APPLICABILITY OF HIGH-DENSITY GENOME-WIDE SNP ARRAYS IN FORENSICS

HELGENOM SNP-ARRAYER I RETTSGENETISKE ANVENDELSER

ANE ELIDA FONNELØP

NORWEGIAN UNIVERSITY OF LIFE SCIENCES
DEPARTMENT OF ANIMAL AND AQUACULTURAL SCIENCES
MASTER THESIS 60 CREDITS 2010

PREFACE

This report is the result of a final thesis project for a master's degree in Biotechnology at the Norwegian University of Life Sciences. The assignment is performed in cooperation between the Center of Integrative Genetics and The Forensic Institute of medicine in Oslo.

I would like to thank my supervisors professor Thore Egeland and professor Sigbjørn Lien for making this project possible and for great support and guidance. A special thanks to Thore for developing a statistical method for this assignment and also for encouraging me to attend a course in R programming. This has been very helpful and R has been used for several of the analysis in this thesis.

I would also like to thank SNP Platform Manager Paul Berg and Scientist Matthew Kent for guidance throughout the project and for teaching me about SNP analysis.

Thanks to lab technician Arne Roseth and Research Assistant Kristil Sundsaasen for help and guidance at the laboratory.

Finally I would like to thank the forensic institute of medicine in Oslo, Senior engineer Bente Mevåg for being supportive about this assignment, Heidi Haltbak and Beate Hellerud for helping me with the STR mixture interpretation, Karianne Molema for preparing the blinded mixtures and the rest of the laboratory group, you have all been very helpful and supportive.

Oslo, 12. mai 2010

Ane Elida Fonneløp

ABSTRACT

This study tests the performance of SNP arrays in forensic analyses. The first part of the thesis is based on finding the utility of SNP arrays when it comes to samples of limited concentration and degraded quality. The second part is conducted to analyze mixtures by raw fluorescence data collected from of SNP arrays. All samples were analyzed by the Illumina GoldenGate 360 SNP test panel.

Dilutions, degraded samples and amplified samples have been analyzed to test the performance of these arrays on DNA samples with low concentrations and degraded quality. This study demonstrates that microarray analyses and cluster based genotype calling does not seem to be suited for analysis of low DNA template samples. It further demonstrates that whole genome amplification of low template samples seems not to produce full genome coverage.

The mixtures were analyzed by two statistical methods, the first based on a genetic distance measurement developed by Homer et al. (2008) the second based on linear regression developed for this assignment by Thore Egeland[1]. The statistical method described in Homer et al. (2008) seems not to function well when the number of SNPs is limited. There are indications that the resolving of individual contributions are depending on the mixture proportions from the contributors. However the findings from the regression based statistics suggest that it microarray analysis can be a helpful tool for mixture interpreting.

[1] Research scientist (statistician), Forensic Institute of Medicine, Oslo.

SAMMENDRAG

Hensikten med denne oppgaven var å undersøke om det er mulig å anvende helgenom SNP microarryer in en rettsgenetisk sammenheng. Første delen av oppgaven omhandler hvilke begrensninger som finnes ved bruk av disse arrayene når det gjelder DNA konsentrasjon og kvalitet. Den andre delen går ut på å analysere blandingsprøver med rå-fluorescensdata samlet in fra microarrayanalyser.

Fortynninger, degraderte prøver og amplifiserte prøver har blitt analysert for å teste resultatene av SNP microarrayanalyser på prøver med lav DNA konsentrasjon og degraderte prøver. Denne studien viser at det er sannsynlig at klyngemetoden brukt til genotype bestemmelse sannsynligvis ikke er egnet for prøver med lav DNA konsentrasjon. Studien viser også at helgenomamplifiserting av prøver med lav DNA konsentrasjon ikke ser ut til å gi full amplifikasjon av hele genomet.

Blandingsprøvene ble analysert med to statistiske metoder. Den første basert på et genetisk avstandsmål utviklet av Homer et al. (2008), den andre metoden basert på lineær regresjon utviklet for denne oppgaven av Thore Egeland[2]. Den statistiske metoden utviklet av Homer et al. (2008) ser ikke ut til å fungere optimalt når antall SNPer analysert er begrenset, samtidig er det funnet indikasjoner på at muligheten for å identifisere et individ i blandingen er avhengig av andelen fra hver bidragsyter i blandingene. Funnene fra den regresjonsbaserte metoden indikerer at microarray analyser kan være et nyttig verktøy for a tolke blandinger.

