beregnes for markørene og inkluderes i analysen. Det ble også foreslått at flere markører bør inkluderes for å oppnå en konklusjon i saker hvor slektskapet dreier seg om et halvsøsken av en forelder eller tilsvarende. I tillegg kommer det frem at bruk av riktig allelfrekvensdatabase i beregningene er avgjørende, spesielt dersom de inkluderte markørene er sensitive for geografisk avstamning.

Det ble til slutt konkludert med at SNPer er velegnet som supplement til STRer i komplekse slektskapssaker, men at videre undersøkelser bør gjennomføres med tanke på å konstruere et panel og en prosedyre som er egnet for bruk i rutinearbeid.

Abbreviations

A	Adenin
AF	Alleged father
AMEL	Amelogenin
AMELX	Amelogenin, X chromosome
AMELY	Amelogenin, Y chromosome
Ancestry	Precision ID Ancestry Panel
bp	Base pair
C	Cytosine
CE	Capillary electrophoresis
CH	Child
cM	CentiMorgan
DNA	Deoxyribonucleic acid
emPCR	Emulsion polymerase chain reaction
Fusion 6C	PowerPlex® Fusion 6C System
G	Guanine
H_1	Hypothesis 1 (main hypothesis)
H_2	Hypothesis 2 (alternative hypothesis)
H+	Proton
HDplex	Investigator HDplex Kit
HWE	Hardy-Weinberg equilibrium
IBD	Identical by decent
IBS	Identical by state
Identity	Precision ID Identity Panel
LD	Linkage disequilibrium
ILS	Internal lane standard
kV	Kilovolt (1000 volt)
LR	Likelihood ratio
MAF	Major allele frequency
MPS	Massively parallel sequencing
PCR	Polymerase chain reaction
PGM™	Personal Genome Machine™
psi	Pounds per square inch
REFA	Oslo University Hospital, Department of Forensic Sciences – Section of Forensic Genetic Kinship and Identity
rfu	Relative fluorescence units
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
T	Thymine
µl	Microliter (0.001 milliliter)

Table of contents

1 Introduction

Any two humans share more than 99.9% of their DNA, even still, unique genetic variants are found in all individuals (Venter et al. 2001). These variants are short, hypervariable regions and are referred to as genetic markers when their chromosomal positions are known. Within forensic DNA analyses, short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) are markers that are frequently used to construct DNA profiles. These profiles can be used in important forensic areas, e.g. cases regarding missing persons, identification of unknown dead bodies, immigration, establishment of paternity or in criminal cases. In such cases, the DNA profiles of the involved persons are compared to a profile from an alleged relative, and it is preferable to obtain samples from close relatives, for instance a parent or a child. However, this is not always possible, and testing of more distant relationships may become relevant. This can complicate the kinship analyses and possibly lead to inconclusive results, which occurs when the probabilities after the analysis do not point in any specific direction - neither for nor against the alleged relationship.

This project has been conducted at Oslo University Hospital, Department of Forensic Sciences – Section of Forensic Genetic Kinship and Identity (henceforth abbreviated REFA). Kinship analyses, mainly paternity, represent a large amount of the cases at REFA. Additionally, the section performs DNA analyses for body identification, measurements of donor chimerism for bone marrow transplanted patients and construction of DNA profiles for the police. Standard procedure at REFA is to analyse all samples with one STR panel, and then supplement with other STR panels in complex cases. Several factors can complicate DNA analyses, e.g. inbreeding, mutations or as in this project – cases including relationships more distant than parent-child or full siblings. In this project, 16 samples from eight complex kinship cases were collected, whereof seven had been reported as inconclusive due to likelihood ratios (LR) between 0.5 and 120. The last case had an LR of around 1000. All samples in this project have been analysed with two STR panels and two supplementary SNP panels, with the purpose of investigating the effect of the SNP panels. This was assessed by comparing the LR calculated based on the two STR panels, with the LR calculated based on the combination of the STR and SNP panels.

The STR analyses were performed by capillary electrophoresis (CE) using the 3500xl Genetic Analyzer (Thermo Fisher Scientific), while the SNPs were analysed with Massively Parallel

Sequencing (MPS) using the Ion Torrent™ Personal Genome Machine™ (PGM™) System (Thermo Fisher Scientific). Unlike the traditional CE approach, SNP sequencing on the PGM™ is not a part of the accredited routine at REFA, and the section does not have a procedure for this method (other than the manufacturers manual). For this reason, the Ion Torrent™ method will generally be more thoroughly evaluated and reviewed in this project.

1.1 The aim of the thesis

SNP based methods have previously been considered in terms of replacing the classical STR approach, but has been demonstrated to currently lead to more inconclusive cases by e.g. Amorim and Pereira (2005). Others, for instance Gill et al. (2004) and Butler et al. (2007), have suggested to use SNPs as supplement to STRs, an approach that has been practiced in relationship case work in Copenhagen for more than a decade (Sanchez et al. 2006; van der Heijden et al. 2017). The main aim of this thesis was to investigate if using SNP panels as a supplement to standard STR panels could lead to more conclusive results in complex kinship cases. Furthermore, the supplementary SNPs could potentially be applied in difficult cases of several areas and contribute in solving more cases than what can be done with only STRs. These cases can for instance include family reunification or identification of dead, where a conclusion due to the DNA analysis may lead to a crucial answer for family members or others concerned by the case. The results will be evaluated based on LRs calculated for both real cases and simulations. Moreover, the applied SNP markers and the Ion Torrent™ method will be evaluated in respect to further work with constructing a procedure that can be used in a routine laboratory. The evaluation of the SNP markers will be based on linkage disequilibrium (LD), linkage and ancestry-sensitivity, and the method will be evaluated in terms of how it performs for the different relationships in the project cases. Ultimately, this will lead to a final conclusion regarding the general effect of SNPs as supplement to STRs in complex kinship cases.

1.2 Genetic markers

Traditional DNA analyses are based on comparing small areas of the genome, since whole genome sequencing generally is expensive, time consuming and less informative. The short, informative areas are known as genetic markers, and generally become useful when several are analysed simultaneously, generating a DNA profile. In traditional identification- and

kinship analyses, genetic markers found in non-coding areas of the genome have been preferred. A reason for this is that possession of gene data including phenotypic information like disease status raises several ethical issues (Samuel and Prainsack 2018). An example could be a genetic variant connected to a serious illness being discovered for a person involved in a kinship case.

As previously mentioned, about 99.9% of the human DNA are identical for all human individuals, i.e. genetic markers must be present in a polymorphic area to be informative. A genetic position that have at least two variants (alleles), and for which the less common one is present in at least 1% of the population, is categorised as a genetic polymorphism (Satya et al. 2011; Karki et al. 2015). Common for polymorphic areas is that a mutation or change has occurred at some point in history and subsequently spread (Karki et al. 2015). Mutations can happen anywhere in the DNA sequence, and the probability of occurrence of a germline mutation, which can be passed on to offspring, is referred to as the mutation rate for a position. Mutations are necessary for the polymorphism in a population, but can also cause problems in DNA analyses, especially in analyses concerning parent-child testing. For instance, 50% of our DNA is inherited from each parent, and it is expected that a parent and a child share, at least, one allele for each genetic marker. If a mutation occurs in a relevant area in a parent's germ cell, this can result in a mismatch in the DNA sequence between the parent and the child. Further, if the possibility of mutations is not taken into account, this can lead to a false exclusion of the parent. Other relatives are not expected to match in all markers, and thus, one mutation will not make such a dramatic impact on calculations in such cases (Egeland et al. 2015, Chapter 2).

1.2.1 Short tandem repeats

Short tandem repeats (STRs) are repeated DNA sequences of two to six nucleotides and are also referred to as microsatellites. The number of repeats for a given STR differ from person to person (Figure 1) and is inherited from parent to child. STRs are found spread in the genome, most commonly in non-coding areas of the DNA (Fan and Chu 2007).

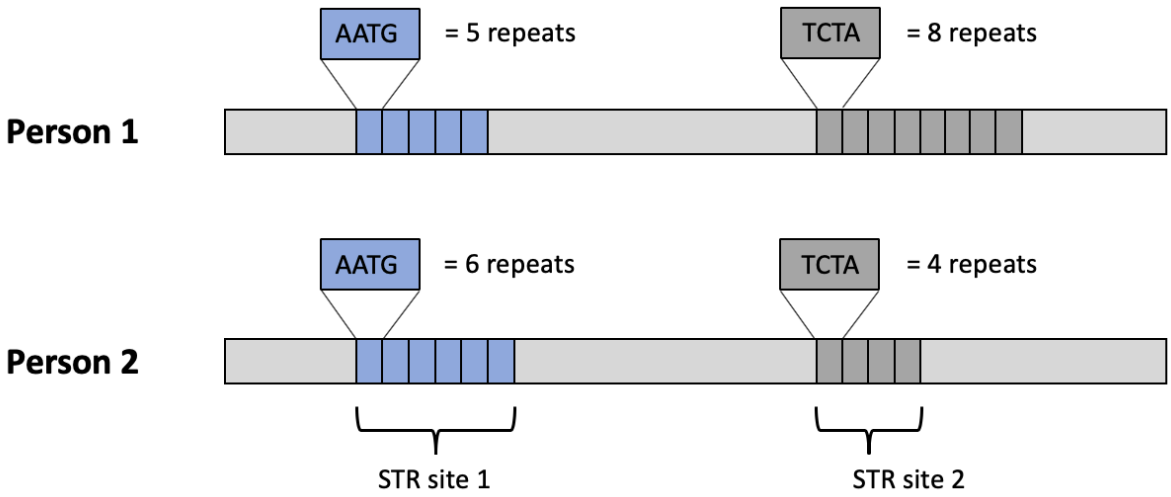


Figure 1: *The repeat number in an STR site can differ between individuals, while the repeated DNA sequence is the same. This concept is exemplified with two individuals (Person 1 and -2) for two different sites (STR site 1 and -2) in the figure.*

An STR analysis reveals the number of repeats for the STR markers and can for instance be used to map relationship or to predict a person's origin. The STRs have high mutation rates and are suitable markers in genetic analyses due to their polymorphism.

The STR markers are generally named by a given standard, as for example D7S820, where D represents DNA, 7 means chromosome 7 on which the STR marker is located, S stands for STR, and 820 is the markers unique identity (Fan and Chu 2007). The alleles of an STR refer to different variants of repeat numbers for the particular STR. The alleles are generally named by the number of repeats which they contain, e.g. 6, 16 or 17. Some alleles consist of an incomplete repeat in addition to a number of complete repeats. In such case, the name of the allele should be designated by the number of complete repeat units followed by a decimal point and the number of base pair (bp) of the partial repeat (Fan and Chu 2007), e.g. the 10,1 allele of the D7S820 marker.

Amelogenin (AMEL) is an STR marker found on both the X- and Y chromosomes and is used in sex detection. A deletion of 6 bp in AMEL on the X chromosome (AMELX) makes it possible to distinguish between AMELX and AMELY. This results in a homozygote top for female samples (two AMELX) and two heterozygote tops for male samples (one AMELX and one AMELY) (Butler 2005, Chapter 5).

STR Panels

STRs are the main markers in current forensic DNA analyses. As previously mentioned, REFA analyses all samples with a single STR panel, PowerPlex® Fusion 6C System (Fusion 6C). In more complex cases, additional STR panels, such as Investigator HDplex (HDplex) are applied. In this project, all samples have been analysed with Fusion 6C and HDplex, commercially available from Promega and Qiagen, respectively.

Fusion 6C includes 27 STRs, of which 23 are autosomal. The four remaining STRs are three Y chromosomal markers and AMEL. The panel consists of a great number of both common and informative STRs, resulting in a great discriminatory power (Cisana et al. 2017). HDplex consists of AMEL and 12 autosomal STRs, most of these not commonly used in standard STR panels. The non-standard STRs make HDplex a well suited supplementary panel that enables higher discrimination in complex cases (Westen et al. 2012; Phillips et al. 2014). The combination of Fusion 6C and HDplex constitutes 32 unique markers when overlapping markers are taken into account.

1.2.2 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are, as the name implies, positions in the DNA sequence where a single nucleotide is exchanged with another (Figure 2).

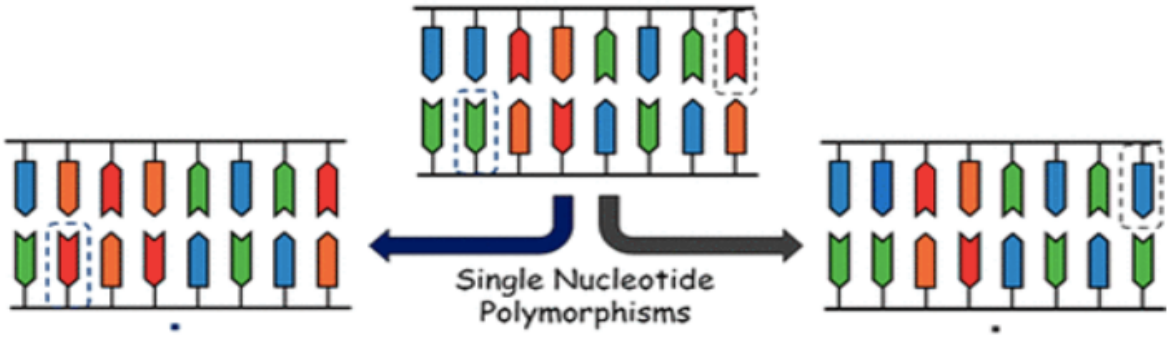
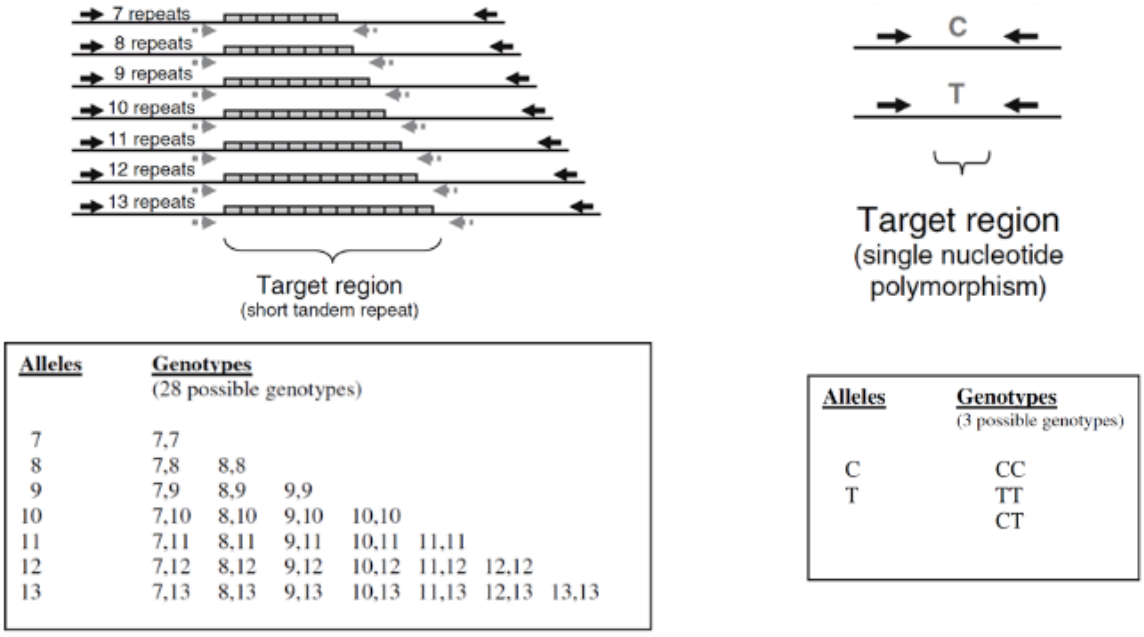


Figure 2: *Mutations can occur during DNA replication, and when this takes place in a germ cell (meiosis) new genetic variants can appear in the population. The mutations in the figure include exchange of single nucleotides, resulting in two new SNPs. Figure modified from: Ericson and Haskell-Luevano (2018).*

SNPs are the most abundant form of genetic variation between human (Hütt 2014). Today, SNPs are commonly analysed with MPS technologies or microarray platforms (Bentley et al. 2008; Goodwin et al. 2016). All known SNPs are given unique reference SNP ID numbers

(rs ID), such as rs1005533 (Bethesda (MD): NCBI (US) 2005). The ID number is not related to the position of the SNP.

SNPs can be used for many of the same purposes as STRs, but an important difference between the markers is that different SNP alleles have the same length and are distinguished based on the occurring nucleotide, rather than size. Thus, SNPs have much fewer possible alleles compared to STRs (Figure 3).



Alleles	Genotypes (28 possible genotypes)						
7	7,7						
8	7,8	8,8					
9	7,9	8,9	9,9				
10	7,10	8,10	9,10	10,10			
11	7,11	8,11	9,11	10,11	11,11		
12	7,12	8,12	9,12	10,12	11,12	12,12	
13	7,13	8,13	9,13	10,13	11,13	12,13	13,13

Alleles	Genotypes (3 possible genotypes)
C	CC
T	TT
	CT

Figure 3: *Comparison of STRs and SNPs in terms of the number of possible alleles, allele combinations (genotypes) and size of the target region. Figure modified from: Butler et al. (2007).*

SNPs have several advantages over STRs, e.g. lower mutation rates, which makes them more stable markers (Gray et al. 2000). Moreover, SNPs provide advantages that simplify the analysis method itself (Kwok and Chen 2003; Sobrino et al. 2005). An important reason for this is that very short amplicons are needed as the polymorphism only includes one single nucleotide. This last property generally enables satisfactory results to be achieved, despite highly degraded DNA, to a greater extent than STR analyses. The fact that the sequencing reaction is independent of the length of the DNA fragments allows several of the fragments to have the same length without this impacting the genotype result (Kayser and De Knijff 2011). Nevertheless, it is shown in studies that SNP analyses have a higher rate of inconclusive cases, than STR analyses, when these are run separately (Amorim and Pereira 2005). However, SNP analyses' ability of multiplexing give the potential of compensation, but many

more markers would be needed. It was for instance reported by Krawczak (1999) that approximately 4.2 SNPs, with allele frequencies of 0.5, were needed to achieve the same exclusion power as one STR marker. In a study by Gill (2001) it was reported that 50 SNPs with allele frequencies of 0.2-0.8 resulted in the same LRs as 12 STRs.

SNP panels

In complex kinship cases, satisfactory results cannot always be achieved based on only STR analyses, despite use of additional STR panels. SNP panels as supplementary markers can be a potential solution to this problem (Meiklejohn and Robertson 2017). In this project, two SNP panels, commercially available from Thermo Fisher Scientific, have been used: Precision ID Identity Panel (Identity) and Precision ID Ancestry Panel (Ancestry). The two panels are further outlined below.

Identity panel

The Identity panel includes 34 Y-chromosome SNPs and 90 autosomal SNPs compiled from literature by Phillips et al. (2007a) and Pakstis et al. (2010). The SNPs in the panel show very low global allele frequency variation and are well suited markers with respect to identifying individuals independent of ancestry (Kidd et al. 2006). The small DNA amount required also contribute in making this panel suitable for forensic samples (Guo et al. 2016).

Ancestry panel

The Ancestry panel includes 165 autosomal markers combined from the Kidd panel (Kidd et al. 2014) and the Seldin panel (Kosoy et al. 2009). In contrast to the Identity SNPs, these show large allele frequency divergences between major ethnic groups and based on the observed alleles, and occurrence of these in different ethnic groups, individuals' ancestry can be predicted. Besides this, and what will be further investigated in this project, the Ancestry SNPs can also be useful in identification cases (Phillips et al. 2007b; Pereira et al. 2017).

1.3 Polymerase chain reaction

Polymerase chain reaction (PCR) is a common laboratory technique that amplifies target regions of extracted DNA in a cyclical process. Most DNA analyses require PCR of the target region prior to the analysis. A primer is a short DNA sequence necessary for PCR, this is complementary to an area in the 3´ end of the target DNA sequence. Two primers are needed to copy one target region, each complementary to either the sense or the antisense strand (Figure 4).

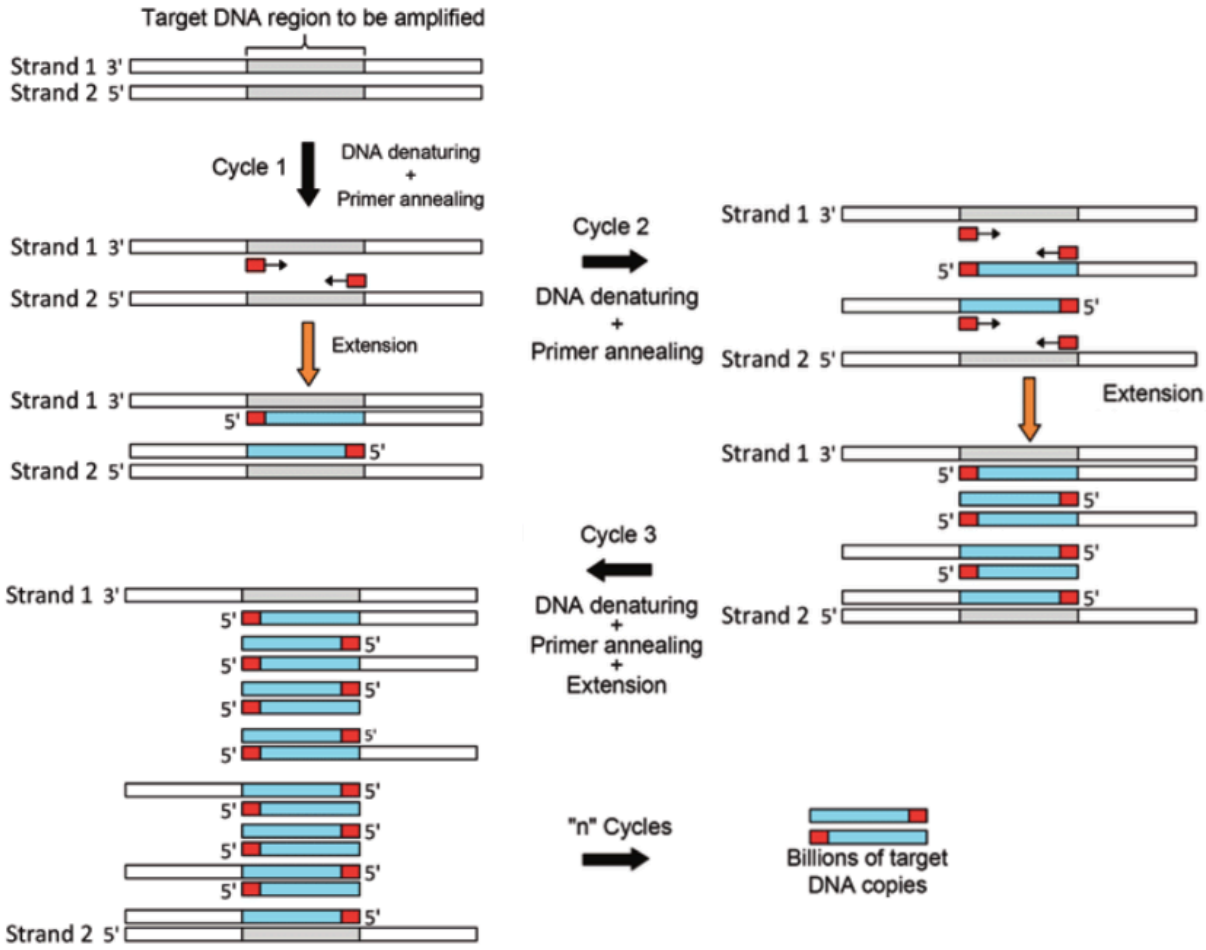


Figure 4: *The different steps in a PCR cycle. After "n" (a selected number) of cycles the result is an exponential increased in the number of copies of a target DNA sequence. Figure modified from: RescearchGate. Available from: https://www.researchgate.net/figure/Principle-of-the-PCR-in-which-small-specific-DNA-sequences-primers-are-used- (Access date: 12/5/18).*

Several components are needed for a PCR to take place: DNA template, forward- and reverse primers, nucleotides and temperature-mediated DNA polymerase. $MgCl_2$ and a buffer solution is added to the components to keep the right conditions during the PCR. The reaction

solution is placed on a thermo cycler where a cyclical temperature change leads to an exponential increase in the number of target DNA sequences. In a traditional PCR, there are three main steps in one cycle: denaturation, annealing and extension. Furthermore, it is common to add an initialization step prior the first cycle and a final elongation post to the last cycle. The initialization step is performed only when the applied DNA polymerase requires "Hot start" to be activated. The activation usually takes place at between 90°C and 95°C. During denaturation, double stranded DNA become single stranded, and the optimal temperature for this reaction is usually 94°C. The annealing step comprises primers attaching to the complementary templates, usually at between 40°C and 65°C. In the extension step, DNA polymerase attaches to the complexes of primer and template and synthesises of new double stranded DNA takes place. This occurs at approximately 72°C, which is the optimal temperature for replication mediated by the thermostable DNA Polymerase. Final elongation is an optional step to make sure all target DNA copies are completely amplified (Pelt-Verkuil et al. 2008). Furthermore, it is possible to perform a PCR in two steps instead of three. In this case, the denaturation step takes place at between 92°C and 97°C, and are followed by a combined annealing and extension step at between 50°C and 70°C (Siebert et al. 1995).