[2] Forsker (statistiker), Rettsmedisinsk Institutt, Oslo.

ABBREVATIONS

A	Adenin
ASO	Allele Specific Oligoes
bp	Base pair
C	Cytosine
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
EDTA	Ethylenediaminetetraacetic acid
GCS	Gene Call Score
G	Guanine
HCL	Hydrochloric acid
LSO	Locus Specific Oligo
MALDI-TOF-MS	Matrix-assisted-laser-desorption-ionization time-of flight mass spectrometry
MDA	Multiple Displacement Amplification
ng	Nanogram ($1x10^{-9}$ gram)
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
RNAse	Ribonuclease
SAP	Shrimp Alkaline Phosphate
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
T	Thymine
TBE	Tris Bor EDTA
TE	Tris EDTA
Tris	2-Amino-2-(hydroxymetyl).1,3.propandiol
UV	Ultra Violet
VNTR	Variable Number of Tandem Repeat
WGA	Whole Genome Amplification
µL	Microliter (0.001 milliliter)

Table of contents

1 INTRODUCTION

The first DNA profiling technique was described in 1985 by Dr. Alec Jeffreys. He had discovered certain regions in DNA sequences that were repeated over and over again, and that the amount of repeats present in the DNA varied between individuals. These regions became known as VNTRs (Variant Number of Tandem Repeats). To perform an identity test by these markers an RFLP (Restriction Fragment Length Polymorphism) technique was used. Restriction enzymes were used to cut the areas surrounding the VNTR and fragments were separated by electrophoresis. In 1986 DNA testing was used for the first time in a forensic setting (Butler 2005).

Since 1986 techniques for analyzing DNA have developed rapidly and new and improved tools for forensic analysis are now available. Today the most common way to analyze DNA in a forensic context is by short tandem repeat (STR) markers.

1.1 Classical markers

Repeated DNA sequences are common in eukaryotic genomes and exist in many variants. These repeats are referred to as satellite DNA and are separated by the length of the core sequence and the number of times it is repeated. The number of bases in each repeat can be as high as several thousand basepair (bp). The VNTRs previously mentioned are referred to as mini satellites were the core repeat ranges from 10-100 bp. The STR markers used in the forensic analysis are microsatellites, were the core sequence range from 2-6 bp. The numbers of repeats are highly variable among individuals, and therefore these markers are very applicable for forensic science. STR markers are distributed throughout the genome and occur on average every 10 000 nucleotide. There are expected to be more than a million microsatellite loci in the human genome (Butler 2005).

The forensic STR analyses are based on a polymerase chain reaction (PCR) technique where the flanking regions of the repeated DNA sequence are used to design primers. PCR is a technique that can amplify a single or few copies of a particular fragment in the genome and generate thousands to millions of copies of a particular DNA sequence. This results in the radical decrease in genetic material needed for forensic analysis, from 500ng for the VNTR analysis to 1ng or less to for STRs (Buckleton et al. 2005). Moreover, STRs can be analyzed

simultaneously in multiplexes which give the analysis a high discriminating power (Buckleton 2005).

1.2 The DNA samples in forensic cases

A main issue in the forensic DNA analysis is the samples originating from two or more contributors (mixtures), which create the need for a genotyping technology where individual contributions can be identified with a high probability.

Because each chromosome in the human genome has two copies (one inherited from the mother and one from the father) each locus in a DNA profile can have one or two alleles. If one allele is shown, the individual is homozygous at the locus and if two alleles are present the individual is heterozygous. Examples of a heterozygous and a homozygous locus are given in figure 1, which represents an electropherogram (epg) from a STR DNA analyses. The peaks in the epg represent the alleles present in the sample. The heights of the peaks are measured in relative fluorescence units (rfu) and are proportional to the amount of DNA present in the sample.

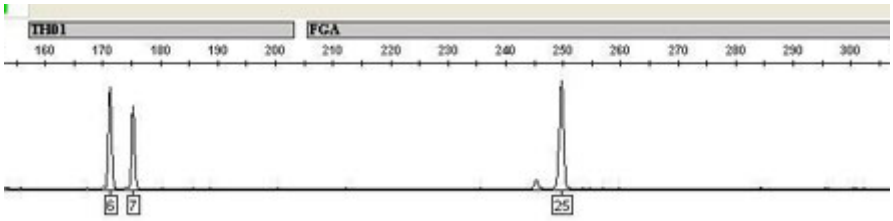


Figure 1. The figure displays an epg from a STR analyses, the peaks in the epg represent the alleles present in the sample. The heights of the peaks are measured in relative fluorescence units (rfu) and are proportional to the amount of DNA present in the sample. Here the individual is heterozygous at the TH01 locus and homozygous at the FGA locus. The small peak at the FGA locus that is not labeled is expected to be a stutter band.

If more than two alleles are present at a locus in an STR profile, it may indicate that the sample is a mixture (Buckleton et al. 2005). In many cases the individuals present in the mixture have contributed different amounts of DNA and a major and minor contributor can be assigned. Figure 2 is an epg from three loci in a two person mixture were the major and minor contributor can be visually separated by the TH01 and FGA loci. For these markers the contributors are heterozygous.

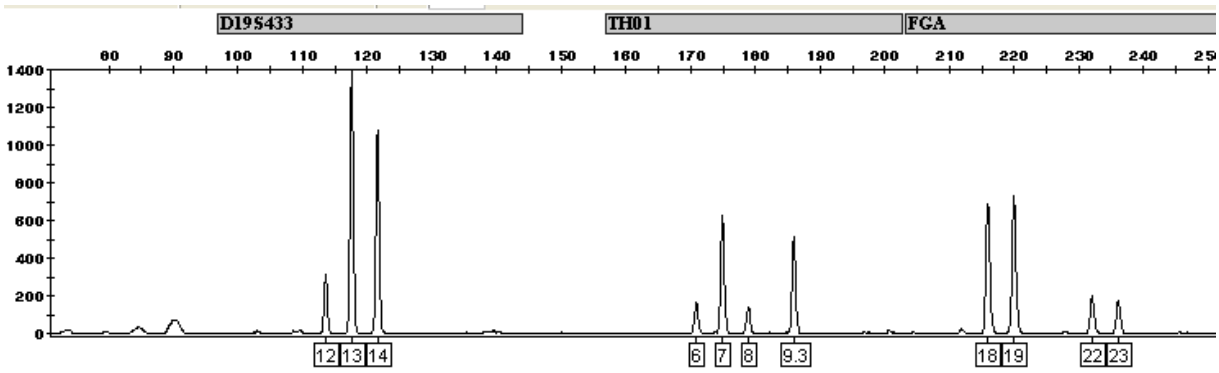


Figure 2. The epg from a two person mixture where the major and minor contributor can be visually separated at two loci. The peaks in the epg represent the alleles present in the sample. The heights of the peaks are measured in relative fluorescence units (rfu) and are proportional to the amount of DNA present in the sample from each contributor. At the TH01 locus the genotype of the major contributor is 7/9.3 and the minor contributor 6/8. For the FGA locus the major is 18/19 and minor 22/23.

A mixture can only be interpreted if the minor contributor is above background noise, this means that the size of the contribution from the minor must be approximately 10% or more of the major. When the contribution from the minor is small, artifacts like stutters and allelic drop outs must also be taken under consideration during analysis of data.

A stutter is a result of slippage during the PCR process and produces peaks at other positions than the parent allele position, as shown in figure 1. A stutter is known to be less than 15% of the parent allele and stutter bands should be excluded from analysis of mixture samples. However it is not always possible to exclude stutters since they may be in the same size as the minor contributor and may be difficult to distinguish (Buckleton et al 2005). Allelic drop out is the phenomena where an allele cannot be seen in the DNA profile. It is typically observed in a heterozygote and will lead to the false impression that the individual is a homozygote. The occurrence of drop out is mostly observed were peak heights are very low like in alleles under 150 rfu (Buckleton et al 2005).

The number of contributors to a mixture in a STR profile is conventionally estimated as the minimum number of contributors by finding the maximum number of alleles present at any locus in the mixture (the maximum allele count method). For instance the presence of four alleles at a locus tells us that at least two people contributed. If five alleles are present the minimum number of contributors will be three. (Buckleton et al. 2005).

STR mixture interpretation is a challenging and time consuming procedure that requires training and experience. The complexity of the interpretation will increase when the number of contributors gets higher. Mixtures consisting of four or more individuals will in many cases

not be analyzed because many individuals by chance could fit into such a complex mixture, especially if there are no clear major and minor components that could be separated by the size of the alleles. The results from mixture interpretation can also vary between labs due to different interpretation methods (Budowle et al. 2009).