When PCR is performed for several target regions in one run, it is known as a multiplex PCR. To enable this, multiple primer pairs are added to one single reaction mix and primers for overlapping target regions are marked with fluorescence of different wave length. This is necessary for the separation of the fragments during detection in the following analysis (Pelt-Verkuil et al. 2008).

1.4 Capillary electrophoresis

Capillary electrophoresis (CE) is a subgroup of electrophoresis and is a size based molecule separation method (Butler 2005, Chapter 12). CE can, among other purposes, be used to perform fragment analyses of fluorescence marked STRs, where fragment size is called based on migration time through a capillary.

A CE consists of an anode (positive charge) and a cathode (negative charge) placed in two separate buffer cartridges. An electric field supplied by a high voltage power source is applied between the anode and the cathode, and a capillary filled with a polymer connects the two buffers cartridges. DNA molecules have a negative charge due to the phosphate groups in the backbone, and the electric field initiates the STRs to travel from the cathode to the anode,

using the capillary as a bridge. The samples are first collected from their wells and pulled up in the beginning end of the capillaries. The capillaries are then moved and located in the cathode buffer. The moving STRs are detected when they passes a glass window located near the end of the capillary (Butler 2005, Chapter 12). The chemical principle of CE is illustrated in Figure 5.

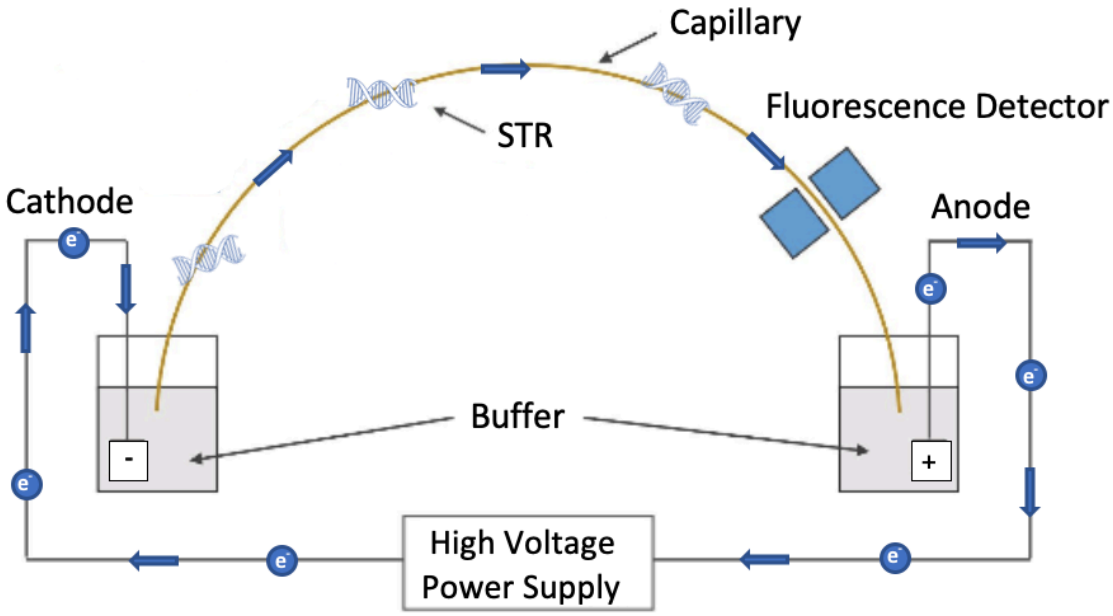


Figure 5: *Schematic illustration of DNA migration in CE. Due to the electric field between the anode- and cathode cartridges, the negative charged DNA fragments move through the capillary. Small fragments move faster than larger, and the migration time is detected near the end of the capillary. Figure modified from: Burri (2015).*

Small fragments move faster through the pores of the polymer than larger fragments, i.e. the migration time through the capillary is proportional to the size of the fragments and can be used to separate the STRs. Factors that affect the STR separation are pore size in the polymer, applied voltage in the electric field, charge of the STRs (all DNA fragments have the same electric charge per bp) and fragment size (Butler 2005, Chapter 12; Lipfert et al. 2014). However, size is the only factor that does not affect all STRs to the same extent.

Through the glass window near the end of the capillary, the fluorescence marked STRs are hit by a narrow beam of laser light. This results in excitation of fluorescent molecules, followed by spontaneous emission of these. The emitted light is detected, and the software generates an electropherogram for the DNA fragments based on the migration time (Gooijer et al. 2000; Butler 2005). The number of capillaries in the CE instrument correspond to the number of

samples that can be analysed simultaneously. The 3500xL Genetic Analyzer, used in this project, consists of 24 capillaries.

An Internal Lane Standard (ILS) containing DNA fragments of known sizes are included in all wells prior to the CE. From the analysis of the ILS fragments, a function of time and size is constructed and used to call the size of the unknown fragments in the samples. Furthermore, a ladder is included in each injection. This consists of several known alleles of all markers in the present panel, and contributes to correct genotype calling of the unknown fragments (Schumm 1997; Butler 2005).

1.5 Semiconductor sequencing

Semiconductor Sequencing is an MPS technology based on the principal of sequencing-by-synthesis. When a nucleotide is incorporated into a growing DNA strand, protons (H+) are released and detected by an electrochemical detector (Merriman et al. 2012). The technology does not require modified nucleotides or optics, and differs in that way from other MPS technologies (Mascher et al. 2013). The method has a number of applications, including SNP genotyping for use in forensic cases. Before the sequencing can take place, library preparation and template preparation have to be performed on extracted DNA. This can be done in different ways and the preparations necessary prior to sequencing with the Ion Torrent™ technology in explained the next sections.

1.5.1 Library preparation

The result of a library preparation is multiple DNA fragments of similar size with a known adapter sequence attached to both the 3´- and 5´ ends. The principle is illustrated in Figure 6. One library corresponds to a single sample and multiple libraries correspond to multiple samples, each marked with its own unique adapter sequence (Guo et al. 2016).

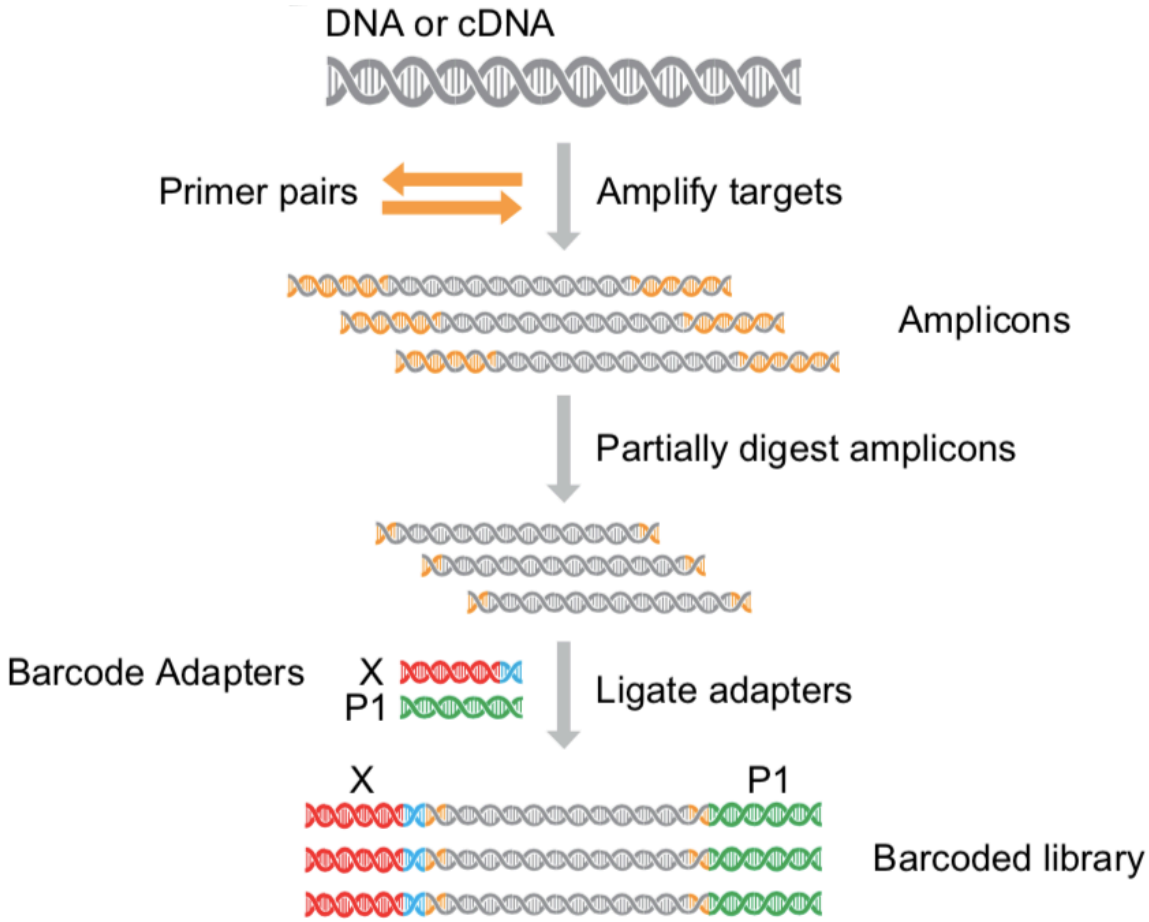


Figure 6: *The workflow in Ion Torrent ™ library preparation. The blue part of the X adapter represents a unique barcode sequence. This is essential to distinguish between the samples in a library. P1 and the red part of X are an anchor- and primer sequences necessary for the following temperate preparation. Unlike the barcode, these sequences are the same for all samples. Figure modified from: Thermo Fisher. Available from: https://www.thermofisher.com/order/catalog/product/4480442 (Access date: 12/12/19).*

The library preparations take place in separate sample wells and consists of several steps. First, forward and revers primers from a desired panel are added to the extracted DNA samples. After a multiplex PCR are performed, the primer sequences located at both ends of the target DNA fragments are partially digest, and the primer fragments and excess primers are washed away. Further, two different adapters, P1 and X, are added to the reaction wells. P1 is a DNA sequence identical for all the samples and complementary to the anchor sequences on the emulsion PCR (emPCR) beads. The X adapter consists of two parts, one that function as a primer binding site, and one that is a unique barcode sequence with the function of marking and separating the different samples. This is necessary as all samples are pooled together after the separate library preparation, resulting in one combined library (Mäki et al. 2016).

1.5.2 Template preparation

The template preparation consists of emPCR and chip loading. The principle of emPCR is illustrated in Figure 7.

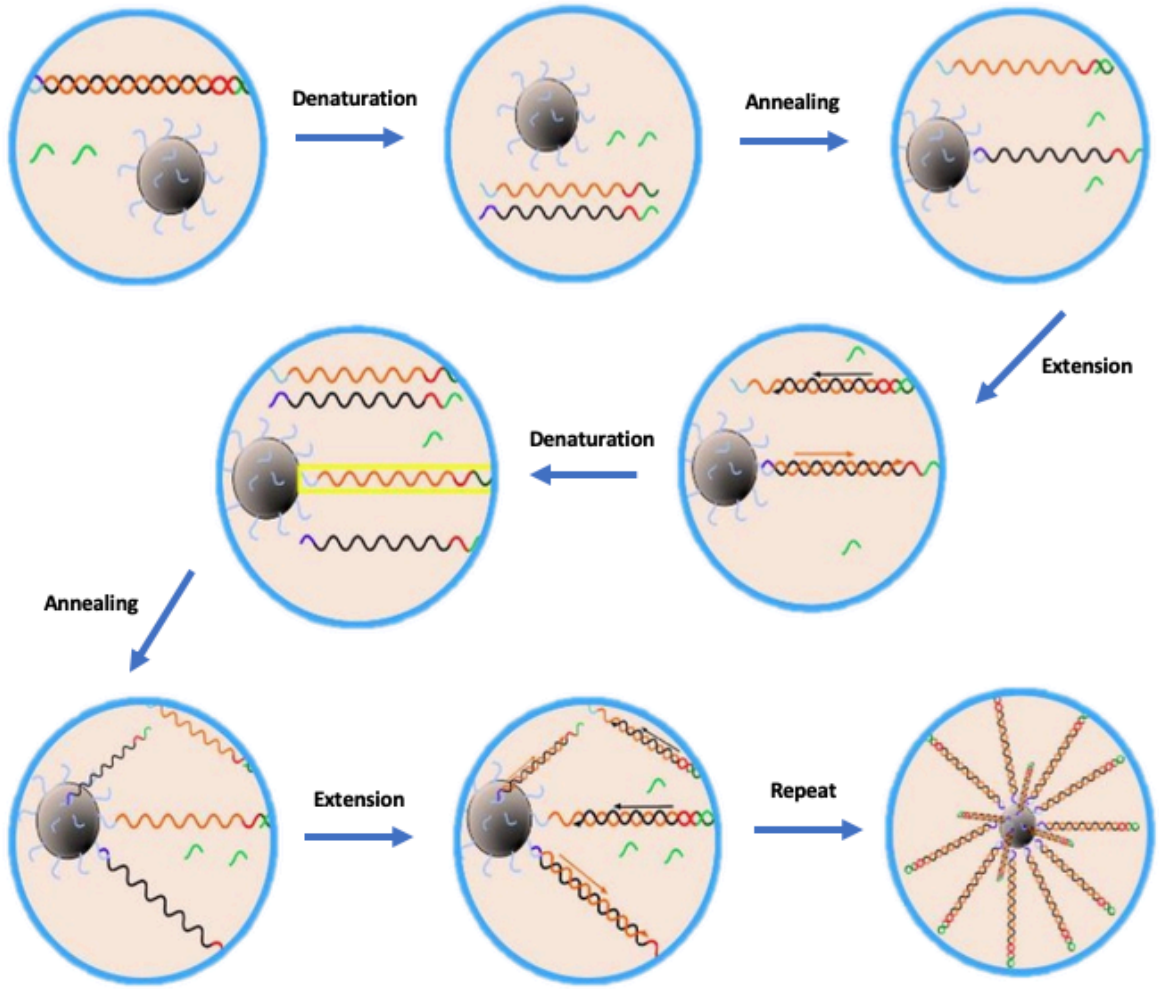


Figure 7: *The different steps of an emPCR. The emPCR reactions take place in separate oil droplets as a part of the template preparation prior to the Ion PGM™ sequencing. Each oil droplet contains a single DNA fragment, one bead, to primers complementary to the library adapter sequences (one free and one attached to the bead) and other reagents necessary for PCR. The emPCR follows the steps in the illustration, and results in a bead covered in identical DNA fragments. Figure modified from: SlideShare. Available from: https://www.slideshare.net/salmanjamil16/emulsion-pcr. (Access date: 12/3/19).*

DNA molecules attached to adapters and barcodes are found in an oil-water emulsion, where the oil droplets constitute separate reaction vesicles. A correct DNA concentration is crucial for an optimal emPCR, as one oil droplet ideally should contain one single DNA fragment. If more than one fragment is present in one droplet, this could lead to polyclonal beads, and further impair the total coverage in the analysis. Additionally, each droplet should consist of one bead, two primers complementary to the library adapter sequences and other reagents

necessary for PCR. One of the primers are bound to the bead, while the other is found free in the solution (Buermans and Den Dunnen 2014). The result of the emPCR is beads covered in plenty of identical DNA fragments, which are finally loaded onto a sequencing chip (Nakano et al. 2003).

1.5.3 Sequencing on the Ion Personal Genome Machine™

The Ion Personal Genome Machine™ (PGM™) by Thermo Fisher Scientific was used for the SNP analysing in this project. The principle of the Ion Torrent™ sequencing technology is to translate chemical signals into digital information. The semiconductor sequencing takes place on a chip consisting of a flow compartment and microwells containing small solid-state pH meters. Each microwell should optimally contain one DNA template covered bead (Buermans and Den Dunnen 2014). The principle of the detection is illustrated in Figure 8.

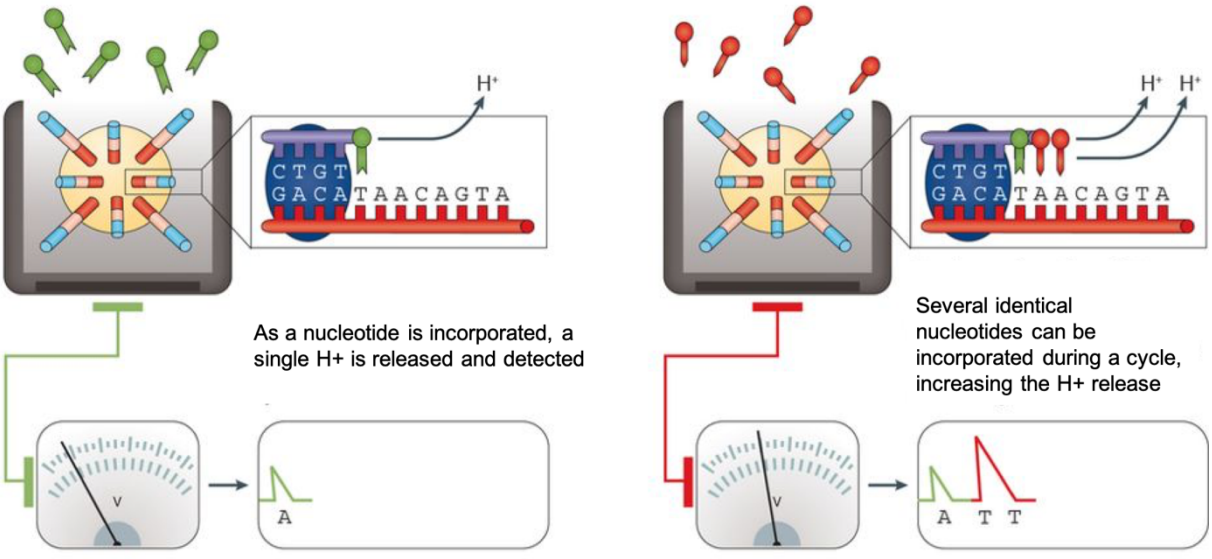


Figure 8: *Ion Torrent™ sequencing principle. A H+ is released with incorporation of a nucleotide and generates a chemical signal which is detected and transformed into digital information about the DNA sequence. In homopolymer regions, several H+ are released and a larger pH increase is generated. The reaction and detection take place in separate wells on the sequencing chip. Figure modified from: Goodwin et al. (2016).*

When a nucleotide is incorporated in a growing DNA strand, a H+ is released. Free H+ results in a change in pH, which is detected by a sensor in the wells. One nucleotide (A, T, G or C) is added to the chip at a time, and in wells where the nucleotides are complementary to the next nucleotide on the template strand, there will be a change in pH. In wells where no nucleotides are incorporated, there will be no pH change (Buermans and Den Dunnen 2014). Incorporation of several nucleotides will occur in homopolymer regions, generating a larger

pH increase. However, the single-base accuracy will decrease for homopolymer regions larger than 6-8 nucleotides (Goodwin et al. 2016).

1.5.4 Data processing

A single sequence of nucleotides representing a template sequence is called a read. The number of reads covering a target site is the coverage of this accurate site. Buchard et al. (2016) imply that generally a minimal coverage from 75 to 200 reads is desirable, but suggest that a lower number can be accepted, especially for research purposes. This will be further discussed later. The total coverage of a run is all reads produced for all analysed samples. This last property is set by the number of sensor wells on the chip and is crucial for the fundamental sequencing capacity. For the Ion Torrent™, there are three chips of different sizes available: Ion 314™, 316™, and 318™, including 1.2, 6.3, and 11.3 million sensor wells, respectively (Merriman et al. 2012). An example of a loaded Ion 314™ chip is shown in Figure 9.

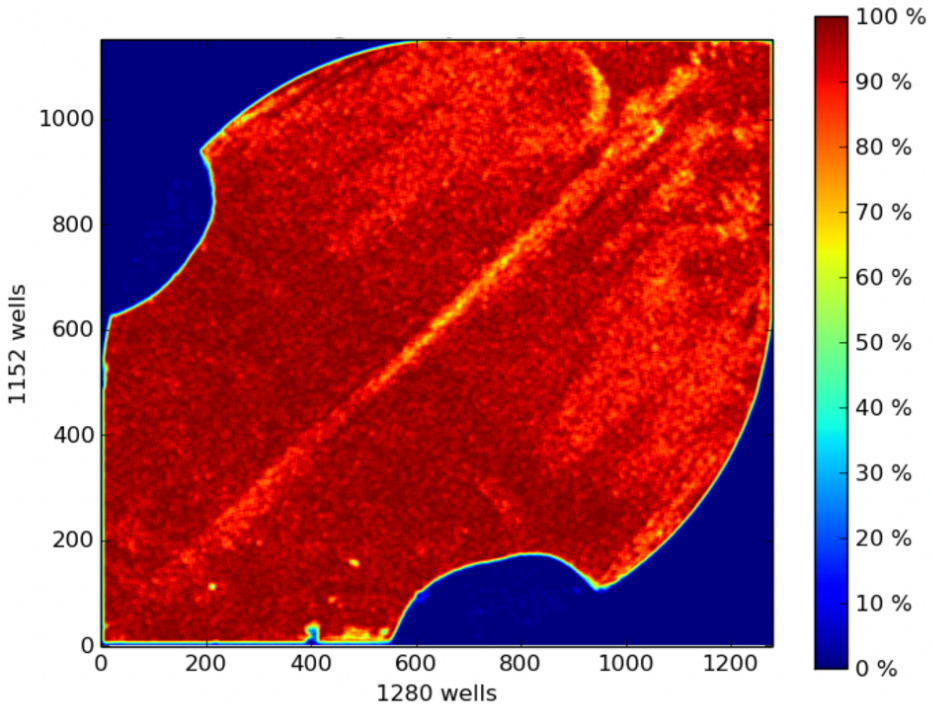


Figure 9: *A loaded Ion 314™ chip. The loading density of this chip is in average ~89%. The red areas have the highest loading density and the blue have the lowest. The "incuts" on each side of the red area should optimally also be loaded and have appeared as an artefact from the Ion Chef™ chip loading. Figure retrieved from: a result report from this project.*

The barcode regions are sequenced in the same way as the rest of the DNA fragments, and the reads are sorted and collected for each sample based on the barcode. Furthermore, a genotype

is called for every marker in each sample based on the registered reads. All sequences are read from both ends and the number of reads is reported as positive (forward)- and negative (reverse) read coverage. A Major Allele Frequency (MAF) is calculated from the nucleotide (A, T, G or C) with the most reads divided by the total number of reads for the marker, i.e. the MAF should optimally be 1.0 for homozygotes and 0.5 for heterozygotes. The background signal is calculated as the number of reads that are different from the called genotype divided by the total number of reads.

1.6 Forensic statistics

When the DNA profiles of two individuals are compared in order to infer a relationship, it is investigated how many alleles the individuals share, and a so-called likelihood ratio (LR) is calculated. Before the actual calculation can take place, there are several steps that must be performed and factors that must be taken into account. This chapter briefly outlines these concepts.

1.6.1 Allele frequency databases

Some alleles are more common than others, and the frequencies may vary considerably in different geographical areas. The tested individual's ethnic origin therefore becomes a key issue when DNA profiles are compared and LRs are calculated. Two individuals sharing a rarely occurring allele in a given area will result in a much higher probability for the potential relationship, than if the tested persons share a more frequently occurring allele. In order to map which alleles occur frequently and which occur rarely in different populations, allele frequency data must be collected, and databases associated with different populations must be constructed. Databases specific for the current populations is then used as references when LRs are calculated (Kidd et al. 2006).

1.6.2 Probabilities of genotypes

It is common to assume that the probability of observing one allele is independent of the probability of observing the other allele in the given genotype. This is called Hardy-Weinberg equilibrium (HWE). To illustrate how probability calculations are performed, imagine two alleles: a and b, with known frequencies: p_a and p_b. The genotypes for a homozygote and a heterozygote can then be calculated as:

- $Pr(a,a) = p_a^2$
- $Pr(a,b) = 2p_ap_b$

Furthermore, this assumes that the population where the allele frequencies are obtained from do not contain subdivision. If this approach cannot be assumed, it can be accounted for by adjusting the genotype probabilities described above, commonly referred to as "theta correction", also known as Wright's fixation index F_{ST} (Wright 1931). This correction increases the probability of homozygotes to a desired extend, depending on the degree of homozygosity in the subpopulation (Council 1996, Chapter 4; Egeland et al. 2015, Chapter 2).