In addition, DNA samples analyzed in forensic cases are often not of ideal quality or quantity. In many cases only small amounts of biological material is found at the crime scenes like a sample collected from a touched surface. The DNA samples in forensic cases may also be exposed to environments that are damaging to cells and DNA, and the samples are sometimes left at the crime scene for a longer period before being collected. DNA can be damaged from environmental exposures like humidity, light, elevated temperatures or nucleases that will break the DNA molecules into smaller degraded DNA pieces, often of a size of 300 bp or smaller. Degraded DNA will in most cases give a partial or no DNA profile, because the DNA needs to be intact where the primer binds as well as between the primers for a successful amplification to occur. The longer alleles have a tendency to be missing because they have a higher probability of being damaged (Butler 2005).

For the STRs to be suitable for forensic analysis several features has to be met. First it is important the markers have a high level of variability within a locus so that there is a low probability of two persons from the same population having similar DNA profiles. The length of alleles should range between 100-400bp which will be better suited for degraded DNA. They should be located far apart to avoid linkage and linkage disequilibrium which will lead to a non random association between the alleles. Robustness and reproducibility are also important and loci that do not markedly stutter are desirable to make the interpretation easier.

There is a European agreement to use a set of standardized STR markers for identification (Gill et al 2005). AmpFℓSTR® SGM Plus® PCR Amplification Kit (Applied Biosystems) is currently used for DNA profiling in the forensic laboratory of medicine in Oslo, and in several other European countries (Gill et al 2005). This kit consists of primer sets for amplifying 10 STR loci and a sex marker, and the estimated probability of match between two unrelated people is usually set to less than 10^{-9} (Bucklelton et al. 2005).

The STR markers are proven to be appropriate for identification purposes and also provide reliable genotyping from small DNA quantities. Still there are challenges to the forensic

casework for example when it comes to degraded DNA and mixtures. The implementation of single nucleotide polymorphism (SNP) markers is a widely discussed theme in the forensic society because as explained in the next section the SNPs have several qualities that may lead to improvements in forensic analysis.

1.3 Alternative markers

A SNP is a single base variation occurring at a specific chromosome position in the genome. Typically, the variation is diallelic (i.e. one of only two possible bases A or B will be found at the loci position), but since in diploid organisms the variation will be present on both autosomal chromosome copies an individual can be classified (or genotyped) as AA, AB or BB. To be considered a SNP the variation most occur in at least 1% of the population. SNPs can be accounted for 90% of the variation and will occur every 100-300 bases in the 3 Billion-base pair human genome (http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml). Depending on the assay technology SNPs can be genotyped in target sequences less than 100bp, leading to the higher probability for successful amplification from degraded DNA. In contrast the target sequences needed for STR analysis normally ranging from ~100-400bp. However, because most SNPs are dialleic, the discrimination power for each SNP will be lower than for the multi allelic STR markers. As a consequence analysis of approximately 50 SNPs are required to reach the same level of discrimination between individuals as for 10 STR markers (Gill 2001). Also, because most SNPs only have two alleles at a locus it will be difficult to recognize and interpret a mixture compared to the methods explained previously for the STR markers. This is one of the main arguments why SNP markers are claimed not suitable for forensic analysis (Butler et al. 2005).

Ease of genotyping, their large number and distribution in the genome have made SNPs the preferred marker in most generic analysis. In turn this has stimulated the development of technologies for accurate and high throughput genotyping. The newest development is the SNP Genotyping arrays which can genotype more than 2 million SNPs markers per sample (The HumanOmni2.5-Quad, Illumina[3]).

[3] http://www.ilumina.com

1.4 A new method for forensic mixture interpretation

Homer et al (2008) claimed that by analyzing mixtures by SNP genotyping arrays with up to 500K SNPs all individuals in highly complex mixture can be successfully identified. Moreover they suggest that it should be possible to identify individuals contributing less than 0.1% of the total genomic DNA present in a sample. The method is based on directly utilizing the raw fluorescence measurements available from microarray analysis, and transforming these measurements to allele frequency estimates for each SNP in the mixture. The possibilities described in the paper could, if true, be revolutionizing to mixture interpretation in a forensic context.

1.5 The aim of this study

The aim of this study was to evaluate the performance of SNP microarray genotyping on forensic samples. Two main questions are addressed; (i) what DNA quality is necessary for microarray analysis (DNA quality), -and (ii) is it possible to improve the methods currently used for mixture analysis by SNP microarray analysis (Mixture interpretation). The array platform chosen for the analysis was the Illumina GoldenGate 360 SNP test panel in which 96 samples can be analyzed simultaneously.