1.6.3 Rules of inheritance

The term "Identical by State" (IBS) is used to describe two identical alleles that do not necessarily originate from the same ancestor. If, on the other hand, the alleles are inherited from the same ancestral allele, these are also referred to as "Identical by Descent" (IBD). The principle of inheritance of ancestral alleles is illustrated in Figure 10.

Figure 10: *The inheritance of alleles in a genetic position is illustrated in the pedigree. The B-allele found in both siblings are inherited from the same ancestor (the mother), i.e. this allele is IBD (and IBS). The A-allele shared by the siblings are identical, but not inherited from a common ancestor, i.e. this allele is only IBS.*

For a given genetic position, 0, 1 or 2 alleles can be shared, and if a wide range of markers are analysed, pairwise relationships without inbreeding are expected to follow a particular inheritance pattern. For instance, we expect full siblings to share two alleles IBD from their parents in 25%-, one allele in 50%- and no alleles in 25% of the investigated genetic markers. The expected inheritance pattern is given for some common relationships in Table 1.

Table 1: *Expected IBD=0, 1 and 2 probabilities for different relationships (Russel et al. 2011).*

Relationship to child	**Pr(IBD=0)**	**Pr(IBD=1)**	**Pr(IBD=2)**
Monozygotic twin	0	0	1
Parent	0	1	0
Full sibling	0.25	0.5	0.25
Half sibling Full sibling of parent Grandparent	0.5	0.5	0
First cousin Half sibling of parent	0.75	0.25	0
Unrelated	1	0	0

The table above shows that for instance half sibling, full sibling of parent (uncle/aunt) and grandparent have the same IBD probabilities. A further explanation of how IBD probabilities can be used to infer kinships is detailed later.

1.6.4 Dependent markers

Genes or markers that are found close on a chromosome and show dependent assortment are known to be linked. During meiosis, homologous chromosomes (pair of one paternal and one maternal chromosome) undergo crossover. Linked markers are less likely to be separated by crossover and are often observed together, this is illustrated in Figure 11 (Lesk 2017, Chapter 3).

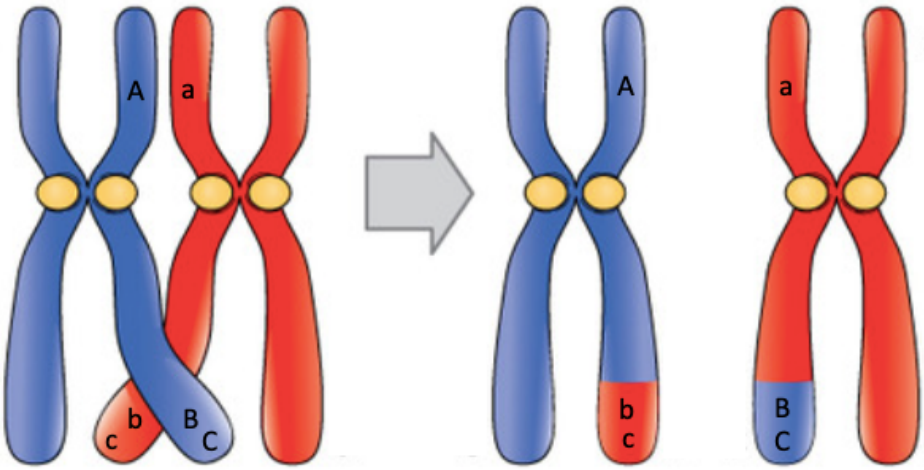


Figure 11: *Illustration of a crossover resulting in a recombination. Linked markers (B/C and b/c) are found close on the chromosome and are inherited dependent of each other. Marker A and a are not found close to the other markers and are independent of these. Marker A was found on the same chromosome as marker B and C before the crossover, but not after, i.e. a recombination has occurred. Figure modified from: Lumen learning. Available at: https://courses.lumenlearning.com/boundless-biolog.*

Recombination occurs when crossover leads to new arrangements on a chromosome, as shown in the figure above. If several crossovers occur, resulting in the three markers being reunited on the same chromosome again (known as a double crossover), recombination has not occurred. The distance between chromosomal positions for which the expected number of crossovers in a single generation is 0.01, is known as a centiMorgan (cM) (Sala and Verpelli 2016). Furthermore, alleles at different loci can show a non-random association unrelated to their physical linkage, referred to as linkage disequilibrium (LD). When alleles are in LD these occur together, at population level, more or less frequently than expected by chance (Tillmar and Phillips 2017). If linked markers and markers in LD are not taken into account when relationship probabilities are calculated, it can lead to incorrect results. The more markers applied in a relationship calculation, the greater becomes the chance of linkage and LD between some of the markers. The risk is further increased when several panels are combined in an analysis. If information about linkage and LD is not available for the markers of interest, it is important to take this into account when the results are evaluated.

1.6.5 Formulation of test hypotheses

Before performing calculations in a kinship case, it is common to formulate two competing hypotheses: H_1 (main hypothesis) and H_2 (alternative hypothesis). In most areas, hypotheses are written with parameters of a statistical model, while in forensics, the hypotheses are formulated verbally. In a paternity case, these might be formulated:

- H_1: The alleged father is the true father of the child
- H_2: A random man is the father of the child

Where a "random man" refers to an unrelated man from the same population as the child (Egeland et al. 2015).

Furthermore, in forensics, it is considered equally important to avoid rejection of either hypothesis. The purpose is not to prove a relationship beyond reasonable doubt, but to determine the most likely hypothesis, i.e. there is no null hypothesis. Thus, the two hypotheses are referred to as H_1 and H_2 (Egeland et al. 2015).

1.6.6 The likelihood ratio

The likelihood ratio (LR) compares the probability for the genetic data given that H_1 is true, with the probability for the data given that H_2 is true. If one look back at the example hypotheses above, a high LR is achieved if the genotypes of the alleged father and the child generate a high probability for the alleged relationship, and a small probability for them to be unrelated. The LRs are formed as (Egeland et al. 2015):

$$LR = \frac{\Pr\,(data|\mathrm{H}_1)}{\Pr\,(data|\mathrm{H}_2)}$$

Where *data* represent the evidence, e.g. the genotypes of the individuals in a kinship case. An example of LR calculation based on the test hypotheses for the alleged father (AF) and the child (CH) is illustrated below:

We assume that we only have DNA data for the two stated persons. For simplicity only one marker is used in the calculation. For this marker, AF has the genotype: a,b, and CH has the genotype: a,a. The frequencies of the alleles are: p_a= 0.3568 and p_b = 0.1296.

$$LR = \frac{P(G_{ch}, G_{AF}|H_1)}{P(G_{ch}, G_{AF}|H_2)}$$

$$= \frac{\Pr\,(G_{CH}|G_{AF}, H_1)}{\Pr\,(G_{CH}|H_2)} * \frac{\Pr\,(G_{AF}|H_1)}{\Pr\,(G_{AF}|H_2)}$$

Where G_{CH} represents the genotype of the child, and G_{AF} represents the genotype of the alleged father. In most cases the last part of this formula ($\frac{\Pr\,(G_{AF}|H_1)}{\Pr\,(G_{AF}|H_2)}$), is equal to 1, and

therefore not mentioned. When assuming that the probability of observing G_{CH} and G_{AF} is independent, the following rule is applied: $\Pr(G_{CH}|G_{AF}, H_2) = \Pr(G_{CH})$. This leads to the final formula (Egeland et al. 2015):

$$LR = \frac{\Pr(G_{CH}|G_{AF}, H_1)}{\Pr(G_{CH})}$$

For the test example this gives:

$$LR = \frac{0.5 * p_a}{{p_a}^2} = \frac{1}{2 * 0.3568} = 1.40135$$

By applying the "product rule", a combined, total LR for several independent markers can be achieved. To get to this, the independent LRs are simply multiplied together (Gjertson et al. 2007):

$$LR_1 * LR_2 \ldots * LR_n = \text{combined LR}$$

Where LR_1 is based on the first marker, LR_2 on the second and *n* refers to the number of markers. A prerequisite for applying this rule is that all included markers are independent of each other, i.e. they are not linked or in LD.

If the LR result in a value >1, the data are more likely if H_1 is true. LR <1 indicate that the data are more likely if H_2 is true. An LR of 1 is achieved if the data are just as likely for both hypotheses. By applying a larger number of markers, one can expect the LR to increase for cases where H_1 is true, and to decrease when H_2 is true.

"Bayes theorem" may be used to convert the LR results (ratio of probabilities of data given the hypotheses) into Essen-Möller index (W) (Essen-Möller 1938; Egeland et al. 2015, Chapter 2). This reports the probabilities of the hypotheses given the genetic data. To do the conversion, it is a prerequisite that the prior probabilities are specified. In most cases it is appropriate to apply an equal prior probability: $\Pr(H_1) = \Pr(H_2) = 0.5$. LR can then be converted to a posterior probability by the following formula (Egeland et al. 2015):

$$\Pr(H_1 \mid data) = \frac{\Pr(data \mid H_1)}{\Pr(data \mid H_1) + \Pr(data \mid H_2)} = \frac{LR}{LR + 1}$$

The Essen-Möller index gives values in the interval: 0-1, allowing the result to be formulated as: the probability of H_1 is x%, e.g. there is a 95% probability that AF is the father of CH.

When case results are reported, it is common to interpret the LRs in terms of a verbal scale. REFA's standard formulations are given in Table 2.

Table 2: *Overview of LRs, W values and associated standard comments for result reporting in DNA analyses. The verbal expressions are translated from an original Norwegian version that are being used as an indication in result reporting at REFA. Additionally, some small modifications have been made in regard to the intervals.*

LR	W value	Verbal expression
> 1,000,000	> 99.9999%	very substantial weight (conclusive)
999,999 – 100,000	99.999%	substantial weight (conclusive)
99,999 – 10,000	99.99%	very great weight (conclusive)
9999 – 1000	99.9%	great weight (conclusive)
999 – 100	99%	moderate weight
99 – 0.01		inconclusive
0.009 – 0.001	1%	cannot exclude, relative chance of 1%
0.0009 – 0.0001	0.1%	cannot exclude, relative chance of 0.1%
0.00009 – 0.00001	0.01%	cannot exclude, relative chance of 0.01%
0.000009 – 0.000001	0.001%	cannot exclude, relative chance of 0.001%
< 0.000001	< 0.0001%	cannot exclude, relative chance less than 0.0001%

As shown in the table, no relationship indications are given for cases with LRs between 99 and 0.01 (inconclusive). Cases consisting of LRs in the ranges 100 to 999 can indicate a more likely hypothesis, but with some uncertainty. In order to conclude that H_1 is true, it is desirable to achieve an LR >1000 (great evidence weight). In cases resulting in LRs <0.01 and >0, it is reported that the alleged relationship cannot be excluded, but has a relative chance of e.g. 0.1% (LR=0.0009).

1.6.7 Calculations in Familias

Familias is a software that can perform probability calculations for several purposes based on DNA data (Egeland et al. 2000; Kling et al. 2014; Kling and Füredi 2016; Kling et al. 2017). The software is generally used to infer relationships between a set of persons. In this matter, the accurate allele frequency database is imported, as well as the case specific DNA data. Two, or more, competing hypotheses are formulated, and Familias calculates an LR for the

alleged relationship. Familias is especially useful for calculations in complex cases, e.g. cases concerning distant relationships, inbreeding or cases where mutations are necessary to explain the data. Familias can perform calculations with or without regard to inbreeding and mutation rates, and several mutation models are available. Furthermore, features such as simulations and blind search are available. These are further explained below.

Simulations

Simulation is an approach where genotypes are constructed based on allele frequency data from a relevant population (Egeland et al. 2015, Chapter 2). In Familias, simulations of specific relationships can be performed. Prior to this, allele frequencies are defined, and if relevant, a mutation model is applied before the test persons and hypotheses are stated. The pedigrees defined in the hypotheses are simulated a given number of times, e.g. 10,000, and for each simulation an LR is calculated. Furthermore, results are given for when H_1 is true and when H_2 is true, and are reported as median LR and proportion of cases exceeding an LR limit of 100, 1000 and 10,000. Thus, the simulation results can give an impression of which LRs to expect for true- and false relationships in the given population.

Blind search

A blind search in Familias can be performed for a group of individuals with known DNA profiles and unknown relationships. Prior to the search, relationship hypotheses are selected, e.g. half siblings vs. unrelated. An LR threshold is set to limit the results and exclude irrelevant matches, and the result list includes combinations of individuals from the group with LRs above the defined threshold. This approach can for instance be useful in respect to discover unknown relationships among a group of individuals, or as a control of performed calculations.

In addition to LRs, the blind search provides result parameters independent of the hypotheses. For instance, the degree of 0, 1 and 2 shared alleles between individuals, i.e. alleles IBS are reported. This can be used in the evaluation of which relationships are most likely for two individuals. Furthermore, Familias uses a maximum likelihood approach to infer the IBD = 0, 1 and 2 probabilities (Fisher 1922; Aldrich 1997). A range of different combinations of IBD probabilities are tested, and the most likely combination, considering the DNA data and the allele frequency database, are reported. From the IBD probabilities, the most likely relationship (hypothesis) can be indicated by comparing these to the expected values for

different relationships (Table 1). Due to SNPs' low number of potential alleles and higher possibility of sharing an allele not IBD, these markers have a greater chance of inflating the IBS values, compared to STRs. This is taken into account when IBD probabilities are estimated in Familias, i.e. this parameter constitutes a more a realistic measure of how much shared genetic material that originate from a common ancestor.

2 Material and methods

A general explanation for standard methods were given in the introduction chapter. In this chapter a more specialised description is given for the methods used in this project.

2.1 Sampling

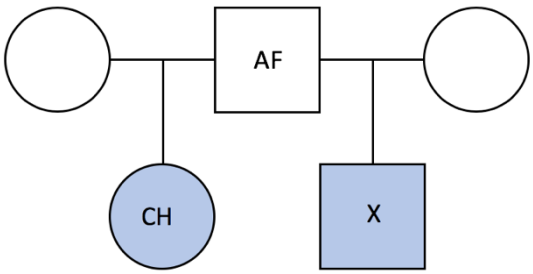
16 samples from eight complex kinship cases were analysed in this project. The cases included relationships such as half siblings and had previously been analysed with the standard STR panels at REFA. Seven of the cases had LRs in, or very close to, the inconclusive range of 0.01-99. One of the cases had an LR of about 1000. In this project, the cases were analysed with additional SNP markers with the aim to achieve more conclusive LR results. The DNA was obtained from FTA cards with buccal cell samples. All markers applied in this project are listed in Appendix A and reagents and analysis components mentioned in this chapter can be found with additional information (LOT and supplier) in Appendix B.

2.2 Hypotheses of project cases

Two competing hypotheses were formulated for all project cases, and relevant kinships were illustrated by family trees, see Figure 12.

- Half sibling vs. unrelated: case 1, 2, 3 and 6

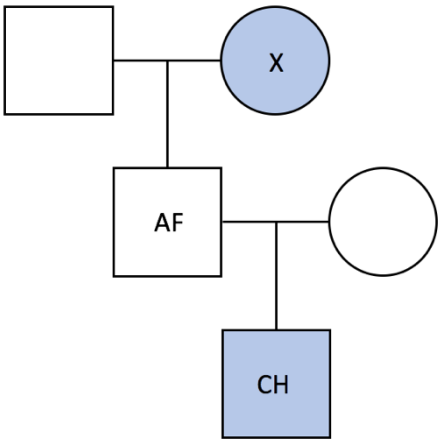
  H_1: X is a half sibling of CH H_2: X is unrelated to CH



- Grandparent vs. unrelated: case 4

H_1: X is a grandparent of CH

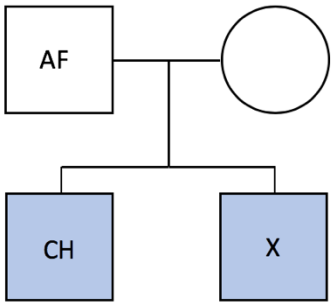
H_2: X is unrelated to CH

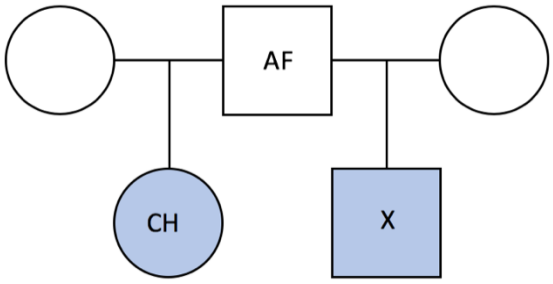


- Full sibling vs. half sibling: case 5

H_1: X is a full sibling of CH

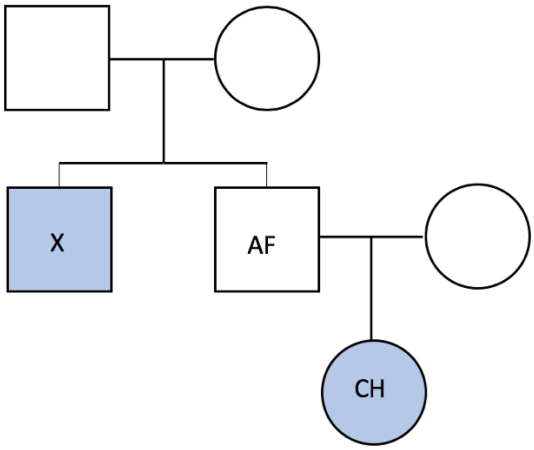
H_2: X is a half sibling of CH





- Full sibling of parent vs. unrelated: case 7

  H_1: X is a full sibling of CH's parent H_2: X is unrelated to CH



- Half sibling of parent vs. unrelated: case 8

  H_1: X is a half sibling of CH's parent H_2: X is unrelated to CH

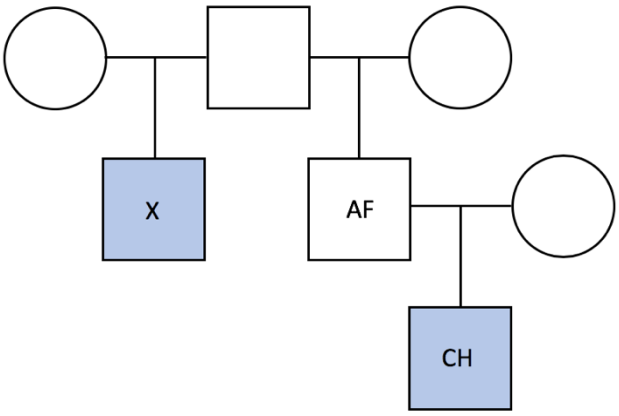


Figure 12: *Main hypothesis (H_1) and alternative hypothesis (H_2) for the eight project cases illustrated by family trees. All cases include DNA samples form a child (CH) and an alleged relative (X), these are illustrated with blue figures in the family trees. X are claimed to be unrelated to the child in all alternative hypotheses, except in case 5. The alleged father of CH is referred to as "AF".*

2.3 Short tandem repeats

As previously mentioned, REFA analyses all samples with PowerPlex® Fusion 6C System (Promega). In complex cases, supplementary STR panels, such as Investigator HDplex Kit (Qiagen), are applied. In this project, all samples have been analysed with both Fusion 6C and HDplex. The panels include 23 and 12 autosomal STRs, respectively, and constitute 32 unique markers when overlaps are taken into account.

2.3.1 Pre-PCR sample processing

FTA cards were automatically «punched» using BSD 600 Duet puncher (Microelectronic Systems). This instrument scans the barcodes of the samples and records in which well of the PCR plate the punches are located. Each sample was punched once with a diameter of 1.2 mm, and a cleaning punch was performed between each sample. To avoid cross contamination between samples included in the same case, the samples were punched in random order. All samples were analysed in duplicates and a positive control, human DNA from a male with a known profile for the current markers (2800M Control DNA, Promega), was added to the plate. Negative controls consisted of only PCR-mix. 5X AmpSolution™ Reagent (Promega) was added to the PCR mix, replacing the otherwise necessary isolation step. The reagent enables the amplification to take place while the DNA is still attached to the FTA punch (direct amplification), and result in amplicons found free in the solution.

2.3.2 Polymerase chain reaction

The PCRs for Fusion 6C and HDplex were performed at different instruments, and with small differences in the setups:

PowerPlex® Fusion 6C System:

The PCR setup contained volumes according to the supplier's protocol[1] (Table 3).

Table 3: *Reagent volumes added to each sample pre-PCR (Fusion 6C).*

Reagent	**Volume**
PowerPlex® Fusion 6C 5X Master Mix	2.5 µl
PowerPlex® Fusion 6C 5X Primer Pair Mix	2.5 µl
5X AmpSolutionTM Reagent	2.5 µl
Nuclease-free water	5.0 µl
Total	**12.5 µl**

For the Fusion 6C analyses, a two-step PCR were performed on the Veriti Thermal Cycler (Thermo Fischer Scientific) with the following setup:

1 min at 96°C, followed by 27 cycles of 96°C for 5 seconds and 60°C for 1 minute. At last 60°C for 10 minutes, and then held at 4°C until further processing.

The PCR plate was spun down to avoid contamination between the samples during removal of the seal. A mix of formamide and ILS were added to a new PCR plate, followed by PCR product (including positive and negative controls) or allelic ladder (Table 4). The plate was placed on a heating block (95°C) for 3 minutes, followed by 3 minutes on a cooling block (4°C) to denaturate the DNA.

Table 4: *Reagent volumes added to the PCR products pre-CE (Fusion 6C).*
**One well contains allelic ladder or PCR product*

Reagent	**Volume per well**
WEN Internal Lane Standard 500	0.5 µl
Formamide	7.5 µl
PowerPlex® Fusion 6C Allelic Ladder Mix*	1.0 µl
PCR product*	1.0 µl
Total	**9.5 µl**

[1] PowerPlex® Fusion 6C System for Use on the Applied Biosystems® Genetic Analysers, TMD045, Thecnical Manual, Promega.

The setup for the duplicate STR analyses prior to the CE is illustrated in Figure 13.

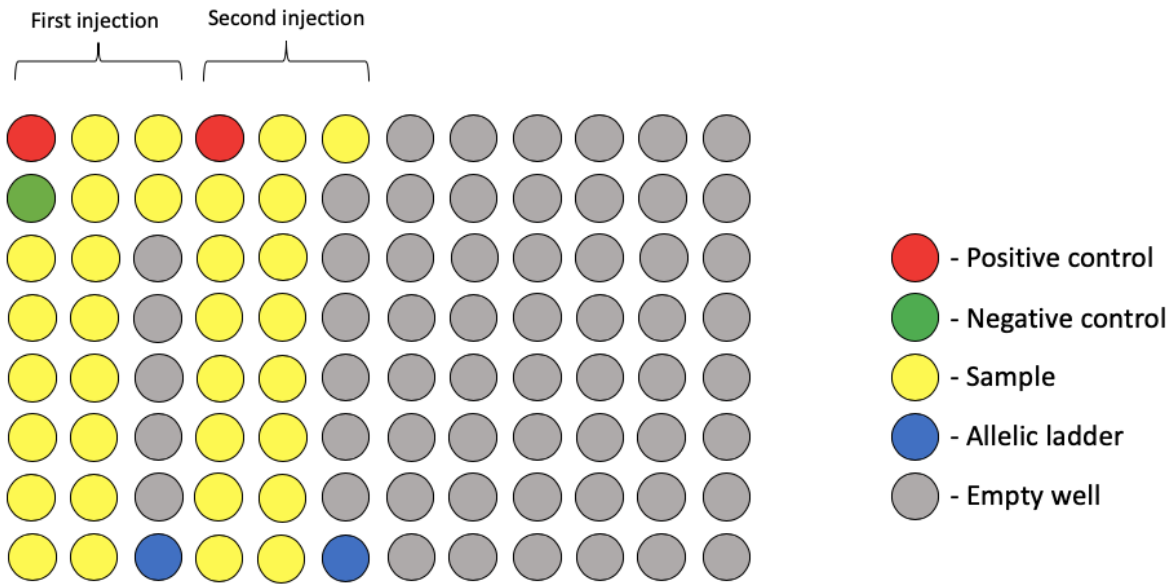


Figure 13: *CE setup for 16 samples analysed in duplicates. A positive control is included for each injection (24 wells), this must be accepted for the analyses to be considered reliable. A negative control is included once per mix batch to reveal possible contamination of the reagent mix. Furthermore, an allelic ladder is included in each injection. No wells included in an injection can be empty, i.e. the "empty" wells of the first and second injection (19-23 and 37-47, respectively) contain formamide and ILS.*

Investigator HDplex Kit:

The PCR contained volumes according to REFA's protocol, except one adjustment of replacing 2.5 µl nuclease-free water by a corresponding volume of 5X AmpSolution™ Reagent to facilitate direct amplification (Table 5).

Table 5: *Reagent volumes added to each sample pre-PCR (HDplex).*

Reagent	**Volume**
Reaction Mix A	2.50 µl
Primer Mix	1.25 µl
MultiTaq2 DNA Polymerase	0.30 µl
5X AmpSolution™ Reagent	2.50 µl
Nuclease-free water	5.95 µl
Total	**12.50 µl**

For the HDplex analysis, a three-step PCR were performed on the GeneAmp™ PCR System 9700 thermal cycler (Thermo Fisher Scientific) with following setup:

4 minutes at 94°C, followed by 27 cycles of 94°C for 30 seconds, 60°C for 120 seconds and 72°C for 75 seconds. After this the samples were held at 68°C for 60 minutes, then at 10°C until further processing.

The PCR products were processed as described for Fusion 6C, but with the reagents listed in Table 6.

Table 6: *Reagent volumes added to the PCR products pre-CE (HDplex).*
**One well contains allelic ladder or PCR product*

Reagent	**Volume per well**
DNA Size Standard 550 (BTO)	0.5 µl
Formamide	7.5 µl
Allelic Ladder HDplex*	1.0 µl
PCR product*	1.0 µl
Total	**9.0 µl**

2.3.3 Capillary electrophoresis

The CE was performed using the Genetic Analyzer 3500xl (Thermo Fisher Scientific), with DataCollection v.2.0. Preparations of the instrument, as well as analysis settings, were performed according to REFA's protocols. Preparations included checking the level of polymer (POP-4®), anode buffer and cathode buffer. The pump chamber and the channels were checked to be free from bubbles. The setup for the run was imported, and the prepared 96-well plate was put on to the instrument.

2.3.4 Data processing

The STR data was imported as a HID-file to GeneMapper® ID-X 1.4 (Thermo Fisher Scientific). This genotyping software provides DNA sizing and quality allele calls.

The results of the STR analyses are visible as peaks for the detected fluorescence. High peaks correspond to strong fluorescence signal, however, small peaks due to background signals will always occur. An analytical threshold is applied to avoid calling peaks that are not associated with true alleles. In this project the threshold was set to 30 relative fluorescence

units (rfu) for weak samples and negative controls, and 100 rfu for the remaining samples and positive controls. The called peaks were further analysed with regard to several general thresholds before they were accepted as true alleles. In terms of heterozygotes or triplets, true alleles can hide among background noise in weak samples, and so-called "dropouts" can occur. To avoid this, homozygote peaks had to exceed 500 rfu and heterozygotes 250 rfu. Additionally, the heterozygote peaks had to be balanced, i.e. the minor peak had to be at least 50% of the major peak height. This threshold was set to make sure the minor peak was an actual allele, and not background signal to a homozygote peak. Furthermore, an upper threshold was set at 27000 rfu. Peaks exceeding this height are often wide and have high stutter peaks, which are both factors that can lead to incorrect allele calling. However, these thresholds should be considered general guidelines, and after a final visual analysis, some peaks were accepted and considered true alleles, even if they did not satisfy the thresholds. After the analyses, concordance was checked for the duplicates.

2.3.5 Ion Torrent™ extract control analysis

An additional STR analysis was performed for the later described Ion Torrent™ extracts. These samples were only analysed with Fusion 6C, and as the DNA were already extracted, 5X AmpSolution™ Reagent was replaced by a corresponding volume of nuclease free water. Since this analysis was only performed for control, these samples were analysed in singles and constituted only one injection in the Genetic Analyzer 3500xl. Beyond this, the analysis was performed in the same manner as described for the Fusion 6C samples above and the genotype results were checked to match the previous duplicate results.

2.4 Single nucleotide polymorphisms

In addition to the STR analyses, all samples in the project were analysed with SNP markers from two Thermo Fisher Scientific panels: Precision ID Identity Panel (Identity) and Precision ID Ancestry Panel (Ancestry). The panels include 90 and 165 autosomal SNPs, respectively. Unless other is stated, the library preparations, template preparations and sequencing were performed according to the supplier's protocol[2]. Thermo Fisher Scientific is the producer of the reagents and supplies if others are not specified.

[2] Precision ID Panels with Ion PGM™ System, Application Guide, MAN0015830, Thermo Fisher Scientific.

2.4.1 Pre-PCR sample processing

For all project samples, three 1.2 mm punches were manually punched from the FTA card and deposited into a sterile 1.5 mL Eppendorf tube. Between the sample punches, two cleaning punches were performed. 200 µl buffer G2 and 10 µl proteinase K enzyme (Qiagen) were added to each tube, before this was placed on a heating block at 56°C for 10 minutes, with a short vortex half way in the heating. Finally, the tubes were incubated at 95°C for 5 minutes. The DNA was extracted using EZ1 Advanced XL BioRobot (Qiagen), with EZ1® DNA Investigator Kit (Qiagen). The isolated DNA was eluted in 50 µl nuclease-free water. An STR control analysis were performed on the DNA extracts (2.3.5).

2.4.2 Preparing libraries using the Ion Chef™ Instrument

Library preparation was performed using the Ion Chef™ Instrument (Thermo Fisher Scientific). The preparation for the Identity – and the Ancestry analyses was performed in the same way. The 16 extracted DNA samples were transferred to two Precision ID DL8 IonCode™ Barcode Adapters 96 Well PCR Plates (IonCode™ PCR Plates), eight samples per plate. 15 µl DNA solution was deposited in well A1 to H1 as shown in Figure 14. Each well in column 6 of the IonCode™ PCR Plate contained different dried-down IonCode™ barcodes.

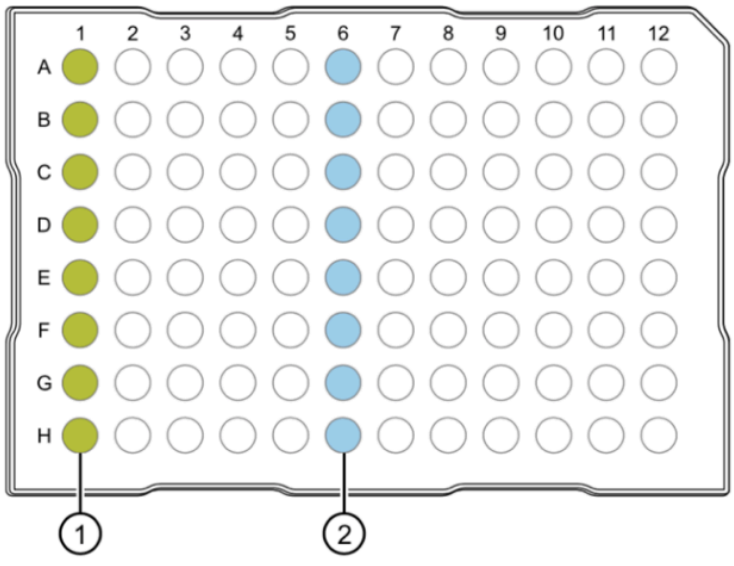


Figure 14: *An IonCode™ PCR Plate ready for library preparation on the Ion Chef.*
1: Wells containing samples.
2: Wells containing unique barcode sequence.
Figure retrieved from supplier's protocol[2].

[2] Precision ID Panels with Ion PGM™ System, Application Guide, MAN0015830, Thermo Fisher Scientific.

The Ion Chef™ Instrument was loaded with the components listed in Table 7. The respective primer panel was added to the specified position. The library preparation was performed for one IonCode™ PCR plate at a time, i.e. two rounds of preparation were performed for each of the two panels.

Table 7: *Components applied in the library preparation on the Ion Chef™.*
** One of the panels are applied in one library preparation.*

Components
Ion AmpliSeq™ Chef Supplies DL8
Ion AmpliSeq™ Chef Reagents DL8
Ion AmpliSeq™ Chef Solutions DL8
IonCode™ 0101-0132 in 96 Well PCR Plates
Precision ID Identity Panel*
Precision ID Ancestry Panel*

For the library preparations, 1 primer pool was used, 25 target amplification cycles were performed, and anneal and extension time was set to 4. After the preparation, the tube containing the barcoded library was capped and stored at -18 until the next preparation step.

2.4.3 Preparing the template on the Ion Chef™ Instrument

16 samples, i.e. two libraries, were included in one template preparation, thus, one preparation was performed for each panel. 25 µl of the prepared libraries were thawed (if frozen) and pipetted to the appropriate tubes. The Ion PGM™ Hi-Q™ Chef Reagents cartridge was thawed at room temperature for 45 minutes before it was loaded onto the Ion Chef™ Instrument. All components used in this step are listed in Table 8.

Table 8: *Components applied in the template preparations on the Ion Chef™.*

Components
Ion 314™ Chip V2 BC
Ion PGM™ Hi-Q™ View Reagents
Ion PGM™ Hi-Q Chef Solutions
Ion PGM™ Hi-Q™ Chef Supplies

Finally, the Ion 314™ Chip V2 BC chips were placed in the centrifuge buckets and loaded onto the chip-loading centrifuge. Each of these chips holds eight samples, i.e. two chips were used for each panel. A planned run was created with the Torrent Server (via the Torrent Browser).

2.4.4 Sequencing on the Ion Personal Genome Machine™

Prior to the sequencing, both water- and chlorite cleaning were performed on the PGM™ System (Thermo Fisher Scientific). The dNTP stock solutions were thawed on ice, and the nitrogen gas pressure was checked to be above 500 psi. 100 µl of 100 mM NaOH was added to the wash 2 bottle, and the pH was adjusted to 8.45 during the initialisation. Correct pH is essential for proper sequencing by the current method. All required components for the sequencing step can be found in Table 9.

Table 9: *Components applied in the Ion PGM™ sequencing.*

Components
Ion PGM™ Seq Supplies
Ion PGM™ Hi-Q™ Sequencing Reagents
Ion PGM™ Hi-Q™ Sequencing Solutions
Ion PGM™ Hi-Q™ Seq dNTP
Wash 2 Bottle kit

After the ended template preparation on the Ion Chef™ Instrument, the loaded chips were centrifuged short in a minifuge and put onto the Ion PGM™ Sequencer, one at a time. The project samples were sequenced and aligned according to the explanation in the introduction (1.5.3 and 1.5.4).

2.4.5 Data processing

SNP raw data were analysed on the Torrent Suite Server using the HID_SNP_Genotyper (v5.2.2) plugin, with the targets and hotspot regions defined (Identity: IISNPv3.20140429.Designed.bed and iiSNP_FP_v1_hotspot.bed; Ancestry: AIMv1.20140429.Designed.bed and aiSNP_FP_v1_hotspot). Default settings were applied, and the data was imported to Excel for further processing. The sample results were accepted/rejected for further analysis based on total coverage, MAF, background noise and number of positive and negative reads. There is currently no procedure for this at REFA, and

the thresholds were selected based on an average of what has been applied in similar studies. In a Danish study by Børsting et al. (2014) the minimal coverage threshold was set to 75, while in other similar studies, 20 was proved to be a propriate threshold (Bentley et al. 2008; Quail et al. 2012; Daniel et al. 2015; Zhang et al. 2017). In this project 50 was applied as threshold, allowing the other analytical thresholds to be slightly less strict, and the maximum threshold for MAF for heterozygotes was set to 0.7, to avoid exclusion of too many results. The aforementioned studies have applied 0.6-0.65. The minimum threshold for MAF for homozygotes applied in this project (0.9) was in concordance with the other studies. The maximum background signal threshold was set to 0.02 in this project and was found between 0.01 and 0.03 in the other studies, the variation explained by the use of other analytical thresholds. Furthermore, a requirement of at least 10 reads of both positives and negatives were applied. The parameter thresholds applied in this project are summarized in Table 10.

Table 10: *Applied thresholds for the different analysis parameters in the data processing of the Ion PGM™ sequencing results. MAF corresponds to major allele frequency.*

Analysis parameter	**Threshold**
Minimal total coverage	50
MAF - exclusion range	0.7-0.9
Maximum background noise	0.02
Minimum number of positive and negative reads	10 of each

Results not compatible with the parameters given in the table above were removed. SNPs were removed for all sample calculations if more than eight of the samples were found outside the restrictions for at least one parameter. Y markers were removed, as they were irrelevant in this study.

2.5 Weighting the evidence

This section covers selection of genetic markers, construction of allele frequency databases and further data analysis and statistical evaluation of STR- and SNP data performed in this project. All STR- and SNP markers applied in the analyses, with additional information, can be found in Appendix A.

2.5.1 Selection of genetic markers

Genetic positions (in cM) for both SNP- and STR markers were collected from Rutgers Map v.3 (Matise et al. 2007; Nato et al. 2012). The positions in the map were obtained with a method presented by Kosambi (1944). While all relevant SNP positions were found in Rutgers Map, the positions could not be obtained for eight of the STR markers (D1S1656, SE33, D10S2325, D13S317, Penta E, D18S51, D21S11, Penta D). Bp positions for the missing STRs were obtained from NIST STRbase[3] except D10S2325 which could not be obtained from this source and was collected from ALFRED[4]. Positions in cM for the mentioned STRs were then interpolated from the Rutger Map based on SNPs found very close on the chromosomes. According to previous studies (Kling et al. 2012b; Tillmar and Phillips 2017), a distance of 0.5 cM is adequate to obtain marker panels without significant LD between any markers. Position data for all markers were imported to Excel and sorted by chromosome and cM position. 13 pairs of markers were found less than 0.5 cM from each other on the same chromosomes and 12 Ancestry markers were excluded due to this. The last marker pair consisted of two Identity markers found 0.3 cM from each other on chromosome 7 (rs6955448 and rs917118). In an article by Tillmar and Phillips (2017), "HapMap" (Consortium 2007) have been applied for position detection, and it is claimed that the two markers are found 1.3 cM from each other and with no notable LD. Furthermore, a web-based application, SNPsnap[5] by Machiela and Chanock (2015), was used in this project to confirm that the two SNPs were not in LD. A European population was selected for these calculations, and the SNPs were included in the allele frequency databases in this project, despite the small distance according to Rutgers map. Furthermore, some of the Ancestry markers were excluded as only one allele was observed in the relevant populations and thus contributed with no information to resolving the kinship.

[3] NIST - National Institute of Standards and Technology. STRBase (SRD-130). Available from: https://strbase.nist.gov/str_fact.htm#Original (Access date: 3/2/19).
[4] ALFRED - The ALlele FREquency Database. Available from: https://alfred.med.yale.edu/alfred/recordinfo.asp?UNID=LO000546P (Access date: 3/2/19).
[5] Broad Institute - SNPsnap. Available from: https://data.broadinstitute.org/mpg/snpsnap/match_snps.html (Access date: 3/2/19).

2.5.2 Allele frequency databases

The sample donors in this project were anonymised and the ethnicities were originally unknown. However, an ancestry inference tool GenoGeographer [6], developed by Tvedebrink et al. (2017), were applied on all project samples, based on the Ancestry SNP results. The genotype probabilities in a number of defined populations were computed and the most likely ancestry were reported. Further, a Nordic population database was found fit for the subsequent calculations in the project (Appendix C). To demonstrate the importance of applying an appropriate database, all calculations were performed based on both Nordic (Danish and Norwegian) and East-African (Somali) allele frequencies, individually. Danish allele frequencies were used for Identity and Ancestry, as a Norwegian database does not currently exist for all the SNP markers. STR markers from Fusion 6C and HDplex were collected in one common database, and "STR" refers to both panels in this section.

Six different allele frequency databases were applied in the calculations; STR Norwegian, STR Somali, SNP Identity Danes, SNP Identity Somali, SNP Ancestry Danes and SNP Ancestry Somali. The two STR databases are the same that are used in the routine at REFA (Dupuy et al. 2013). The Danish Identity SNP database was obtained from a research article by Buchard et al. (2016). The Somali Identity SNP database was based on a publication by van der Heijden et al. (2017). The two SNP Ancestry databases were constructed from allele frequencies published by Kosoy et al. (2009) and Kidd et al. (2014). SNPs overlapping between the two publications were merged, and SNPs that were not included in the Ancestry panel were removed. The databases were constructed in Excel to a format compatible with Familias.

2.5.3 Calculations in Familias

All cases in this project consist of extended kinships, i.e. mutation rates were not expected to affect the results notably, and mutation rates were not applied in the calculations. Furthermore, HWE is assumed for the populations (Nordic and East-African) in this project, and theta correction was not applied in the calculations.

[6] GenoGeographer - A tool for genogeographic inference. Available from: http://apps.math.aau.dk/aims/ (Access date: 3/6/19).

Calculation of LRs for the project cases

The six frequency databases and the processed STR- and SNP data of the project samples were imported to Familias. LRs for the hypotheses of the eight project cases were calculated. Calculations were performed individually for the panels: STR, Identity and Ancestry. LRs were then calculated based on combinations of these: STR+Identity STR+Ancestry and STR+Identity+Ancestry. Calculations were performed for the Nordic- and the East-African databases respectively.

Simulations

The following relationships were simulated with Familias: full sibling vs. half sibling; half sibling vs. unrelated; half sibling of parent vs. unrelated. A person theoretically shares the same amount of unlinked autosomal markers with a grandparent, a full sibling to a parent and a half sibling. Thus, the LRs remains the same, regardless of which of these relationships that are simulated, i.e. case 4, 5 and 7 were all covered by half sibling vs. unrelated in this project. 10,000 simulations were performed separately on the following databases/combinations of databases: STR, Identity, Ancestry, STR+Identity, STR+Ancestry, STR+Identity+Ancestry. Simulations were only performed for the Nordic databases.

Blind search

Prior to the blind search, the Nordic database for STR+Identity+Ancestry and DNA data for the project cases were imported to Familias. For the same reason explained in the section above, searches were performed for the following relationships: full sibling, half sibling and half sibling of parent. Half sibling of parent was not an option in Familias blind search, and as cousins have the same IBD probabilities, it was selected instead. The result of the blind search is an exhaustive list of candidate relationships that exceeds the pre-set LR limit. In this project, the blind search was performed mainly to confirm the calculated LRs. The limit was set to LR>1 to include all the project cases. In a real situation, were the purpose is to reveal unknown relationships, this should might be set higher, e.g. LR>100.

2.5.4 Calculations in FamLink

FamLink is an alternative to Familias, that can account for linkage in LR calculations (Kling et al. 2012a). The software is a complement to Familias, which in contrast do not provide linkage adjusted calculations. Adjusted calculations for all project cases were performed in

FamLink. The Familias projects were simply imported to FamLink, together with a file containing the marker positions using the "Quick analysis" feature. FamLink does not provide calculations with mutation rates, theta correction and other versatile functionality, e.g. blind search and simulations, which are available in Familias. For this reason, and since Familias is being used in the routine at REFA, LR results from Familias were used in the further processing in this project.

2.5.5 Evaluation of sequencing coverage

All SNPs were evaluated with respect to coverage. Several statistical tests were performed in R to disclose whether the coverage for the SNP markers was affected differently by the individual samples or not. Differences may occur as a consequence of different DNA concentrations in the samples. First, an ANOVA test was performed. This is a parametric test and three assumptions must be met: the residuals are normally distributed, equal variance between the treatments and the samples are independent (Urdan 2011, Chapter 10). This was assumed for the data in this project, as the samples were independent of each other, and as a high number of markers were included. An additional t-test was performed. This does not adjust for the increasing probability of falsely rejecting an H_0 when multiple comparisons are made (David and Nagaraja 2004, Chapter 9; Urdan 2011, Chapter 10). To compensate for this, Bonferroni correction was applied. This method corrects the applied significance level, and is found by dividing the original significant level (0.05) by the number of test hypotheses (120), resulting in an adjusted significance level of 4.17E-04 (Bonferroni 1936; De Muth 2014, Chapter 11).

2.6 Ethics

REFA has an agreement («Databehandleravtale») with customers to use their samples for research purposes. The analysed samples were anonymised before the start of the project, new case numbers were constructed, and the donors of the samples were never traceable.

3 Results

The results will be divided as follows, first technical results for sequencing- and fragment analyses will be reviewed. These are more thoroughly evaluated for the Ion Torrent™ method, as this is not standardised by REFA. Additionally, the results of this method are assessed with regard to ancestry. Secondly, LR results for the project cases are reviewed for each of the panels individually and together. Further, simulation- and blind search results are presented, and finally, linkage adjusted LR results are reviewed. STR refers to both Fusion 6C and HDplex in this chapter.

3.1 3500xL Genetic Analyzer

Full profiles were achieved for all project samples, and full concordance was found between the duplicates. All controls were accepted.

3.2 Personal Genome Machine

Sequencing on this instrument is an expensive technique and the samples were not analysed in duplicates. However, the identity of the DNA extracts used in the Ion Torrent™ method were confirmed by STR analyses, and full concordance was found between the control analyses and the previous STR results. The results of the PGM™ analyses are further reviewed below.

3.2.1 Excluded SNPs

SNPs were excluded according to the description in 2.4.5 Data processing and 2.5.1 Selection of genetic markers. The excluded SNPs and the reason for their exclusions are given in Table 11.

Table 11: *SNPs excluded from the sequencing results and reason for exclusion. All excluded markers are Ancestry SNPs, except two SNPs marked: "(Identity)". The parameters above the bold line are independent of the sequencing results. The parameters below the line were made based on the sequencing results and SNPs were removed from all calculations if more than eight of the samples were found outside the restrictions for at least one of the four parameters. MAF corresponds to major allele frequency.*

Reason for exclusion	**SNP ID**
<0.5 cM distance from another marker	rs1834619, rs3827760, rs12498138, rs1229984, rs10954737, rs3814134, rs4411548, rs2033111, rs881728, rs3916235, rs1876482, rs671 (12)
Allele frequency=1, Nordic database	rs1871534, rs2814778, rs705308, rs7226659, rs7657799, rs9291090 (6)
Allele frequency=1, East-African database	rs12130799, rs12544346, rs12657828, rs12913832, rs1569175, rs174570, rs2042762, rs214678, rs3737576, rs4880436, rs6422347, rs6754311, rs818386 (13)
Allele frequency=1, Nordic and East-African database	rs1800414, rs3811801 (2)
Coverage<50	rs2504853, rs1296819, rs2306040, rs2196051, 16891982, rs192655, rs2986742, rs1369093, rs6548616, rs12439433, rs37369, rs3823159, rs10007810, rs32314, rs1407434, rs4833103, rs6990312, rs260690, rs719366 (Identity) (19)
MAF in the range 0.7-0.9	rs7520386 (Identity) (1)
Background noise>0.02	rs7722456 (1)
<10 positive or negative reads	(0)*

*****_The exclusions are made in the order of which the parameters are listed, and the markers are only reported under the reason for the first exclusion. This explains why no markers are listed under this requirement. However, individual results exclusions have been made on the basis of this requirement._

The table reveals that Coverage <50 is the mayor reason for exclusions in this project. For the results based on Nordic databases, 41 SNPs have been excluded for all cases, 39 Ancestry markers and two Identity markers. For the results based on the East-African databases, 48 SNPs were removed. SNPs were also excluded from individual cases if they did not meet the requirements given in Table 10. Originally, the Identity and Ancestry panels include 255 autosomal SNPs in total. After exclusion of both the SNPs listed in Table 11 and individual results not meeting the requirements given in Table 10, calculations for the main results were based on 169-208 SNPs in the different cases.

3.2.2 Coverage

The results for the Ancestry panel achieved an overall average coverage of 188 reads, while for Identity the average (for autosomal markers) were 261 reads.

The ANOVA test gave $p<0.05$ for both Identity and Ancestry, which implies a significant difference in coverage between at least two samples. Furthermore, the t-test with Bonferroni corrected significance level (4.17E-04) rejected H_0 for about 50% of the sample pairs for both panels. This indicated that the coverage values for the markers were dependent on which sample these were obtained from. Due to this finding and with respect to minimize outlier values, the coverage values were normalised in Excel. First, coverage result data were sorted by sample and the mean coverage for each sample was found. All individual coverage values were divided by the mean coverage for the accurate sample and multiplied by 100. Thus, coverage values found below the mean for the sample resulted in normalised values <100, and the opposite counted for coverage values found above the mean for the sample. Boxplots for all SNP marker were then constructed in R, each based on 16 normalised values (one from each sample). These can be found in Figure 15 (Identity), Figure 16 and Figure 17 (Ancestry). Furthermore, equivalent boxplots can be found for the non-normalised coverage values in Appendix D. The values in these boxplots are much more spread, and it is more difficult to say something over-all about the performance of the markers based on these. On the other side, the actual coverage values can be found on y-axis in these plot, and the SNPs found with the lowest coverage in this plot are the SNPs that have been excluded (Table 11). Moreover, minor differences can be seen in terms of the marker order in the normalized- and the non-normalised plots, but the same tendency is observed in terms of which markers perform poorest.