1.5.1 DNA quality

A possible drawback with the method described in the paper by Homer et al. (2008) is that although the statistical method seems revolutionizing this technology will not perform well on the real forensic samples. The amounts of DNA required for high-density SNP array analyses (500ng for Affymetrix and 250ng for Illumina) are much higher than the average DNA yield available in forensic cases. The samples collected from crime scenes will many times after extraction turn out to be of lower DNA quantity than 1ng/µL. The DNA quality of a typical forensic sample will thus not be good enough for high-density microarray analyses.

The DNA quality analysis in this thesis sought to test three variables, dilution, degradation and amplification. First, to find the limitations of the genotyping analysis on samples of different concentration, a dilution series was genotyped. The concentration of the samples was classified as good, medium and poor (where good corresponds to the manufactures recommendations and poor are the concentration used for STR analysis). Secondly the SNPs are expected to perform better on degraded samples than the STR markers. To test this statement an artificially degraded sample was prepared and genotyped. The degraded sample

was also diluted to different concentrations in the same range as the dilution series. Finally, because it is expected that genotyping results will be negatively affected by low DNA concentration samples, a whole genome amplification (WGA) technique was performed in an effort to boost assayable DNA. The WGA technique was performed on samples with good, medium and poor concentration (were good corresponds to the concentrations recommended for WGA analysis and poor the lowest concentration that can give an STR profile).

Constrained by available funding it was only possible to perform one Illumina GoldenGate run. Therefore it was important to avoid consuming valuable Illumina GoldenGate reaction on samples that were likely to fail due to extreme variable testing. As a consequence samples were pre-tested to identify the approximate limits to concentration, degradation and amplification using the per run less expensive Sequenom massARRAY platform with two iPLEX SNP genotyping multiplexes consisting of 10 and 29 SNPs.

1.5.2 Mixture interpretation

The currently used mixture interpretation method in forensics has limitations when it comes to complex mixtures and minor contributors that are less than 10% of the major. The mixture interpretation may be eased by a procedure that can overcome these limitations.

The suggested statistical method (Homer et al. 2008) was tested, first by a simulation routine and, secondly on real mixtures. The purpose was to test the limitations of the method when it comes to the number of contributors in a mixture and the mixture compositions. Twenty-five mixtures composed of 2-5 contributors at different proportions were analyzed. The true mixture compositions were blinded to avoid possible effect on the interpretation result.

An alternative statistical method based on regression analysis was developed. This method was tested by simulation and on real mixtures by the same parameters as the suggested statistics.

2 MATERIALS AND METHODS

The material and methods section is divided into five parts. The first part is the preparation of the samples, which includes dilutions, degraded samples whole genome amplification and mixtures. The second part is genotyping by the iPLEX technology (sequenom), were a set of dilutions, degraded samples and whole genome amplified samples are analyzed. The purpose of this part of the study is to serve as a guideline for which samples that will perform well in SNP array analysis. The third part is the Illumina GoldenGate analysis in which samples representing variables of dilutions, degraded samples and amplification (found appropriate after the iPLEX analysis) are presented. In addition a set of twenty-five mixtures will be analyzed on the Illumina GoldenGate platform. The fourth part is the STR analysis of the mixtures and the fifth part is the statistical analyses of the mixtures were the theory behind the mixture interpretation is explained.

The methods chapter gives a brief introduction to the theory behind this assignment.

A complete list of vendors for the reagents and equipment used in this study is given in Appendix F.

2.1 DNA samples

An extensive table of all samples analyzed can be found in appendix A and B.

The samples are blood samples from 8 anonymous unrelated individuals, 4 females and 4 males.

All samples were extracted according to the manufactures protocol (Qiagen, EZ1® DNA Blood Handbook) for the EZ1 Blood 350µL blood kit and the DNA blood card on the BioRobot EZ1 (Qiagen). The elution volume was 200 µL. There were in total made 4 extractions at different times, between extractions the blood samples were frozen.

After extraction the concentrations of all samples was measured by Quantifiler® Duo DNA Quantification Kit (Applied Biosystems) on the 7500 Real-Time PCR System (Applied Biosystems). The manufactures protocol was followed (Applied Biosystems, Quantifiler® Duo DNA Quantification Kit, User's Manual).

2.1.2 Diluted samples

For the sequenom iPLEX genotyping two dilution series were prepared in three technical replicates from different samples. The concentrations of the first series were made to be 25, 12.5, 6.25, 3.125, 1.56 and 0.78ng/µL. The concentrations of the second series were 5.56, 1.85, 0.62, 0.21, 0.068, 0.023 and 0.015ng/µL.