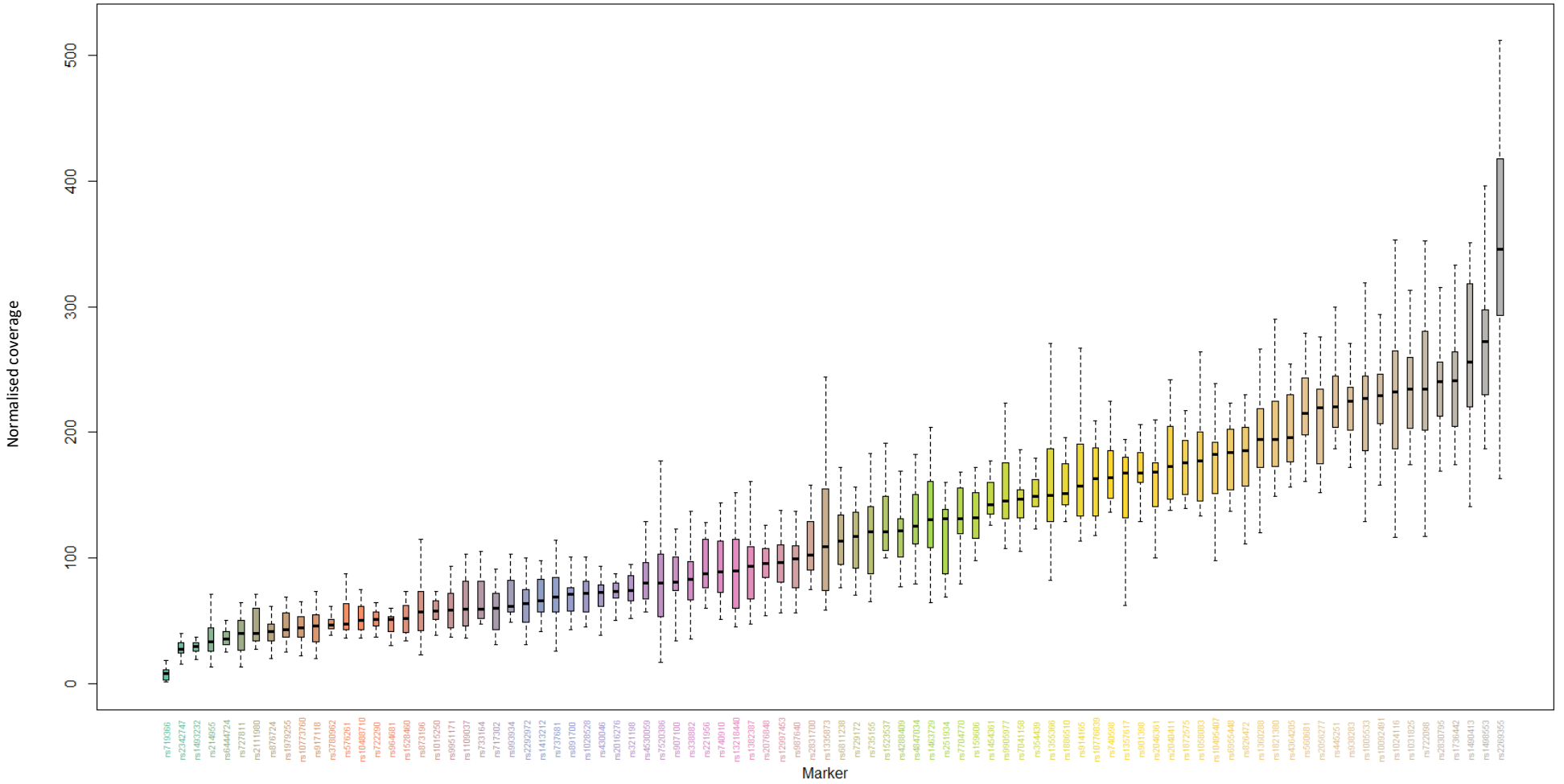


Figure 15: *Boxplot illustrating how the different (autosomal) Identity SNPs perform compared to each other in terms of coverage. The box for each marker is based on 16 normalised coverage values (in terms of which sample they were obtained from). The normalised values were found by dividing the individual coverage values by the mean coverage for the accurate sample and multiplying this by 100. Thus, SNPs with an average normalised coverage value >100 perform better than the average SNP.*

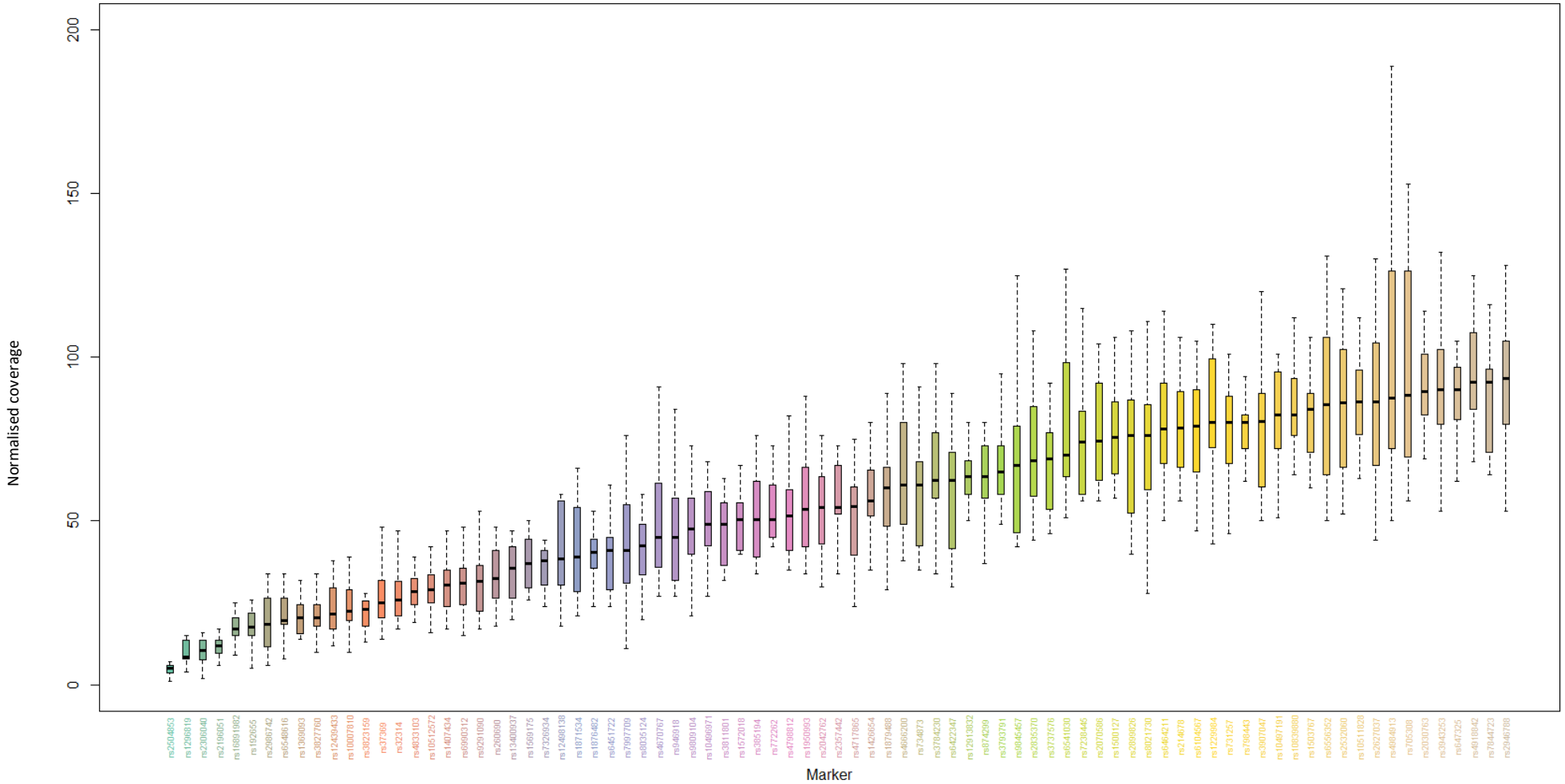


Figure 16: *Boxplot illustrating how 83 Ancestry SNPs perform compared to each other in terms of coverage (see Figure 17 for the remaining 82 Ancestry SNPs). The box for each marker is based on 16 normalised coverage values (in terms of which sample they were obtained from). The normalised values were found by dividing the individual coverage values by the mean coverage for the accurate sample and multiplying this by 100. Thus, SNPs with an average normalised coverage value >100 perform better than the average SNP.*

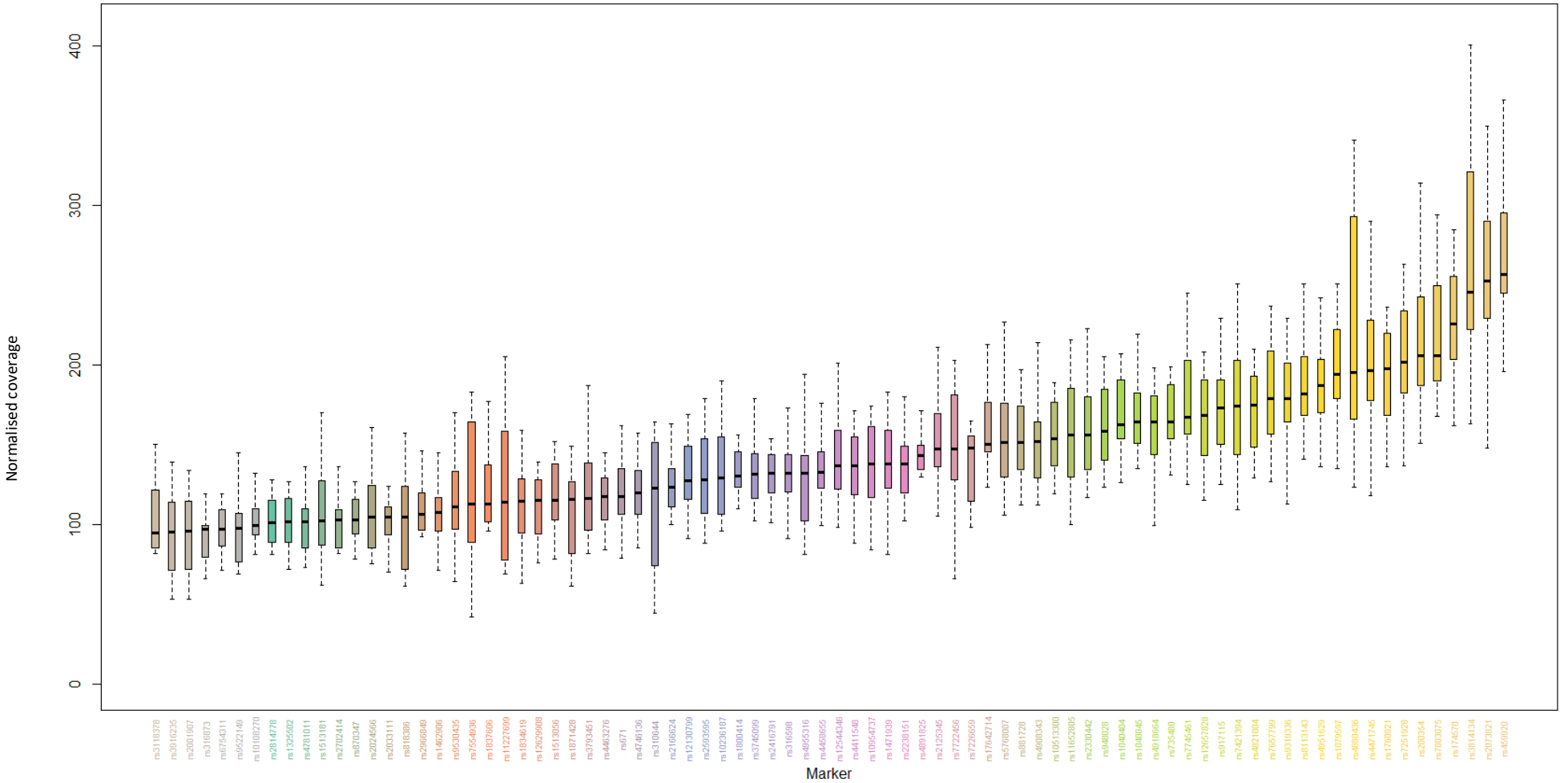


Figure 17: *Boxplot illustrating how 82 Ancestry SNPs perform compared to each other in terms of coverage (see Figure 16 for the remaining 83 Ancestry SNPs). The box for each marker is based on 16 normalised coverage values (in terms of which sample they were obtained from). The normalised values were found by dividing the individual coverage values by the mean coverage for the accurate sample and multiplying this by 100. Thus, SNPs with an average normalised coverage value >100 perform better than the average SNP.*

3.2.3 Ancestry

The results generated with GenoGeographer can be found in Appendix C. The analyses were performed to investigate if it was proper to apply a Nordic allele frequency database in the subsequent LR calculations. All samples were found most likely to originate from Europe, except sample 47 which achieved a slightly greater probability of originating from the Middle-East, compared to Europe. However, a Nordic allele frequency database was found fit in this project. This assumption is further reviewed in the discussion.

3.3 Project cases

3.3.1 Nordic allele frequency databases

In Table 12 the LR results using the Nordic allele frequency databases in the calculations are presented.

Table 12: *LRs for the eight project cases with use of the different marker panels/combinations of these. The calculations are based on Nordic allele frequency databases. H_1 correspond to the main hypothesis-, and H_2 to the alternative hypothesis in the case.*

Case (samples)	H_1	H_2	STR	Identity	Ancestry	STR +Identity	STR +Ancestry	STR+Identity +Ancestry
1 (34+35)	Half sibling	Unrelated	1064.4	236.6	116.8	2.51E+05	1.24E+05	2.94E+07
2 (36+37)	Half sibling	Unrelated	100.3	48.7	58.7	4877.8	5881.2	2.86E+05
3 (38+39)	Half sibling	Unrelated	119.7	0.2	23.3	24.2	2783.1	561.6
4 (40+41)	Grandparent	Unrelated	24.3	2.1	1.9	50.8	46.2	96.7
5 (42+43)	Full sibling	Half siblings	33.7	3053.2	225.9	1.03E+05	7600.3	2.32E+07
6 (44+45)	Half sibling	Unrelated	8.9	11.3	421.8	100.3	3745.2	4.23E+04
7 (46+47)	Full sibling of parent	Unrelated	4.8	296.0	9.5	1405.0	45.3	1.34E+04
8 (48+49)	Half sibling of parent	Unrelated	0.6	2.0	6.0	1.1	3.4	6.8

The results given in Table 12 are visually presented in Figure 18.

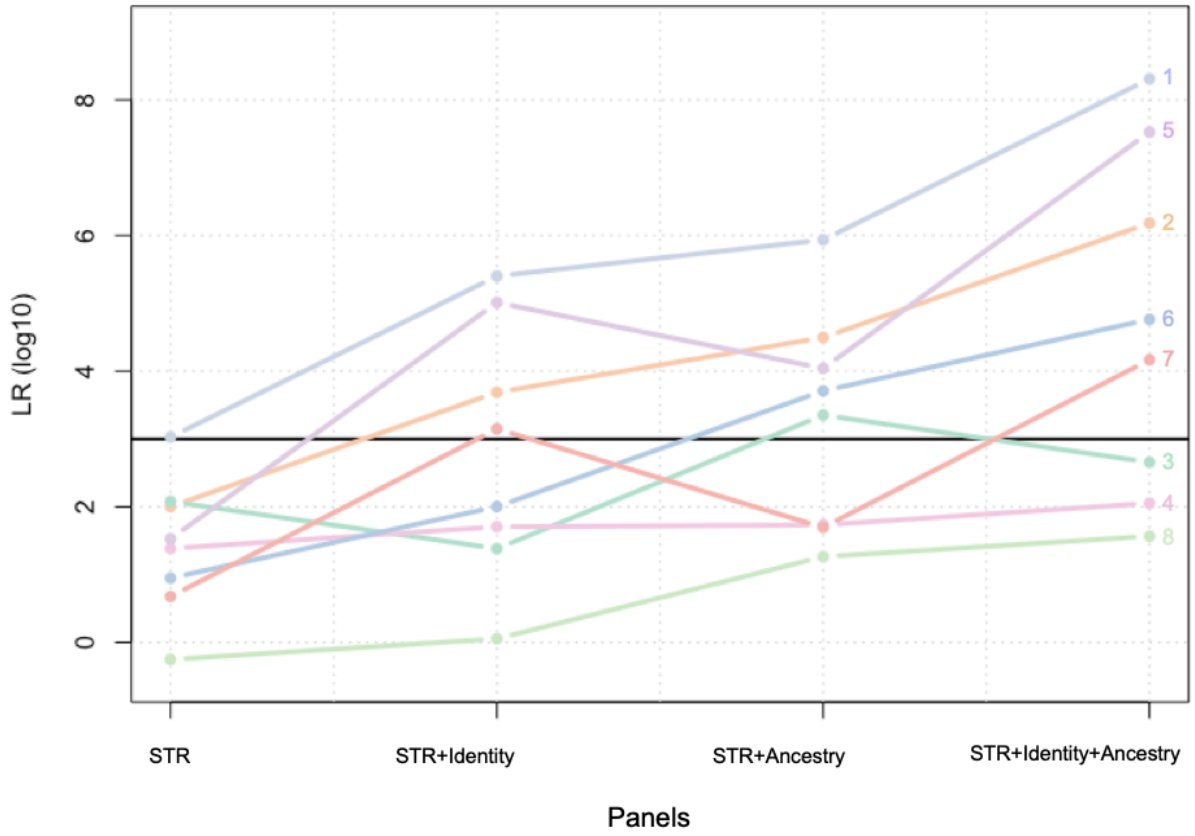


Figure 18: *LR results with use of different combinations of panels. Each line represents one case and each point represent the LR achieved for the different combinations of panels included in the calculations (x-axis). The case numbers are given in the end of each line. Lines crossing the black horizontal line illustrate LRs exceeding 1000 (log10(1000) = 3). The LRs are given as log10 (y-axis) to collate the values and make them presentable.*

The effect of supplementing STR panels with SNP panels is clearly illustrated. When the eight cases were analysed with STRs alone, one of them barely exceeded the LR threshold of 1000. The number of cases exceeding LR=1000 increases to four when Identity is added as supplement, and to five cases when all panels are applied.

Based on the guidelines in Table 2, case 1, 2, 5, 6 and 7 are found in, or above, the conclusive LR range, corresponding to (at least) great evidence weight and a 99.9% probability that H_1 is true (when Bayes Theorem is applied as described in 1.6.6). Case 3 are found in the interval comprising moderate weight and a 99% probability for H_1 being true. That is, no conclusion can be stated, but indications can be made that H_1 are the most likely hypothesis. Furthermore, case 4 and 8 are found in the interval for inclusive results.

3.3.2 East-African allele frequency databases

The results for the calculations based on the East-African allele frequency databases are presented in Table 13. This shows an LR >1000 for all cases with use of both SNP- and STR panels. Compared to the Nordic database results, the greatest difference in LR are seen for the Ancestry results which are in average 10 million times as high for the East African results. In respect to the Identity- and STR results, the East-African database gives LRs in average 100 times higher compared to the Nordic equivalents.

Table 13: *LRs for the eight project cases with use of the different marker panels/combinations of these. The calculations are based on East-African allele frequency databases. H_1 correspond to the main hypothesis-, and H_2 to the alternative hypothesis in the case.*

Case (samples)	H_1	H_2	STR	Identity	Ancestry	STR +Identity	STR +Ancestry	STR+Identity +Ancestry
1 (34+35)	Half sibling	Unrelated	8758.7	9.27E+04	6.37E+09	7.20E+07	5.58E+13	5.17E+18
2 (36+37)	Half sibling	Unrelated	2109.9	2.46E+04	8.33E+08	3.15E+06	1.76E+12	4.32E+16
3 (38+39)	Half sibling	Unrelated	552.8	2.7	2.22E+06	208.6	1.23E+09	3.30E+09
4 (40+41)	Grandparent	Unrelated	1.11E+04	30.7	4.32E+05	1190.7	4.80E+09	1.47E+11
5 (42+43)	Full sibling	Half siblings	190.5	3.53E+05	4.19E+07	1.66E+07	7.97E+09	2.81E+15
6 (44+45)	Half sibling	Unrelated	527.4	113.6	6.68E+09	1362.7	3.52E+12	4.00E+14
7 (46+47)	Full sibling of parent	Unrelated	931.5	7947.3	5.52E+07	5.84E+04	5.14E+10	4.09E+14
8 (48+49)	Half sibling of parent	Unrelated	3.1	10.8	3.30E+04	7.5	1.01E+05	1.09E+06

3.4 Simulations

The results from the simulations performed in Familias are given for the different relationships in Table 14, Table 15 and Table 16 respectively. As described before, half sibling, grandparent and full sibling of parent follow the same inheritance pattern and will all be referred to as half siblings in this section.

Table 14: *Median LR and proportion of simulated kinship cases exceeding LRs of 100, 1000 and 10,000. Simulation results for full sibling vs. half sibling, where full siblings are the true relationship. Pr(LR>1000) are marked in red as this is the LR limit for conclusive cases. 10,000 simulations were performed, based on Nordic allele frequency databases.*

Full sibling vs. Half sibling

Panel	Median LR	Pr(LR>100)	Pr(LR>1000)	Pr(LR>10,000)
STR	1217.0	72.6%	52.0%	31.6%
STR+Identity	1.33E+05	91.9%	82.4%	69.2%
STR+Ancestry	2.16E+05	92.6%	84.3%	71.3%
STR+Identity+Ancestry	2.18E+07	98.1%	95.5%	90.5%

Table 15: *Median LR and proportion of simulated kinship cases exceeding LRs of 100, 1000 and 10,000. Simulation results for half sibling-, grandparent- and full sibling of parent vs. unrelated, where one of the mentioned relationships are the true. Pr(LR>1000) are marked in red as this is the LR limit for conclusive cases. 10,000 simulations were performed, based on Nordic allele frequency databases.*

Half sibling vs. Unrelated

Panel	Median LR	Pr(LR>100)	Pr(LR>1000)	Pr(LR>10,000)
STR	510.2	66.7%	43.0%	22.5%
STR+Identity	1.05E+04	86.0%	70.1%	50.5%
STR+Ancestry	3.77E+04	89.8%	78.1%	60.7%
STR+Identity+Ancestry	6.30E+05	95.5%	88.9%	78.5%

Table 16: *Median LR and proportion of simulated kinship cases exceeding LRs of 100, 1000 and 10,000. Simulation results for half sibling of parent vs. unrelated, where half sibling of parent is the true relationship. Pr(LR>1000) are marked in red as this is the LR limit for conclusive cases. 10,000 simulations were performed, based on Nordic allele frequency databases.*

Half sibling of parent vs. Unrelated

Panel	Median LR	Pr(LR>100)	Pr(LR>1000)	Pr(LR>10,000)
STR	5.1	8.8%	1.5%	0.2%
STR+Identity	11.1	18.3%	4.0%	0.7%
STR+Ancestry	15.3	23.2%	6.0%	1.3%
STR+Identity+Ancestry	31.2	33.8%	11.5%	2.8%

Based on the results, 52.0% of cases including (true) full siblings, are expected to achieve an LR >1000 when the alternative hypothesis is half siblings, based on only STRs (Table 14). When the Identity and Ancestry SNPs are added to the analyses, the proportion of cases increases to 95.5%. For half sibling vs. unrelated this corresponds to an increase from 43.0% to 88.9% (Table 15), and for half sibling of parent vs. unrelated the amount increase from 1.5% to 11.5% (Table 16). The results given in the tables above are graphically illustrated in Figure 19.