One dilution series was prepared for the Illumina GoldenGate analyses, the concentrations in the series were 20, 15, 10, 5, 2 and 1ng/ µL.

All dilutions were made in TE buffer.

After extraction the concentrations of all samples was measured by Quantifiler® Duo DNA Quantification Kit (Applied Biosystems) on the 7500 Real-Time PCR System (Applied Biosystems). The manufactures protocol was followed (Applied Biosystems, Quantifiler® Duo DNA Quantification Kit, User's Manual).

2.1.3 DNAse degradation

A reaction buffer was made from 5.5 µL $CaCl_2$ 1M, 11µL TRIS-HCL 1M, 1µL $MgCl_2$ 25mM and 92.5 µL distilled H_2O. DNA (1000ng) was added to the reaction 110µL buffer together with RNAse-Free DNAse (Qiagen). A sample from the reaction was collected immediately and subsequently after 5min, 10min, 15min and 20min. The reaction was stopped by adding 1.5µL 0.5M EDTA and heating at 60° for 10min. The length of the fragments after degradation was controlled by electrophoresis at 100 V in a 1.5% TBE agarose gel containing ethidium bromide.

The concentration was measured for the first sample and for the samples collected after 5min and 15min by Qubit™ flurometer (Invitrogen) following the manufactures guidelines (Invitrogen, Quant-iT™ dsDNA HS assay). A dilution series of five dilutions was made from the samples degraded for 0min, 5min, and 15min to have the concentrations 13.5, 6.5, 3, 1.5 and 0.5ng/µL.

2.1.4 UV-light degradation

For testing of DNA degradation using the Sequenom iPLEX technology, UV degradation of DNA was performed by directly exposing four samples (50µL DNA (69ng/µL)) to a UV-light

source with sampling for 30minutes, 1, 1.5 or 2 hours. The length of the fragments after degradation was controlled by electrophoresis at 100 V in a 1.5% TBE agarose gel containing ethidium bromide. The concentration after degradation was measured with NanoDrop ND-1000 Spectrometer (Thermo Scientific). A dilution series of five dilutions were made from the samples degraded for 1, 1.5 and 2hours. The concentration of the series was 15, 7, 3, 1.5 and 0.5ng/ µL.

To prepare degraded samples for the Illumina GoldenGate analysis, a sample (50µL DNA (60ng/µL)) was exposed to UV light for 1.5 h. The concentration after degradation was measured with NanoDrop ND-1000 Spectrometer (Thermo Scientific). A dilution series of five dilutions were and the concentration of the series was 50, 10, 5, 2 and 1ng/µL.

2.1.5 WGA samples

Six samples with concentration 4.897, 4.915, 0.021, 0.014, 0.005 and 0.003ng/µL were whole genome amplified by the Repli-g® UltraFast mini kit (Qiagen) following the manufactures protocols (REPLI-g® UltraFast Mini Handbook, Protocol: Amplification of Purified Genomic DNA). After the amplification it is recommended to dilute the samples 1:25, not all achieved concentrations after WGA were adequate for this dilution. Instead 5µL from the samples were diluted to have the concentration of 5ng/µL. In addition the 10µL of the samples were purified by the DNA Clean & Concentrator™-25 (Zymo Research). The clean up reaction was performed according to the manufactures protocol (Zymo Research, DNA Clean & Concentrator-25™ instruction manual).

The concentration for all WGA samples was measured by Qubit™ flurometer (Invitrogen) following the manufactures guidelines (Invitrogen, Quant-iT™ dsDNA HS assay).

For the Illumina GoldenGate analysis seven samples with concentration 6.0, 3.0, 1.6, 0.7, 0.5, 0.2 and 0.05ng/µL were WGA as in the previously mentioned procedure. The total volume of the samples (18µL) was purified after the WGA reaction. No samples were diluted.

2.1.5.1 WGA technology

WGA is a technique used to amplify DNA in low concentration samples. Different from PCR which amplify specific DNA sequences, the purpose of WGA is to amplify the entire genome.

There are several variants of WGA, but the method proven to be the most advantageous is called multiple displacement amplification (MDA) (Dean et al. 2002, Ballantyne et al. 2006). The difference between MDA and other WGA methods is that MDA does not require thermal cycling. Instead the amplification provided by the polymerase from the bacteriophage φ29 will be achieved at 30° C. (Coskun and Alsmadi 2007). In the first step of the procedure the DNA will go through a denaturation process, but no other denaturation and therefore no PCR cycling is needed (http://www.qiagen.com/products/bytechnology/wholegenomeamplification/tutorial/techniques.aspx). The reason is that when moving along the DNA strand the φ29 polymerase displaces the complimentary DNA strand (Lasken and Egholm 2003). As shown in figure 3 the displaced stand can be template for further DNA amplification.