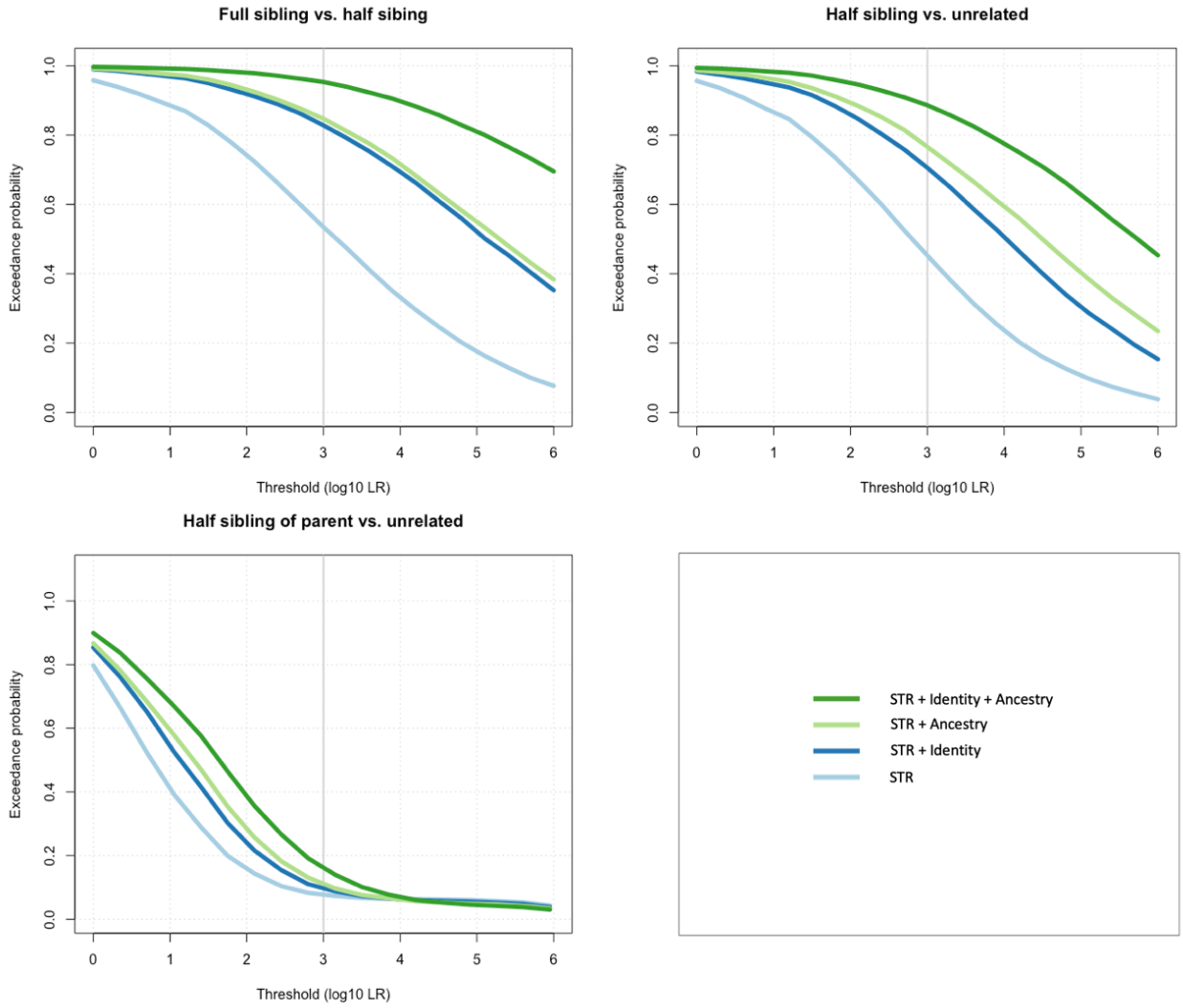


Figure 19: *Graphically illustrated simulation results for the different relationship cases, where the first mentioned relationship in each case is the true. Each line represents the proportion of simulated cases (y-axis) exceeding the LR thresholds given by the x-axis. The grey, vertical lines represent LR=1000. Each relationship was simulated 10,000 times, and the LRs are given as log10 to collate the values and make them presentable.*

When all panels are applied, the number of cases exceeding an LR of 1000 increases for all relationships. STR+Ancestry are found slightly above STR+Identity in all three graphs. The graph illustrating simulations of half sibling of parent vs. Unrelated reveals that only a small part of the total cases achieves conclusive results, even when all panels are applied.

3.5 Blind search

In Table 17, results from the Familias blind search is reported for all project cases.

Table 17: *Results from the Familias blind search for the eight project cases. LR and Pr(IBD=0, 1 and 2) are reported for the relevant relationships. LR were calculated with unrelated as alternative hypothesis for all cases except case 5 where the alternative hypothesis was half sibling. Nordic allele frequency data bases are used in the calculations. The IBD probabilities were calculated with a previously explained maximum likelihood approach.*

Case (Samples)	Alleged relationship (H_1)	Pr(IBD=0)	Pr(IBD=1)	Pr(IBD=2)	LR
1 (34+35)	Half sibling	0.44	0.45	0.11	2.94E+07
2 (36+37)	Half sibling	0.46	0.54	0.00	2.86E+05
3 (38+39)	Half sibling	0.63	0.33	0.04	561.6
4 (40+41)	Grandparent	0.63	0.37	0.00	96.7
5 (42+43)	Full sibling	0.10	0.68	0.22	2.32E+07
6 (44+45)	Half sibling	0.53	0.45	0.02	4.23E+04
7 (46+47)	Full sibling of parent	0.55	0.45	0.00	1.34E+04
8 (48+49)	Half sibling of parent	0.74	0.26	0.00	6.8

The LRs obtained from the blind search are confirmed to be identical with the already calculated LRs for the hypotheses of the project cases.

The maximum likelihood estimates of the IBD probabilities in case 1, 2, 6 and 8 show values close to what is expected for the alleged relationships in these cases (explained in 1.6.3). This is also reflected by the higher LRs in these cases, except in case 8. Furthermore, case 5 fits the expected values for full siblings, but shows a slightly elevated Pr(IBD=1) and lower Pr(IBD=0), which is also reflected in the LR. Case 3 and 4 deviate more from what is expected for the alleged relationships, with slightly elevated Pr(IBD=0) and reduced Pr(IBD=1), which is again reflected in the LR.

3.6 FamLink

The LR results calculated with FamLink and Familias, and the relative differences between these, are given in Table 18.

Table 18: *Different LRs are achieved for calculations performed in FamLink and Familias. These are given for both methods in the table. The relative difference was found by dividing the LR from FamLink by the LR from Familias. A relative difference of 1 corresponds to identical LR for both methods. Relative difference >1 corresponds to a higher LR from FamLink, while the opposite counts for relative difference <1.*

Case (samples)	H_1	H_2	FamLink LR	Familias LR	Relative difference
1 (34+35)	Half sibling	Unrelated	2.86E+07	2.94E+07	0.97
2 (36+37)	Half sibling	Unrelated	1.27E+06	2.86E+05	4.44
3 (38+39)	Half sibling	Unrelated	1706.0	561.6	3.04
4 (40+41)	Grandparent	Unrelated	1816.0	96.7	18.78
5 (42+43)	Full sibling	Half sibling	2.50E+06	2.32E+07	0.11
6 (44+45)	Half sibling	Unrelated	1.53E+05	4.23E+04	3.62
7 (46+47)	Full sibling of parent	Unrelated	2.87E+04	1.34E+04	2.14
8 (48+49)	Half sibling of parent	Unrelated	7.4	6.8	1.09

For most cases, the LR increased when the calculations were performed in FamLink, compared to Familias. However, the LR for case 5 decreased notably. This is the only case were unrelated is not the alternative hypothesis, which is further reviewed in the discussion.

4 Discussion

In this project the value of supplementing STRs with SNPs in complex kinship cases has been investigated. The Familias software developed by Egeland et al. (2000) and Kling et al. (2014) was used to calculate likelihood ratios (LRs), and it was observed that the number of cases with LR >1000 increased from one when only STRs were applied, to five when STRs and SNPs were applied in the calculations (Table 12 and Figure 18). This supports what is earlier reported by several others, that more conclusive results are achieved in forensic DNA analysis when the traditional STR panels are supplemented by SNP panels (Amorim and Pereira 2005; Sanchez et al. 2006; Børsting et al. 2012; Pontes et al. 2015). Furthermore, this is corroborated by the simulations performed in this project (Table 14, Table 15 and Table 16). Case 8, which concerns a relationship more distant than the rest of the cases, achieved the lowest LR. As described in 1.6.3, child-half sibling of parent has a 0.75 probability of sharing 0 alleles Identical by decent (IBD) for each marker, and cases concerning such relationships are not currently accepted at REFA. Moreover, the simulation results show that only 11.5% of the cases with (true) child-half sibling of parent were expected to result in an LR >1000 when all panels were applied (Table 16). This implies that a larger or better suited panel are needed to solve cases like this. As a consequence, in cases where the proposed relationship is this distant, one should try to obtain samples from multiple- or closer relatives, rather than to supplement with the SNP panels investigated in this project.

The results generated with the GenoGeographer tool developed by Tvedebrink et al. (2017), pointed in the direction of European ancestry for all sample donors, except for individual 47 (Appendix C). The genotypes in this profile was found slightly more common in the Middle-East, compared to Europa, perhaps as a consequence to admixed ancestry. In the article by Tvedebrink et al. (2017), it is stated that there is a general need for an improvement of the GenoGeographer, especially with regard to calculations of admixed descent. With this being said, the GenoGeographer results should only be considered as an indication and knowledge of where the samples are obtained from should be included in the evaluations. Nordic databases were found fit in this project. The importance of basing LR calculations on appropriate allele frequency databases was emphasised in the introduction. Furthermore, comparison of the LRs based on Nordic- and East-African allele frequency databases (Table 12 and Table 13, respectively) illustrates that major differences is observed. Far higher LRs

are seen for all cases when they are based on the East-African databases, probably as a consequence of the individuals sharing alleles frequently occurring in Norway, but rarely in Africa. The greatest deviation was seen for the Ancestry SNPs were LRs in average were 10 million times as high for the East-African results, compared to the Nordic results. For Identity and STR, the values were in average 100 times as high. A possible reason is that the ancestry-sensitivity of the markers are different for the panels applied in this project. The Ancestry panel is developed with the purpose of inferring ancestry, i.e. the allele frequencies for the markers should vary considerably between different geographic areas (Kosoy et al. 2009; Kidd et al. 2014). The STR panels and the Identity panel are developed with the purpose of identifying, and the markers are selected minimizing the ancestry-sensitivity (Pakstis et al. 2010; Westen et al. 2012; Ensenberger et al. 2016). The results from this project indicate that overestimation of LRs is a general risk when Ancestry SNPs are applied in relationship calculations. In terms of further development of a SNP panel with the purpose to identify or establish relationships, the results show that ancestry-insensitive markers, such as Identity SNPs, are most suitable.

As an alternative method to the traditional LR approach, estimates of Pr(IBD=0), Pr(IBD=1) and Pr(IBD=2) can be evaluated to infer the most likely relationship for two individuals. This method differs from how LRs are calculated, where two hypotheses, and IBD probabilities according to the claimed relationships, are set prior to the calculations. Evaluation of IBD probabilities is not a part of the standard routine at REFA, as the cases often include specific questions with only two relevant alternatives, making the standard LR method the preferred. However, in cases with several possible alternative hypotheses, it can be advantageous to evaluate the IBD probabilities independent of pre-set hypotheses, e.g. in paternity cases where it is possible that the true father is a relative to the alleged father. When IBD probabilities from the project results were evaluated, it was discovered that when these fitted with the expected probabilities, the cases generally also had a high LR (Table 17). Case 5 achieved a high LR despite a slightly elevated Pr(IBD=1) and lower Pr(IBD=0), compared to the expected for full siblings (Table 1). This could be a result of the alternative hypothesis, where the individuals are half siblings, which has a worse fit for the expected IBD probabilities than for full siblings (main hypothesis). Case 8 differs from the other cases, as the IBD probabilities fits very well with the expected, but still the case has achieved a low LR. This supports what has already been suggested, that relationships this distant may not be resolved, with the currently applied markers. Moreover, none of the project cases showed exact IBD

distribution according to what was excepted for the different relationships. Deviations are expected as the number of markers applied in the calculations were limited, and the IBD probabilities (for cases where the alleged relationship is true) would probably fit better with the true relationship if more markers were added to the analyses. Moreover, the investigated individuals were from real cases, and the only time the true relationships can be known with 100% certainty is when the cases are simulated. Half siblings can for instance be related beyond this relationship because of an additional common ancestor far back in time. This can e.g. occur if the individuals are half siblings on the mother's side and second cousins on the father's side, which can elevate the IBD probabilities for sharing alleles above what is expected for half siblings.

The probability of linkage increases when large marker panels are applied in DNA analyses, and traditionally, scepticism has been associated with use of linked markers in forensic genetics. This is reflected by the commonly used STRs, which are found spread on the chromosomes (Hares 2012). FamLink developed by Kling et al. (2012a) was used to calculate an additional, linkage-adjusted LR for all cases in this project (Table 18). Generally, the LR increased when linkage was accounted for, indicating a general underestimation or conservative weight of the evidence. Furthermore, the probabilities for a (true) relationship will generally increase when linkage is accounted for (Gill et al. 2012). Related individuals are expected to share the same number of alleles in closely linked markers, resulting in an increased LR compared to what would be achieved for matching alleles in unlinked markers. When individuals are unrelated (alternative hypothesis in 7 of the project cases), there are no expectations in terms of matching alleles in the linked markers. Thus, the probability for this hypothesis will not be affected by the fact that linkage is accounted for. However, this is a general explanation, and the exact impact of linkage is affected by several factors, e.g. allele frequencies, recombination rates and genotype constellations (Gill et al. 2012). Case 5 was the only case with a notably decrease in LR after linkage was taken into account. This is also the only case where the alternative hypothesis does not state that the individuals are unrelated. If these individuals are related, the probability for both the main- and the alternative hypothesis will generally increase, unlike for the other cases, which could be the reason for the decreased LR. Moreover, LR were found above 1000 for both the Familias- and FamLink results for this case, i.e. the conclusion is not affected by the change. Nevertheless, accounting for linkage

was crucial in terms of the conclusion in case 3 and 4, indicating that if a large number of markers are to be applied in routine work, linkage should be accounted for.

All markers included in this thesis were checked and possibly removed with respect to linkage disequilibrium (LD). It was desirable to check for actual LD between all markers, but as several of the markers were missing from the SNPsnap application by Machiela and Chanock (2015), a distance of 0.5 cM was applied as a threshold instead (Table 11) (Kling et al. 2012b). Disregarding the limiting capacity of this project it would have been appropriate to calculate the degree of LD between all markers if a new panel is to be constructed.

REFA does not have a procedure for the Ion Torrent™ method, and a manual punching- and extraction method was found appropriate in this project. Several studies have investigated different methods for library preparation and found the results from automatic workflows performed by the Ion Chef™ Instrument at least as good as for the manual workflows (Mogensen et al. 2015; van der Heijden et al. 2017). However, the automatic library preparation only allows eight samples to be prepared at a time, i.e. another method may be preferred in routine laboratories where a great number of samples are being handled. This project included only 16 samples, and the Ion Chef™ Instrument was an appropriate choice for achieving results with small variations.

Experience at REFA has shown that DNA extracted with the method used in this project are not measurable with the available equipment in the laboratory where the project samples were analysed. This was assumed to be a consequence to low DNA concentrations, and 25 cycles (recommended for weak samples of <1 ng DNA) were applied for the PCRs in the library preparations. The high cycle number could potentially have a negative impact on samples with normal and high DNA concentration, but as a trade-off it reduces the chance of weak samples not achieving results at all. This latter mentioned approach was prioritised in this project, and it cannot not be ruled out that some of the PCRs might have been sub-optimal, which could further have led to polyclonal beads and a lower number of reads (1.5.2). Thus, disregarding the limitations of the project, it would be advantageous to measure the DNA concentration for all DNA extracts in order to optimize the emPCR and possibly achieve better results with respect to coverage. This might lead to less result exclusions without this compromising the credibility of the analyses, which is even more relevant if the method is to be used in routine work, where more conservative thresholds most likely will be applied. Furthermore, it is reasonable to consider the maximum loaded 314 chips as a source of

improvement in terms of coverage in this project. In a study by Guo et al. (2016) it is shown that satisfactory results are achieved by applying six samples on the 314 chip, while Børsting et al. (2014) suggest that no more than four samples should be loaded onto this chip. Thus, it is reasonable to assume that better quality results would be achieved in this project by replacing the 314 chips with 316 chips, constituting a fivefold increase with regard to the number of sensor wells (Merriman et al. 2012). This is also supported by the fact that Identity (autosomal markers), where the total number of reads are divided on a smaller number of markers, has achieved a higher average coverage compared to Ancestry, 261 versus 188 reads respectively. The theoretical total number of reads for the 314 chip is 1.2 million, which amounts to a per marker number of 1250 reads for Identity and 909 for Ancestry. However, one cannot expect 100% loading of the chip, 0% polyclonal reads and 0% low quality reads. In this project the total number of reads per chip was roughly 270,000 (23% of the theoretical number). The lower read numbers for the Ancestry panel could also be a possible explanation for the predominance of Ancestry SNPs among the excluded markers.

Standard thresholds for the analytical parameters applied in the Ion Torrent™ analyses in this project does not exist. The applied thresholds were obtained from an average of what was used as in similar projects, this is described in 2.4.5. Disregarding the limiting capacity of this project it would be appropriate to analyse a larger dataset and evaluate which thresholds would be optimal for the purpose investigated in this project.

A positive- and negative control was not included in the sequencing analyses in this project. This was partly due to cost and partly due to the purpose of the project. The applied 314 chips were the only chips available in this project, and it was prioritised to include all the 16 samples rather than to include controls and duplicates. However, this needs to be addressed if the method is to be implemented in future routine work, where controls are a necessity. Nevertheless, the extracts in this project were controlled by STR analysis, confirming that the right results are reported for the correct individuals.

5 Conclusion and future directions

The main aim of this thesis was to investigate if SNPs could be used as supplementary markers to STRs and lead to more conclusive results in complex kinship cases, where STRs alone were not sufficient. The results from both the real cases and the simulations showed a notable decrease in inconclusive cases when SNPs were included in the analyses. However, throughout this project it has been revealed that there are several important aspects that can affect the final conclusion in complex kinship cases, especially when a large number of markers are applied. These aspects must be taken into account and be further investigated if the procedure of supplementing with SNPs is to be used in routine work. First, a thorough evaluation of a large number of SNP markers should be made to find those that are best suited for use with complex kinship cases. The evaluation should take into account that the selected markers should be ancestry-insensitive, not in LD with each other or the STR markers and linkage should be calculated and included in the analyses. It was also suggested that larger- or better suited panels are needed to solve cases where the proposed relationship is half sibling of parent or equivalent. Additionally, it is shown in this project that the use of the correct allele frequency databases in the calculations is crucial, especially if ancestry-sensitive markers are applied. This further emphasizes the importance of using ancestry-insensitive markers when samples of unknown ancestry are analysed. It will probably be appropriate to use a larger sequencing chip in addition to perform concentration measurements and dilute accordingly prior to the sequencing. This can have several advantages, firstly, one will probably be able to achieve better quality results with respect to coverage. Secondly, a larger chip could enable inclusion of analysis controls and duplicate samples, i.e. provide more credibility to the analyses. Finally, the conclusion is that SNPs are well suited as supplement to STRs in complex kinship cases, but that further investigations should be performed in respect to constructing a panel and a procedure that is suitable for routine work.

References

Aldrich, J. (1997). "RA Fisher and the making of maximum likelihood 1912-1922." Statistical science **12**(3): 162-176.

Amorim, A. and Pereira, L. (2005). "Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs." Forensic science international **150**(1): 17-21.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L. and Bignell, H. R. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53.

Bethesda (MD): National Center for Biotechnology Information (US). (2005). "SNP FAQ Archive." Retrieved 01/09/19, from https://www.ncbi.nlm.nih.gov/books/NBK3848/.

Bonferroni, C. (1936). "Teoria statistica delle classi e calcolo delle probabilita." Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze **8**: 3-62.

Buchard, A., Kampmann, M. L., Poulsen, L., Børsting, C. and Morling, N. (2016). "ISO 17025 validation of a next-generation sequencing assay for relationship testing." Electrophoresis **37**(21): 2822-2831.

Buermans, H. and Den Dunnen, J. (2014). "Next generation sequencing technology: advances and applications." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease **1842**(10): 1932-1941.

Butler, J. M. (2005). Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Elsevier.

Butler, J. M., Coble, M. D. and Vallone, P. M. (2007). "STRs vs. SNPs: thoughts on the future of forensic DNA testing." Forensic science, medicine, and pathology **3**(3): 200-205.

Børsting, C., Fordyce, S. L., Olofsson, J., Mogensen, H. S. and Morling, N. (2014). "Evaluation of the Ion Torrent™ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing." Forensic Science International: Genetics **12**: 144-154.

Børsting, C., Mikkelsen, M. and Morling, N. (2012). "Kinship analysis with diallelic SNPs–experiences with the SNPforID multiplex in an ISO17025 accreditated laboratory." Transfusion Medicine and Hemotherapy **39**(3): 195-201.

Cisana, S., Cerri, N., Bosetti, A., Verzeletti, A. and Cortellini, V. (2017). "PowerPlex® Fusion 6C System: evaluation study for analysis of casework and database samples." Croatian medical journal **58**(1): 26-33.

Consortium, I. H. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851.

Council, N. R. (1996). The evaluation of forensic DNA evidence, National Academies Press.

Daniel, R., Santos, C., Phillips, C., Fondevila, M., Van Oorschot, R., Carracedo, A., Lareu, M. and McNevin, D. (2015). "A SNaPshot of next generation sequencing for forensic SNP analysis." Forensic Science International: Genetics **14**: 50-60.

David, H. A. and Nagaraja, H. N. (2004). "Order statistics." Encyclopedia of Statistical Sciences.

De Muth, J. E. (2014). Basic statistics and pharmaceutical statistical applications, Chapman and Hall/CRC.

Dupuy, B. M., Stenersen, M. and Kling, D. (2013). "Frequency data for 35 autosomal STR markers in a Norwegian, an East African, an East Asian and Middle Asian population and simulation of adequate database size." Forensic Science International: Genetics Supplement Series **4**(1): e378-e379.

Egeland, T., Kling, D. and Mostad, P. (2015). Relationship inference with familias and R: statistical methods in forensic genetics, Academic Press.
Egeland, T., Mostad, P. F., Mevåg, B. and Stenersen, M. (2000). "Beyond traditional paternity and identification cases: selecting the most probable pedigree." Forensic science international **110**(1): 47-59.
Ensenberger, M. G., Lenz, K. A., Matthies, L. K., Hadinoto, G. M., Schienman, J. E., Przech, A. J., Morganti, M. W., Renstrom, D. T., Baker, V. M. and Gawrys, K. M. (2016). "Developmental validation of the PowerPlex® fusion 6C system." Forensic Science International: Genetics **21**: 134-144.
Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis; theoretische Grundlagen. Wien, Mitteilungen der Anthroposophischen Gesellschaft
Fan, H. and Chu, J.-Y. (2007). "A brief review of short tandem repeat mutation." Genomics, Proteomics & Bioinformatics **5**(1): 7-14.
Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **222**(594-604): 309-368.
Gill, P. (2001). "An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes." International Journal of Legal Medicine **114**(4-5): 204-210.
Gill, P., Phillips, C., McGovern, C., Bright, J.-A. and Buckleton, J. (2012). "An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees." Forensic Science International: Genetics **6**(4): 477-486.
Gill, P., Werrett, D. J., Budowle, B. and Guerrieri, R. (2004). "An assessment of whether SNPs will replace STRs in national DNA databases--joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM)." Science & justice: journal of the Forensic Science Society **44**(1): 51.
Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., Lessig, R., Mayr, W. R., Pascali, V. L. and Prinz, M. (2007). "ISFG: recommendations on biostatistics in paternity testing." Forensic Science International: Genetics **1**(3-4): 223-231.
Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016). "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics **17**(6): 333.
Gooijer, C., Kok, S. and Ariese, F. (2000). "Capillary electrophoresis with laser-induced fluorescence detection for natively fluorescent analytes." Analusis **28**(8): 679-685.
Gray, I. C., Campbell, D. A. and Spurr, N. K. (2000). "Single nucleotide polymorphisms as tools in human genetics." Human molecular genetics **9**(16): 2403-2408.
Guo, F., Zhou, Y., Song, H., Zhao, J., Shen, H., Zhao, B., Liu, F. and Jiang, X. (2016). "Next generation sequencing of SNPs using the HID-Ion AmpliSeq™ Identity Panel on the Ion Torrent PGM™ platform." Forensic Science International: Genetics **25**: 73-84.
Hares, D. R. (2012). "Expanding the CODIS core loci in the United States." Forensic Science International: Genetics **6**(1): e52-e54.
Hütt, M. T. (2014). "Understanding genetic variation–the value of systems biology." British journal of clinical pharmacology **77**(4): 597-605.
Karki, R., Pandya, D., Elston, R. C. and Ferlini, C. (2015). "Defining "mutation" and "polymorphism" in the era of personal genomics." BMC medical genomics **8**(1): 37.
Kayser, M. and De Knijff, P. (2011). "Improving human forensics through advances in genetics, genomics and molecular biology." Nature Reviews Genetics **12**(3): 179.

Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L., Karoma, N. J., Kungulilo, S., Kim, J.-J., Lu, R.-B. and Odunsi, A. (2006). "Developing a SNP panel for forensic identification of individuals." Forensic science international **164**(1): 20-32.

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R. and Kidd, J. R. (2014). "Progress toward an efficient panel of SNPs for ancestry inference." Forensic Science International: Genetics **10**: 23-32.

Kling, D., Egeland, T., Piñero, M. H. and Vigeland, M. D. (2017). "Evaluating the statistical power of DNA-based identification, exemplified by 'The missing grandchildren of Argentina'." Forensic Science International: Genetics **31**: 57-66.

Kling, D., Egeland, T. and Tillmar, A. O. (2012a). "FamLink–a user friendly software for linkage calculations in family genetics." Forensic Science International: Genetics **6**(5): 616-620.

Kling, D. and Füredi, S. (2016). "The successful use of familial searching in six Hungarian high profile cases by applying a new module in Familias 3." Forensic science international: genetics **24**: 24-32.

Kling, D., Tillmar, A. O. and Egeland, T. (2014). "Familias 3–Extensions and new functionality." Forensic Science International: Genetics **13**: 121-127.

Kling, D., Welander, J., Tillmar, A., Skare, Ø., Egeland, T. and Holmlund, G. (2012b). "DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects." Forensic Science International: Genetics **6**(3): 322-329.

Kosambi, D. D. (1944). The estimation of map distances from recombination values. DD Kosambi, Springer**:** 125-130.

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K. and Belmont, J. W. (2009). "Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America." Human mutation **30**(1): 69-78.

Krawczak, M. (1999). "Informativity assessment for biallelic single nucleotide polymorphisms." ELECTROPHORESIS: An International Journal **20**(8): 1676-1681.

Kwok, P.-Y. and Chen, X. (2003). "Detection of single nucleotide polymorphisms."

Lesk, A. M. (2017). Introduction to Genomics. Oxford, Oxford University Press.

Lipfert, J., Doniach, S., Das, R. and Herschlag, D. (2014). "Understanding nucleic acid–ion interactions." Annual review of biochemistry **83**: 813-841.

Machiela, M. J. and Chanock, S. J. (2015). "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants." Bioinformatics **31**(21): 3555-3557.

Mascher, M., Wu, S., Amand, P. S., Stein, N. and Poland, J. (2013). "Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley." PloS one **8**(10): e76925.

Matise, T. C., Chen, F., Chen, W., Francisco, M., Hansen, M., He, C., Hyland, F. C., Kennedy, G. C., Kong, X. and Murray, S. S. (2007). "A second-generation combined linkage–physical map of the human genome." Genome research **17**(12): 1783-1786.

Meiklejohn, K. A. and Robertson, J. M. (2017). "Evaluation of the Precision ID Identity Panel for the Ion Torrent™ PGM™ sequencer." Forensic Science International: Genetics **31**: 48-56.

Merriman, B., D Team, I. T. and Rothberg, J. M. (2012). "Progress in ion torrent semiconductor chip based sequencing." Electrophoresis **33**(23): 3397-3417.

Mogensen, H. S., Børsting, C. and Morling, N. (2015). "Template preparation of AmpliSeq™ libraries using the Ion Chef™." Forensic Science International: Genetics Supplement Series **5**: e368-e369.

Mäki, A., Rissanen, A. J. and Tiirola, M. (2016). "A practical method for barcoding and size-trimming PCR templates for amplicon sequencing." BioTechniques **60**(2): 88-90.
Nakano, M., Komatsu, J., Matsuura, S.-i., Takashima, K., Katsura, S. and Mizuno, A. (2003). "Single-molecule PCR using water-in-oil emulsion." Journal of biotechnology **102**(2): 117-124.
Nato, A., Buyske, S. and Matise, T. (2012). "The Rutgers map: A third-generation combined linkage-physical map of the human genome." Hum Genet Inst New Jersey Second Res Day.
Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R. and Kidd, K. K. (2010). "SNPs for a universal individual identification panel." Human genetics **127**(3): 315-324.
Pelt-Verkuil, E. v., Belkum, A. v. and Hays, J. P. (2008). Principles and technical aspects of PCR amplification. Dordrecht, Springer.
Pereira, V., Mogensen, H. S., Børsting, C. and Morling, N. (2017). "Evaluation of the Precision ID Ancestry Panel for crime case work: a SNP typing assay developed for typing of 165 ancestral informative markers." Forensic Science International: Genetics **28**: 138-145.
Phillips, C., Fang, R., Ballard, D., Fondevila, M., Harrison, C., Hyland, F., Musgrave-Brown, E., Proff, C., Ramos-Luis, E. and Sobrino, B. (2007a). "Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel." Forensic Science International: Genetics **1**(2): 180-185.
Phillips, C., Fernandez-Formoso, L., Gelabert-Besada, M., García-Magariños, M., Amigo, J., Carracedo, A. and Lareu, M. (2014). "Global population variability in QIAGEN investigator HDplex STRs." Forensic Science International: Genetics **8**(1): 36-43.
Phillips, C., Salas, A., Sanchez, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D. and Lareu, M. (2007b). "Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs." Forensic Science International: Genetics **1**(3-4): 273-280.
Pontes, M. L., Fondevila, M., Laréu, M. V. and Medeiros, R. (2015). "SNP markers as additional information to resolve complex kinship cases." Transfusion Medicine and Hemotherapy **42**(6): 385-388.
Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. and Gu, Y. (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC genomics **13**(1): 341.
Sala, C. and Verpelli, C. (2016). Neuronal and Synaptic Dysfunction in Autism Spectrum Disorder and Intellectual Disability, Academic Press.
Samuel, G. and Prainsack, B. (2018). "Forensic DNA phenotyping in Europe: views "on the ground" from those who have a professional stake in the technology." New Genetics and Society: 1-23.
Sanchez, J. J., Phillips, C., Børsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C. D., Musgrave-Brown, E., Salas, A. and Syndercombe-Court, D. (2006). "A multiplex assay with 52 single nucleotide polymorphisms for human identification." Electrophoresis **27**(9): 1713-1724.
Satya, R. V., Zavaljevski, N. and Reifman, J. (2011). "SNIT: SNP identification for strain typing." Source code for biology and medicine **6**(1): 14.
Schumm, J. W. (1997). "Why Use a Size Marker and Allelic Ladders in STR Analysis?" Profiles in DNA **1**: 11-13.
Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995). "An improved PCR method for walking in uncloned genomic DNA." Nucleic acids research **23**(6): 1087.

Sobrino, B., Brión, M. and Carracedo, A. (2005). "SNPs in forensic genetics: a review on SNP typing methodologies." Forensic science international **154**(2-3): 181-194.

Tillmar, A. O. and Phillips, C. (2017). "Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets." Forensic Science International: Genetics **26**: 58-65.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. and Morling, N. (2017). "GenoGeographer–A tool for genogeographic inference." Forensic Science International: Genetics Supplement Series **6**: e463-e465.

Urdan, T. C. (2011). Statistics in plain English, Routledge.

van der Heijden, S., de Oliveira, S. J., Kampmann, M.-L., Børsting, C. and Morling, N. (2017). "Comparison of manual and automated AmpliSeq™ workflows in the typing of a Somali population with the Precision ID Identity Panel." Forensic Science International: Genetics **31**: 118-125.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A. and Holt, R. A. (2001). "The sequence of the human genome." science **291**(5507): 1304-1351.

Westen, A. A., Haned, H., Grol, L. J., Harteveld, J., van der Gaag, K. J., de Knijff, P. and Sijen, T. (2012). "Combining results of forensic STR kits: HDplex validation including allelic association and linkage testing with NGM and Identifiler loci." International journal of legal medicine **126**(5): 781-789.

Wright, S. (1931). "Evolution in Mendelian populations." Genetics **16**(2): 97.

Zhang, S., Bian, Y., Chen, A., Zheng, H., Gao, Y., Hou, Y. and Li, C. (2017). "Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM." Forensic Science International: Genetics **27**: 50-57.

Appendix A – STR- and SNP markers with positions (cM)

All SNP- and STR markers sorted by chromosome and cM position. Only autosomal markers were used in the statistical calculations, i.e. markers located on the X- and Y-chromosomes are not listed in the table. Genetic positions (in cM) for both SNP- and STR markers were collected from Rutgers Map v.3 (Matise et al. 2007; Nato et al. 2012). Eight of the STR markers (D1S1656, SE33, D10S2325, D13S317, Penta E, D18S51, D21S11, Penta D) could not be obtained from this source and were interpolated from SNPs found very close on the chromosome.

No.	Chr	Marker ID	Marker type	Position cM	Panel
1	1	rs1490413	SNP	11.42	Identity
2	1	rs2986742	SNP	16.29	Ancestry
3	1	rs6541030	SNP	27.87	Ancestry
4	1	rs7520386	SNP	30.53	Identity
5	1	rs647325	SNP	39.40	Ancestry
6	1	rs4908343	SNP	54.21	Ancestry
7	1	rs1325502	SNP	73.36	Ancestry
8	1	rs12130799	SNP	83.49	Ancestry
9	1	rs3118378	SNP	102.08	Ancestry
10	1	rs3737576	SNP	130.01	Ancestry
11	1	rs4847034	SNP	133.43	Identity
12	1	rs7554936	SNP	153.39	Ancestry
13	1	rs2814778	SNP	161.79	Ancestry
14	1	rs560681	SNP	166.11	Identity
15	1	rs1040404	SNP	178.55	Ancestry
16	1	rs1407434	SNP	194.53	Ancestry
17	1	rs4951629	SNP	222.04	Ancestry
18	1	D1S1656	STR	237.5	Fusion 6C
19	1	rs10495407	SNP	257.38	Identity
20	1	rs891700	SNP	260.53	Identity
21	1	rs316873	SNP	267.72	Ancestry
22	1	rs1413212	SNP	269.24	Identity
23	2	rs876724	SNP	0.03	Identity
24	2	TPOX	STR	1.9	Fusion 6C
25	2	rs798443	SNP	16.56	Ancestry
26	2	rs1109037	SNP	22.97	Identity
27	2	rs7421394	SNP	32.13	Ancestry

28	2	rs1876482	SNP	37.77	Ancestry
29	2	D2S1360	STR	37.9	HDplex
30	2	rs1834619	SNP	38.25	Ancestry
31	2	rs4666200	SNP	50.29	Ancestry
32	2	rs4670767	SNP	61.33	Ancestry
33	2	D2S441	STR	89.6	Fusion 6C
34	2	rs13400937	SNP	104.48	Ancestry
35	2	rs3827760	SNP	121.02	Ancestry
36	2	rs260690	SNP	121.07	Ancestry
37	2	rs993934	SNP	134.06	Identity
38	2	rs6754311	SNP	148.23	Ancestry
39	2	rs10496971	SNP	155.29	Ancestry
40	2	rs10497191	SNP	165.89	Ancestry
41	2	rs2627037	SNP	186.22	Ancestry
42	2	rs12997453	SNP	187.96	Identity
43	2	rs1569175	SNP	200.00	Ancestry
44	2	D2S1338	STR	220.4	Fusion 6C
45	2	rs907100	SNP	255.74	Identity
46	3	rs1357617	SNP	1.55	Identity
47	3	rs4955316	SNP	53.87	Ancestry
48	3	rs4364205	SNP	57.22	Identity
49	3	rs9809104	SNP	62.56	Ancestry
50	3	D3S1358	STR	67.9	Fusion 6C
51	3	rs6548616	SNP	106.83	Ancestry
52	3	rs1872575	SNP	121.47	Identity
53	3	rs12629908	SNP	128.74	Ancestry
54	3	rs12498138	SNP	129.08	Ancestry
55	3	rs9845457	SNP	144.95	Ancestry
56	3	D3S1744	STR	155.5	HDplex
57	3	rs734873	SNP	156.36	Ancestry
58	3	rs2030763	SNP	186.80	Ancestry
59	3	rs1513181	SNP	203.55	Ancestry
60	3	rs1355366	SNP	208.79	Identity
61	3	rs6444724	SNP	214.13	Identity
62	4	rs9291090	SNP	8.19	Ancestry
63	4	D4S2366	STR	11.9	HDplex
64	4	rs2046361	SNP	23.27	Identity
65	4	rs4833103	SNP	57.75	Ancestry

66	4	rs10007810	SNP	62.74	Ancestry
67	4	rs1369093	SNP	85.61	Ancestry
68	4	rs385194	SNP	98.10	Ancestry
69	4	rs1229984	SNP	109.98	Ancestry
70	4	rs3811801	SNP	109.98	Ancestry
71	4	rs7657799	SNP	113.90	Ancestry
72	4	FGA	STR	157.3	Fusion 6C
73	4	rs6811238	SNP	171.90	Identity
74	4	rs2702414	SNP	183.11	Ancestry
75	4	rs1979255	SNP	213.38	Identity
76	5	rs316598	SNP	6.87	Ancestry
77	5	rs717302	SNP	8.69	Identity
78	5	rs870347	SNP	20.31	Ancestry
79	5	rs159606	SNP	39.17	Identity
80	5	rs16891982	SNP	57.06	Ancestry
81	5	rs37369	SNP	58.69	Ancestry
82	5	rs6451722	SNP	66.83	Ancestry
83	5	D5S2500	STR	74.9	HDplex
84	5	rs12657828	SNP	94.54	Ancestry
85	5	D5S818	STR	131.8	Fusion 6C
86	5	CSF1PO	STR	157.7	Fusion 6C
87	5	rs6556352	SNP	164.89	Ancestry
88	5	rs7704770	SNP	168.40	Identity
89	5	rs1500127	SNP	175.79	Ancestry
90	5	rs7722456	SNP	185.63	Ancestry
91	5	rs251934	SNP	199.13	Identity
92	5	rs6422347	SNP	204.66	Ancestry
93	5	rs338882	SNP	206.18	Identity
94	6	rs1040045	SNP	13.85	Ancestry
95	6	rs13218440	SNP	30.00	Identity
96	6	rs2504853	SNP	31.07	Ancestry
97	6	rs7745461	SNP	45.84	Ancestry
98	6	SE33	STR	98.3	Fusion 6C, HDplex
99	6	rs192655	SNP	100.59	Ancestry
100	6	D6S474	STR	121.6	HDplex
101	6	rs3823159	SNP	141.67	Ancestry
102	6	rs4463276	SNP	152.09	Ancestry
103	6	rs214955	SNP	163.05	Identity

104	6	rs4458655	SNP	178.20	Ancestry
105	6	rs727811	SNP	180.85	Identity
106	6	rs1871428	SNP	190.99	Ancestry
107	7	rs6955448	SNP	6.10	Identity
108	7	rs917118	SNP	6.42	Identity
109	7	rs731257	SNP	21.91	Ancestry
110	7	rs917115	SNP	42.79	Ancestry
111	7	rs32314	SNP	50.43	Ancestry
112	7	rs2330442	SNP	62.71	Ancestry
113	7	rs4717865	SNP	85.07	Ancestry
114	7	rs10954737	SNP	97.13	Ancestry
115	7	D7S820	STR	97.4	Fusion 6C
116	7	rs705308	SNP	108.36	Ancestry
117	7	D7S1517	STR	128.4	HDplex
118	7	rs7803075	SNP	134.71	Ancestry
119	7	rs321198	SNP	143.04	Identity
120	7	rs10236187	SNP	148.41	Ancestry
121	7	rs6464211	SNP	167.32	Ancestry
122	7	rs737681	SNP	181.56	Identity
123	8	rs10108270	SNP	8.83	Ancestry
124	8	rs3943253	SNP	26.74	Ancestry
125	8	rs10092491	SNP	51.20	Identity
126	8	rs1471939	SNP	52.41	Ancestry
127	8	rs1462906	SNP	55.92	Ancestry
128	8	rs12544346	SNP	97.11	Ancestry
129	8	D8S1132	STR	116.3	HDplex
130	8	rs6990312	SNP	118.55	Ancestry
131	8	rs2196051	SNP	125.36	Ancestry
132	8	rs7844723	SNP	126.48	Ancestry
133	8	D8S1179	STR	133.1	Fusion 6C
134	8	rs4288409	SNP	152.15	Identity
135	8	rs2056277	SNP	157.20	Identity
136	8	rs2001907	SNP	159.89	Ancestry
137	8	rs1871534	SNP	168.81	Ancestry
138	9	rs1015250	SNP	4.68	Identity
139	9	rs7041158	SNP	54.60	Identity
140	9	rs10511828	SNP	55.43	Ancestry
141	9	rs3793451	SNP	67.42	Ancestry

142	9	rs2306040	SNP	97.32	Ancestry
143	9	rs10513300	SNP	127.15	Ancestry
144	9	rs1463729	SNP	135.60	Identity
145	9	rs3814134	SNP	135.91	Ancestry
146	9	rs1360288	SNP	137.27	Identity
147	9	rs2073821	SNP	149.65	Ancestry
148	9	rs10776839	SNP	155.59	Identity
149	10	rs826472	SNP	4.43	Identity
150	10	rs735155	SNP	8.23	Identity
151	10	D10S2325	STR	30	HDplex
152	10	rs3780962	SNP	41.00	Identity
153	10	rs3793791	SNP	67.58	Ancestry
154	10	rs4746136	SNP	92.70	Ancestry
155	10	rs4918664	SNP	112.08	Ancestry
156	10	rs4918842	SNP	131.76	Ancestry
157	10	rs740598	SNP	136.31	Identity
158	10	D10S1248	STR	164.1	Fusion 6C
159	10	rs964681	SNP	169.33	Identity
160	10	rs4880436	SNP	173.42	Ancestry
161	11	TH01	STR	6.2	Fusion 6C
162	11	rs1498553	SNP	14.07	Identity
163	11	rs10839880	SNP	18.07	Ancestry
164	11	rs901398	SNP	21.99	Identity
165	11	rs1837606	SNP	28.96	Ancestry
166	11	rs2946788	SNP	44.86	Ancestry
167	11	rs174570	SNP	72.34	Ancestry
168	11	rs11227699	SNP	76.89	Ancestry
169	11	rs1079597	SNP	118.27	Ancestry
170	11	rs10488710	SNP	121.59	Identity
171	11	rs948028	SNP	129.89	Ancestry
172	11	rs2076848	SNP	160.91	Identity
173	12	vWA	STR	16.6	Fusion 6C
174	12	rs2269355	SNP	18.61	Identity
175	12	rs2416791	SNP	26.93	Ancestry
176	12	D12S391	STR	29.5	Fusion6C, HDplex
177	12	rs1513056	SNP	38.38	Ancestry
178	12	rs214678	SNP	64.27	Ancestry
179	12	rs772262	SNP	74.37	Ancestry

180	12	rs2111980	SNP	122.11	Identity
181	12	rs2070586	SNP	126.40	Ancestry
182	12	rs2238151	SNP	128.46	Ancestry
183	12	rs671	SNP	128.47	Ancestry
184	12	rs1503767	SNP	140.97	Ancestry
185	12	rs10773760	SNP	169.47	Identity
186	13	rs1335873	SNP	1.08	Identity
187	13	rs1886510	SNP	4.29	Identity
188	13	rs9319336	SNP	17.46	Ancestry
189	13	rs7997709	SNP	32.49	Ancestry
190	13	rs1572018	SNP	43.07	Ancestry
191	13	rs2166624	SNP	43.68	Ancestry
192	13	rs7326934	SNP	52.54	Ancestry
193	13	rs9530435	SNP	71.56	Ancestry
194	13	D13S317	STR	76.8	Fusion 6C
195	13	rs1058083	SNP	95.23	Identity
196	13	rs354439	SNP	109.35	Identity
197	13	rs9522149	SNP	124.06	Ancestry
198	14	rs1760921	SNP	1.19	Ancestry
199	14	rs1454361	SNP	16.99	Identity
200	14	rs2357442	SNP	48.11	Ancestry
201	14	rs722290	SNP	48.66	Identity
202	14	rs1950993	SNP	56.07	Ancestry
203	14	rs8021730	SNP	62.84	Ancestry
204	14	rs946918	SNP	81.00	Ancestry
205	14	rs873196	SNP	108.09	Identity
206	14	rs200354	SNP	109.62	Ancestry
207	14	rs4530059	SNP	121.09	Identity
208	14	rs3784230	SNP	122.63	Ancestry
209	15	rs2016276	SNP	4.84	Identity
210	15	rs1800414	SNP	15.45	Ancestry
211	15	rs12913832	SNP	16.02	Ancestry
212	15	rs12439433	SNP	32.93	Ancestry
213	15	rs1821380	SNP	39.01	Identity
214	15	rs735480	SNP	43.13	Ancestry
215	15	rs1426654	SNP	47.11	Ancestry
216	15	rs1528460	SNP	51.74	Identity
217	15	rs2899826	SNP	77.96	Ancestry

218	15	rs8035124	SNP	98.64	Ancestry
219	15	Penta E	STR	109.9	Fusion 6C
220	16	rs4984913	SNP	2.45	Ancestry
221	16	rs729172	SNP	12.22	Identity
222	16	rs2342747	SNP	13.43	Identity
223	16	rs4781011	SNP	27.75	Ancestry
224	16	rs818386	SNP	85.77	Ancestry
225	16	rs430046	SNP	96.27	Identity
226	16	rs1382387	SNP	104.97	Identity
227	16	rs2966849	SNP	122.08	Ancestry
228	16	D16S539	STR	126.5	Fusion 6C
229	16	rs459920	SNP	133.34	Ancestry
230	17	rs1879488	SNP	5.75	Ancestry
231	17	rs9905977	SNP	9.62	Identity
232	17	rs740910	SNP	16.03	Identity
233	17	rs4411548	SNP	70.29	Ancestry
234	17	rs2593595	SNP	70.51	Ancestry
235	17	rs17642714	SNP	77.00	Ancestry
236	17	rs4471745	SNP	82.22	Ancestry
237	17	rs2033111	SNP	82.56	Ancestry
238	17	rs11652805	SNP	94.41	Ancestry
239	17	rs10512572	SNP	104.32	Ancestry
240	17	rs2125345	SNP	114.86	Ancestry
241	17	rs938283	SNP	127.80	Identity
242	17	rs2292972	SNP	136.84	Identity
243	18	rs1493232	SNP	3.33	Identity
244	18	rs4798812	SNP	34.16	Ancestry
245	18	rs9951171	SNP	35.32	Identity
246	18	rs2042762	SNP	63.21	Ancestry
247	18	rs7226659	SNP	67.29	Ancestry
248	18	rs7238445	SNP	77.96	Ancestry
249	18	rs1736442	SNP	82.50	Identity
250	18	D18S51	STR	89.9	Fusion 6C, HDplex
251	18	rs881728	SNP	90.21	Ancestry
252	18	rs3916235	SNP	103.90	Ancestry
253	18	rs4891825	SNP	104.30	Ancestry
254	18	rs874299	SNP	123.20	Ancestry
255	18	rs1024116	SNP	124.35	Identity

256	19	rs7251928	SNP	13.22	Ancestry
257	19	rs719366	SNP	48.25	Identity
258	19	D19S433	STR	49.9	Fusion 6C
259	19	rs8113143	SNP	55.53	Ancestry
260	19	rs576261	SNP	63.53	Identity
261	19	rs3745099	SNP	90.88	Ancestry
262	19	rs2532060	SNP	102.81	Ancestry
263	20	rs1031825	SNP	12.43	Identity
264	20	rs6104567	SNP	30.36	Ancestry
265	20	rs445251	SNP	38.01	Identity
266	20	rs1005533	SNP	60.44	Identity
267	20	rs1523537	SNP	79.99	Identity
268	20	rs3907047	SNP	86.38	Ancestry
269	20	rs310644	SNP	113.91	Ancestry
270	21	rs722098	SNP	6.36	Identity
271	21	D21S11	STR	13.1	Fusion 6C
272	21	rs2830795	SNP	29.40	Identity
273	21	rs2831700	SNP	31.43	Identity
274	21	rs2835370	SNP	44.59	Ancestry
275	21	D21S2055	STR	50.7	HDplex
276	21	rs914165	SNP	54.87	Identity
277	21	rs221956	SNP	60.66	Identity
278	21	Penta D	STR	61.5	Fusion 6C
279	22	rs1296819	SNP	6.26	Ancestry
280	22	rs733164	SNP	33.29	Identity
281	22	rs4821004	SNP	38.44	Ancestry
282	22	rs987640	SNP	40.42	Identity
283	22	D22S1045	STR	49.4	Fusion 6C
284	22	rs2024566	SNP	53.47	Ancestry
285	22	rs2040411	SNP	65.39	Identity
286	22	rs5768007	SNP	66.89	Ancestry
287	22	rs1028528	SNP	67.71	Identity

Appendix B – Reagents and components

Applied reagents and components are listed with producer and LOT.