Figure 3 The process of multiple displacement amplification (MDA). Random hexamer primer anneals to the template strand. ϕ29 DNA polymerase moves along the DNA template strand and displaces the complementary strand. The displaced strand becomes a template for further replication (figure from Qiagen, WGA tutorial).

The φ29 polymerase uses random hexamer primers and is capable of synthesizing up to 100kb before dissolving from the template DNA strand (Lasken and Egholm 2003). The high proofreading quality of the enzyme also leads to a low sequence bias, with an error rate of 1 in 10^6-10^7(Coskun and Alsmadi 2007).

2.1.6 Mixtures

The mixtures were created to resemble samples that are realistic in a forensic setting. The mixtures should range from simple mixtures that are easy to interpret to complex mixtures that are not possible to interpret by today's mixture interpretation methods. In that context a simple mixture is a mixture from two contributors were a clear minor and major contributor can be distinguished, while mixtures of four or more contributors are considered complex.

Twenty-five mixtures were created from 2-5 contributors in different proportions. The compositions of all mixtures are listed in table 1.

Table 1. Mixture compositions for all mixtures prepared for genotyping analysis. The rows represent each mixture while the columns represent the contribution to the mixture from each reference sample.

Sample	Individual F	Individual D	Individual B	Individual H	Individual C
MixtureA	0.5	0.5			
MixtureB	0.33	0.33	0.33		
MixtureC	0.25	0.25	0.25	0.25	
MixtureD	0.2	0.2	0.2	0.2	0.2
MixtureE	0.1	0.9			
MixtureF	0.1	0.45	0.45		
MixtureG	0.1	0.3	0.3	0.3	
MixtureH	0.1	0.225	0.225	0.225	0.225
MixtureI	0.05	0.95			
MixtureJ	0.05	0.475	0.475		
MixtureK	0.05	0.317	0.317	0.317	
MixtureL	0.05	0.238	0.238	0.238	0.238
MixtureM	0.3	0.7			
MixtureN	0.2	0.3	0.5		
MixtureO	0.1	0.2	0.3	0.4	
MixtureP	0.15	0.2	0.25	0.3	0.1
MixtureQ	0.4	0.6			
MixtureR	0.45	0.25	0.3		
MixtureS	0.5	0.1	0.15	0.25	
MixtureT	0.7	0.1	0.01	0.09	0.1
MixtureU	0.8	0.2			
MixtureV	0.8	0.05	0.15		
MixtureW	0.3	0.1	0.25	0.35	
MixtureX	0.2	0.25	0.3	0.15	0.1
MixtureY	0.01	0.99			

All mixtures were made to have a total DNA concentration of 50ng/µL.

The mixture compositions were blinded and the mixtures were prepared by an employee at the forensic Institute of Medicine in Oslo. The mixed samples were given new names Blind1-Blind25.

2.2Sequenom iPLEX™ Assay

The dilutions, degraded samples and WGA samples were genotyped in two different iPLEX multiplexes, a 10-plex and a 29-plex. All genotyping where performed according to the iPLEX™ Assay (Sequenom). All samples analyzed with iPLEX and their measured concentrations are listed in the appendix A. The primer sequences and SNPs in the analyses are listed in appendix C.

2.2.1 iPLEX™ technology

The iPLEX™ assay is a SNP genotyping technology that utilizes the mass difference between different alleles to determine which allele is present. The differences are discovered by matrix assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS). These assays can be multiplexed for up to 40 SNPs in one reaction (http://www.sequenom.com/Genetic-Analysis/Applications/iPLEX-Genotyping/iPLEX-Overview.aspx).

The first step of the reaction is a PCR reaction to amplify the DNA region of interest by specific primers. The sample is then treated with shrimp alkaline phosphate (SAP). SAP cleaves the phosphate from unincorporated dNTPs in the sample and converts them to dNDPs which cannot be used in a further reaction. Next, a linear PCR reaction is performed where mass-modified nucleotides- A, T, C and G are present. During the reaction the primer is extended by one of the nucleotides, the incorporation of one of these nucleotides terminates the extension of the primer. The result is an allele specific extension product of different mass depending on the sequence analyzed (iPLEX™ Gold Application Guide, Sequenom). An example of the different masses between two bases is given in figure 4.

Analytes	Peak description	Length of product [bp]	Calculated mass [Da]
Unextended primer	Extension Primer	20	6163.0
Allele 1	Extension Primer + A	21	6434.2
Allele 2	Extension Primer + G	21	6450.2

Figure 4 An example of the mass difference between two alleles extended by one mass modified nucleotide in the iPLEX™ assay. It is this mass difference that is measured and leads to the determination of which allele that is present in the sample (figure from iPLEX™ Gold Application Guide).

The reaction is then desalted by the use of Clean Resin before the sample is printed on a SpectroCHIP (Sequenom) and analyzed by the MassARRAY analyzer (Sequenom).

The mass differences are discovered by MALDI-TOF MS when the target sequence is ionized by a pulse of energy from a laser and then accelerated trough the flight tube of the mass spectrometer. In the flight tube the kinetic energy of the analyte is identical to the work applied to accelerate the ions. The time of flight down the column will then depend on the mass of the target, and the mass can be calculated (Push et al. 2002).

The resulting spectrum for one allele in a heterozygous individual is shown in figure 5.

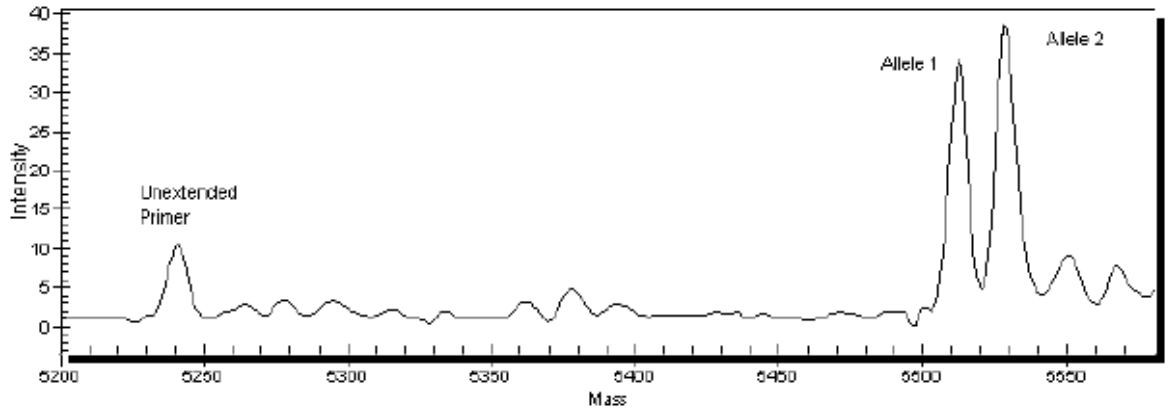


Figure 5. The resulting spectrum from a MALDI-TOF MS analysis. The peaks represent the alleles present far a specific SNP in the sample and the height of the peaks represent the concentration in the sample. Here the individual is heterozygous, both alleles are present. The firs peak is from the unextended primers. (figure from iPLEX™ Gold Application Guide, Sequenom).

2.2 Illumina GoldenGate® Assay

The Illumina GoldenGate® DNA Test Panel includes 360 highly validated single nucleotide polymorphism (SNP) distributed across the genome with all chromosomes represented, including both X and Y for gender verification. All chromosomes are represented with an average 8Mb spacing between loci.

The dilutions, degraded samples, WGA samples and Mixtures were genotyped using the Illumina GoldenGate assay according to the manufacturers instructions (GoldenGate Genotyping Assay Guide, Illumina).

2.3.1 Illumina GoldenGate® technology

The bead array technology from Illumina takes advantage of 3-micron silica beads that can self-assemble in micro-wells contained either on fiber optic bundles or planar silica slides. Several hundred-thousand identical copies of specific oligonucleotides cover each bead, with different oligonucleotide sequences on different beads (http://www.illumina.com/technology/beadarray_technology.ilmn). The beads are mixed in a pool and randomly placed out throughout the array. After the assembly of the beads they are decoded to determine which bead occupies which well (Gundersen et al. 2004).

In the Illumina GoldenGate protocol, SNP specific oligonucleotides will hybridize to the DNA sample prior to the amplification steps. For each SNP locus there are three oligonucleotides, two allele specific oligoes (ASO) and one locus specific oligo (LSO). The two ASOs are perfect compliments to the 20-30 bases upstream from, and including, the SNP site. The LSO is a perfect complement to a series of bases (20-30) starting 1-20 bases downstream