Component	Producer	LOT
PowerPlex® Fusion 6C 5X Master Mix	Promega	0000228304
PowerPlex® Fusion 6C 5X Primer Pair Mix	Promega	0000224598
5X AmpSolutionTM Reagent	Promega	0000182874
WEN Internal Lane Standard 500	Promega	0000283901
PowerPlex® Fusion 6C Allelic Ladder Mix	Promega	0000280539
2800M Control DNA	Promega	0000328717
Reaction Mix A	Qiagen	CH1500564b
Primer Mix	Qiagen	CH1600486
MultiTaq2 DNA Polymerase	Qiagen	CH1600021b
DNA Size Standard 550 (BTO)	Qiagen	160031010
Allelic Ladder HDplex	Qiagen	160031010
POP-4®	Thermo Fisher Scientific	1612119
Anode Buffer Container	Thermo Fisher Scientific	1701368
Cathode Buffer Container	Thermo Fisher Scientific	1703417
Buffer G2	Qiagen	154012845
Proteinase K	Qiagen	154012845
EZ1® DNA Investigator Kit	Qiagen	157011997
Ion AmpliSeq™ Chef Supplies DL8	Thermo Fisher Scientific	1774496
Ion AmpliSeq™ Chef Reagents DL8	Thermo Fisher Scientific	1839273
Ion AmpliSeq™ Chef Solutions DL8	Thermo Fisher Scientific	1824916
IonCode™ 0101-0132 in 96 Well PCR Plates	Thermo Fisher Scientific	1703014
Precision ID Identity Panel	Thermo Fisher Scientific	1612012
Precision ID Ancestry Panel	Thermo Fisher Scientific	1612012
Ion 314™ Chip V2 BC	Thermo Fisher Scientific	P31702.1
Ion PGM™ Hi-Q™ View Reagents	Thermo Fisher Scientific	1777767
Ion PGM™ Hi-Q Chef Solutions	Thermo Fisher Scientific	1774257A
Ion PGM™ Hi-Q™ Chef Supplies	Thermo Fisher Scientific	1771074
Ion PGM™ Seq Supplies	Thermo Fisher Scientific	MHWW500
Ion PGM™ Hi-Q™ Sequencing Reagents	Thermo Fisher Scientific	012572
Ion PGM™ Hi-Q™ Sequencing Solutions	Thermo Fisher Scientific	1853251
Ion PGM™ Hi-Q™ Seq dNTP	Thermo Fisher Scientific	00469888

Appendix C – GenoGeographer results

Ancestry inference results was provided for all project samples using GenoGeographer (Tvedebrink et al. 2017). Statistical calculations have been performed for the metapopulations given in the first column. The allele frequency databases for the different metapopulations are based on "n" individuals. Further, the probability for the genotype in the different metapopulations are given by the «log10 P(G|pop)», "var[log10 P(G|pop)]» refers to the variance in terms of the allele frequencies, and a confidence interval are reported based on this. Finally, a statistical test, explained in the article by Tvedebrink et al. (2017), is performed. A z- and p-value is given for this test and is used to evaluate whether it is reasonable that the genotype originates from the metapopulations included here, or not.

Sample 34

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-35,356	0,024	-35,053	-35,659	0,096	0,462	true
Middle East	382	-39,583	0,077	-39,039	-40,127	1,009	0,156	true
North Africa	235	-44,1	0,18	-43,268	-44,932	2,142	0,016	false
South / Central Asia	489	-48,491	0,076	-47,95	-49,032	3,538	0	false
Greenland	75	-60,372	0,686	-58,749	-61,995	6,99	0	false
Somalia	75	-66,037	2,159	-63,157	-68,916	10,552	0	false
East Asia	622	-85,52	0,717	-83,86	-87,179	16,331	0	false
Sub Sahara	668	-141,342	1,493	-138,948	-143,737	45,305	0	false

Sample 35

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-33,593	0,02	-33,312	-33,873	-0,22	0,587	true
Middle East	382	-38,64	0,075	-38,101	-39,178	0,995	0,16	true
South / Central Asia	489	-42,252	0,058	-41,78	-42,724	1,646	0,05	false
North Africa	235	-44,261	0,171	-43,45	-45,071	2,509	0,006	false
Greenland	75	-52,559	0,458	-51,233	-53,886	4,585	0	false
Somalia	75	-64,891	1,764	-62,288	-67,495	10,361	0	false
East Asia	622	-76,295	0,59	-74,789	-77,801	13,432	0	false
Sub Sahara	668	-146,28	1,515	-143,867	-148,692	47,634	0	false

Sample 36

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-33,947	0,019	-33,674	-34,22	-0,274	0,608	true
Middle East	382	-37,628	0,066	-37,125	-38,131	0,333	0,37	true
South / Central Asia	489	-45,332	0,058	-44,859	-45,804	2,275	0,011	false
North Africa	235	-45,667	0,131	-44,956	-46,377	2,559	0,005	false
Greenland	75	-61,768	0,59	-60,263	-63,274	7,126	0	false
Somalia	75	-69,48	1,572	-67,023	-71,937	11,072	0	false
East Asia	622	-89,885	0,634	-88,325	-91,446	16,935	0	false
Sub Sahara	668	-164,86	1,458	-162,493	-167,226	52,483	0	false

Sample 37

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-34,66	0,022	-34,372	-34,949	-0,87	0,808	true
Middle East	382	-39,144	0,074	-38,612	-39,676	0,028	0,489	true
North Africa	235	-43,403	0,139	-42,673	-44,133	1,002	0,158	true
South / Central Asia	489	-47,727	0,073	-47,196	-48,257	2,275	0,011	false
Greenland	75	-68,82	0,792	-67,077	-70,564	8,689	0	false
Somalia	75	-64,216	2,075	-61,393	-67,039	8,737	0	false
East Asia	622	-92,478	0,817	-90,707	-94,249	17,046	0	false
Sub Sahara	668	-147,566	1,869	-144,887	-150,245	45,352	0	false

Sample 38

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-31,965	0,018	-31,702	-32,227	-1,321	0,907	true
Middle East	382	-36,162	0,064	-35,666	-36,659	-0,353	0,638	true
North Africa	235	-38,091	0,092	-37,494	-38,687	-0,087	0,535	true
South / Central Asia	489	-40,81	0,054	-40,356	-41,264	0,717	0,237	true
Greenland	75	-47,624	0,364	-46,442	-48,806	2,441	0,007	false
Somalia	75	-58,612	1,939	-55,883	-61,34	8	0	false
East Asia	622	-68,569	0,691	-66,94	-70,198	10,356	0	false
Sub Sahara	668	-125,986	1,19	-123,848	-128,125	39,383	0	false

Sample 39

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-35,967	0,024	-35,66	-36,273	0,935	0,175	true
Middle East	382	-38,724	0,07	-38,206	-39,243	1,332	0,091	true
North Africa	235	-41,792	0,132	-41,079	-42,504	1,968	0,025	false
South / Central Asia	489	-44,08	0,068	-43,568	-44,593	2,639	0,004	false
Greenland	75	-56,136	0,527	-54,713	-57,559	6,1	0	false
Somalia	75	-60,311	1,74	-57,726	-62,897	9,367	0	false
East Asia	622	-73,145	0,51	-71,745	-74,545	12,624	0	false
Sub Sahara	668	-123,36	1,364	-121,071	-125,65	39,317	0	false

Sample 40

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-24,492	0,014	-24,263	-24,72	-1,063	0,856	true
Middle East	382	-28,316	0,052	-27,871	-28,761	-0,12	0,548	true
North Africa	235	-30,477	0,081	-29,919	-31,035	0,346	0,365	true
South / Central Asia	489	-33,069	0,045	-32,654	-33,484	1,071	0,142	true
Greenland	75	-46,224	0,423	-44,95	-47,498	5,34	0	false
Somalia	75	-48,449	1,767	-45,844	-51,055	7,98	0	false
East Asia	622	-61,539	0,608	-60,01	-63,067	11,788	0	false
Sub Sahara	668	-106,821	1,335	-104,556	-109,085	37,767	0	false

Sample 41

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-28,027	0,017	-27,773	-28,28	-0,482	0,685	true
Middle East	382	-30,938	0,057	-30,468	-31,407	0,085	0,466	true
South / Central Asia	489	-32,89	0,043	-32,481	-33,298	0,182	0,428	true
North Africa	235	-33,642	0,098	-33,03	-34,254	0,759	0,224	true
Greenland	75	-43,558	0,349	-42,4	-44,715	3,555	0	false
Somalia	75	-50,521	1,928	-47,799	-53,243	8,027	0	false
East Asia	622	-58,099	0,592	-56,591	-59,606	9,675	0	false
Sub Sahara	668	-108,817	1,182	-106,687	-110,948	37,588	0	false

Sample 42

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-37,13	0,025	-36,82	-37,44	0,867	0,193	true
Middle East	382	-38,795	0,072	-38,271	-39,32	0,894	0,186	true
North Africa	235	-41,979	0,125	-41,287	-42,671	1,661	0,048	false
South / Central Asia	489	-46,179	0,076	-45,638	-46,719	2,935	0,002	false
Greenland	75	-60,714	0,582	-59,219	-62,21	7,267	0	false
Somalia	75	-61,807	1,827	-59,158	-64,456	9,476	0	false
East Asia	622	-76,524	0,49	-75,152	-77,897	13,414	0	false
Sub Sahara	668	-124,494	1,289	-122,269	-126,719	39,001	0	false

Sample 43

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-31,592	0,019	-31,322	-31,861	-0,742	0,771	true
Middle East	382	-33,955	0,056	-33,49	-34,421	-0,428	0,666	true
North Africa	235	-40,447	0,119	-39,77	-41,124	1,545	0,061	true
South / Central Asia	489	-43,942	0,071	-43,421	-44,462	2,566	0,005	false
Greenland	75	-55,651	0,503	-54,262	-57,041	5,936	0	false
Somalia	75	-64,282	1,874	-61,599	-66,965	10,774	0	false
East Asia	622	-74,014	0,482	-72,654	-75,374	12,956	0	false
Sub Sahara	668	-139,833	1,672	-137,298	-142,367	45,866	0	false

Sample 44

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-38,764	0,026	-38,446	-39,083	0,03	0,488	true
Middle East	382	-43,722	0,086	-43,146	-44,298	1,055	0,146	true
South / Central Asia	489	-48,096	0,07	-47,578	-48,614	1,916	0,028	false
North Africa	235	-49,614	0,182	-48,778	-50,451	2,575	0,005	false
Greenland	75	-62,673	0,626	-61,122	-64,224	6,094	0	false
Somalia	75	-74,244	2,325	-71,256	-77,233	11,587	0	false
East Asia	622	-85,994	0,779	-84,264	-87,724	14,362	0	false
Sub Sahara	668	-163,499	1,649	-160,982	-166,015	50,587	0	false

Sample 45

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-34,316	0,018	-34,052	-34,58	-1,591	0,944	true
Middle East	382	-37,41	0,054	-36,953	-37,867	-1,161	0,877	true
North Africa	235	-43,139	0,106	-42,502	-43,776	0,35	0,363	true
South / Central Asia	489	-44,691	0,057	-44,223	-45,159	0,677	0,249	true
Greenland	75	-56,844	0,467	-55,505	-58,184	4,053	0	false
Somalia	75	-64,888	1,487	-62,499	-67,278	8,147	0	false
East Asia	622	-80,462	0,489	-79,091	-81,833	12,288	0	false
Sub Sahara	668	-150,921	1,462	-148,551	-153,291	45,48	0	false

Sample 46

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-31,363	0,016	-31,113	-31,613	-1,319	0,906	true
Middle East	382	-33,68	0,047	-33,256	-34,103	-1,148	0,874	true
North Africa	235	-37,819	0,087	-37,24	-38,399	-0,05	0,52	true
South / Central Asia	489	-40,518	0,046	-40,098	-40,937	0,554	0,29	true
Somalia	75	-59,152	0,989	-57,203	-61,101	7,582	0	false
Greenland	75	-64,634	0,673	-63,027	-66,241	7,861	0	false
East Asia	622	-79,207	0,384	-77,993	-80,422	13,244	0	false
Sub Sahara	668	-134,857	1,24	-132,675	-137,039	41,25	0	false

Sample 47

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Middle East	382	-37,645	0,054	-37,191	-38,099	-1,443	0,925	true
Europe	1014	-38,976	0,024	-38,671	-39,28	-0,328	0,628	true
North Africa	235	-43,542	0,111	-42,889	-44,194	0,257	0,399	true
South / Central Asia	489	-46,243	0,083	-45,679	-46,808	0,874	0,191	true
Greenland	75	-60,495	0,621	-58,95	-62,039	4,947	0	false
Somalia	75	-64,441	1,076	-62,409	-66,474	7,75	0	false
East Asia	622	-78,853	0,323	-77,739	-79,967	11,5	0	false
Sub Sahara	668	-138,594	1,304	-136,356	-140,831	40,862	0	false

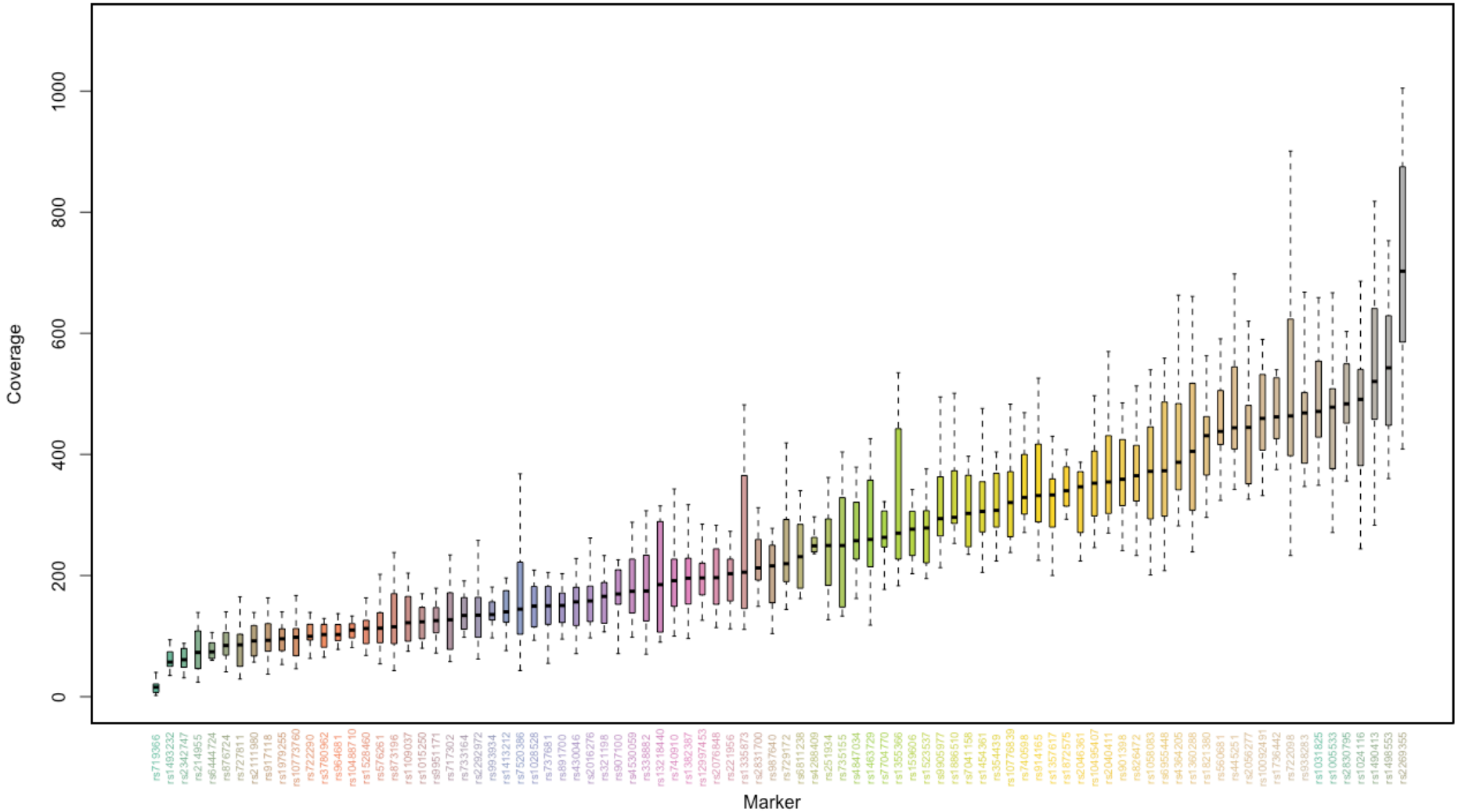
Sample 48

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-34,517	0,023	-34,221	-34,813	-1,115	0,868	true
Middle East	382	-39,766	0,073	-39,238	-40,295	0,287	0,387	true
South / Central Asia	489	-43,075	0,061	-42,589	-43,561	1,02	0,154	true
North Africa	235	-43,678	0,138	-42,95	-44,407	1,3	0,097	true
Greenland	75	-53,576	0,489	-52,205	-54,947	4,171	0	false
Somalia	75	-64,491	2,034	-61,695	-67,286	9,653	0	false
East Asia	622	-74,208	0,709	-72,557	-75,858	11,842	0	false
Sub Sahara	668	-142,758	1,837	-140,101	-145,414	45,583	0	false

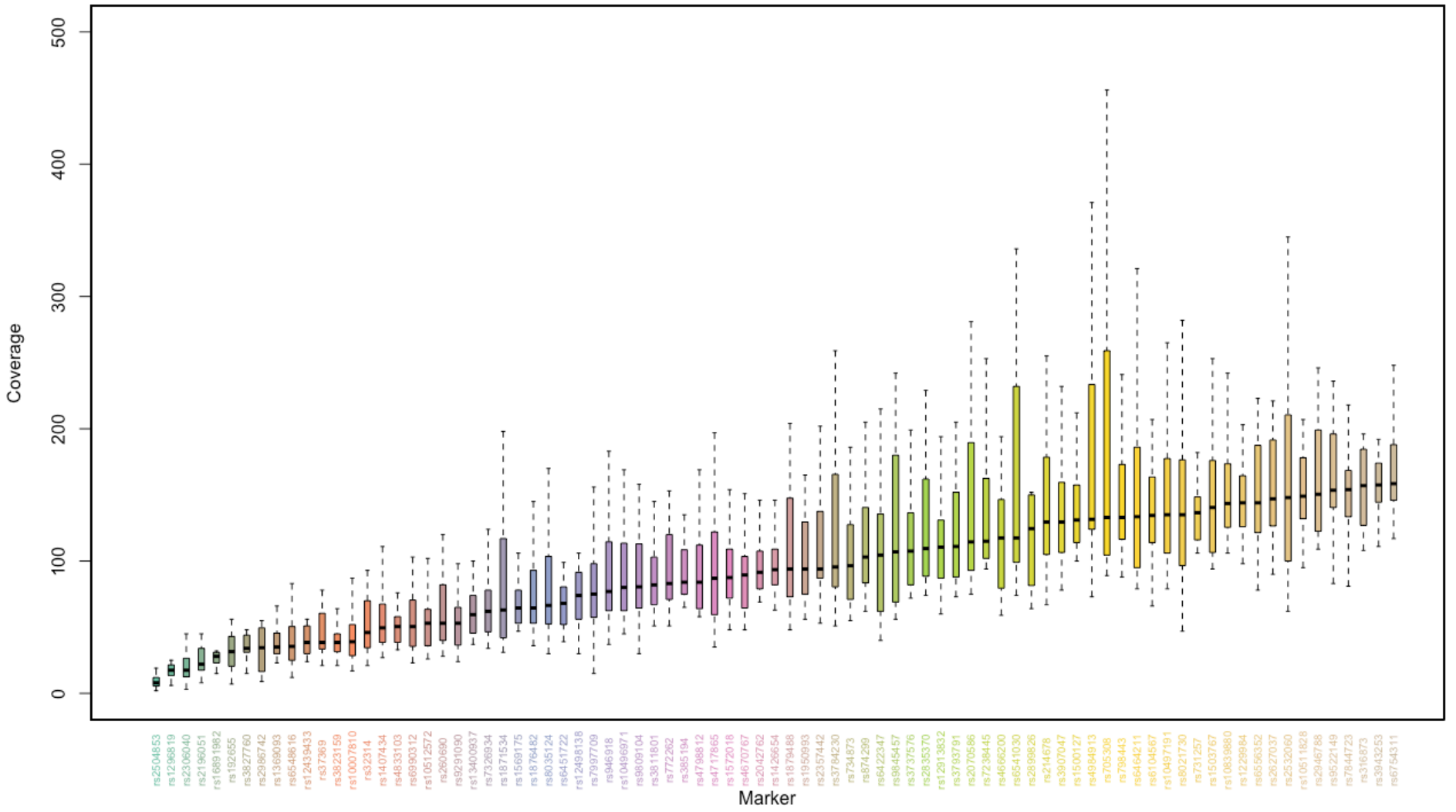
Sample 49

metapopulation	n	log10 P(G|pop)	var[log10 P(G|pop)]	CI[log10 P(G|pop)] upr	CI[log10 P(G|pop)] lwr	z-score	p-value	accept
Europe	1014	-33,134	0,02	-32,854	-33,414	0,738	0,23	true
Middle East	382	-37,296	0,069	-36,782	-37,81	1,705	0,044	false
South / Central Asia	489	-42,116	0,063	-41,625	-42,607	2,963	0,002	false
North Africa	235	-43,472	0,149	-42,716	-44,228	3,491	0	false
Greenland	75	-52,572	0,517	-51,162	-53,981	6,036	0	false
Somalia	75	-65,656	2,192	-62,754	-68,558	12,392	0	false
East Asia	622	-69,705	0,605	-68,181	-71,23	12,439	0	false
Sub Sahara	668	-143,197	2,052	-140,389	-146,005	48,441	0	false

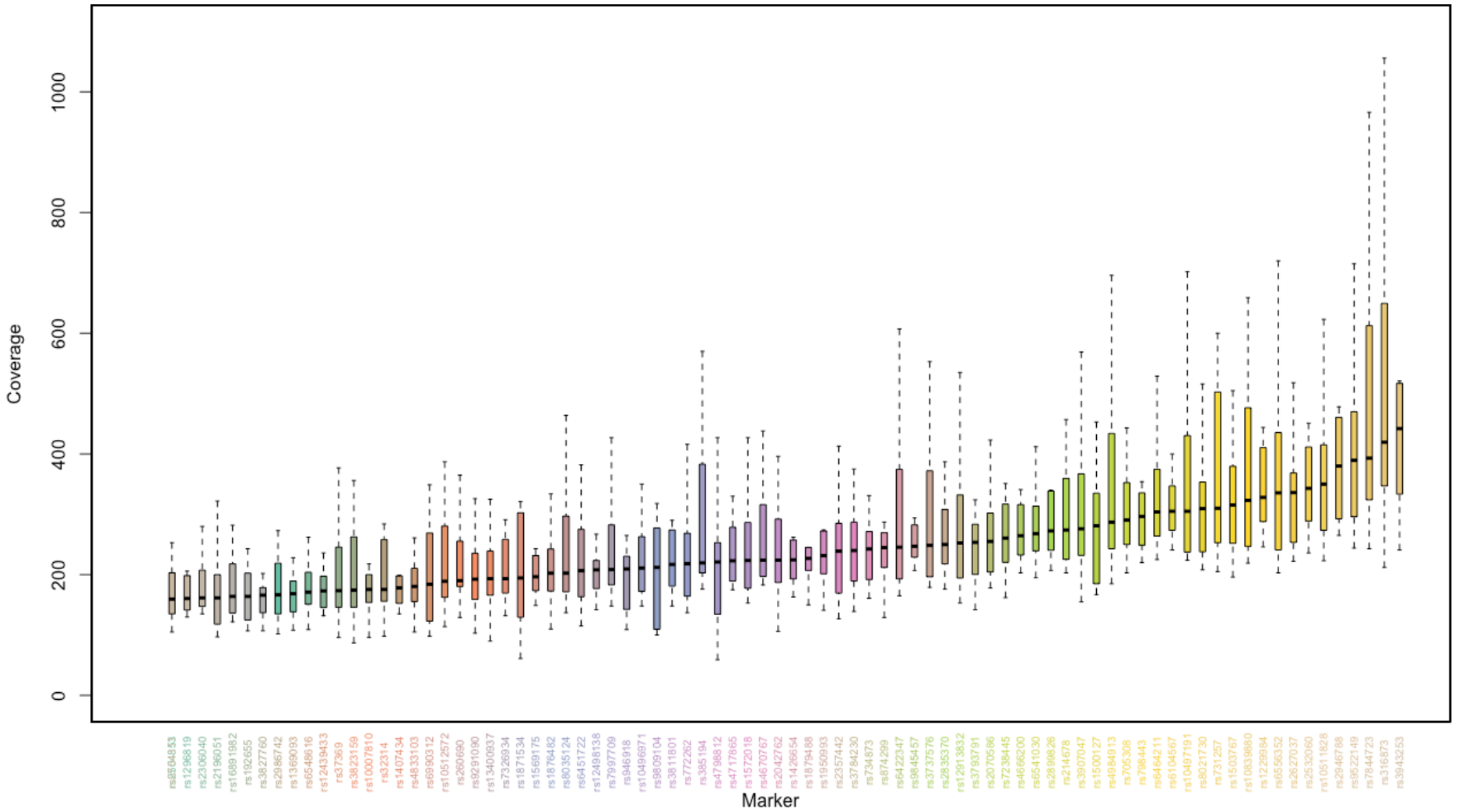
Appendix D – Non-normalised coverage values for the SNP markers



Boxplot illustrating the coverage (y-axis) for the autosomal Identity SNPs (x-axis). The box for each marker is based on 16 coverage values, one from each project sample.



Boxplot illustrating the coverage (y-axis) for 83 Ancestry SNPs (x-axis). The remaining 82 SNPs can be found in the figure below. The box for each marker is based on 16 coverage values, one from each project sample.



Boxplot illustrating the coverage (y-axis) for 82 Ancestry SNPs (x-axis). The remaining 83 SNPs can be found in the figure above. The box for each marker is based on 16 coverage values, one from each project sample.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway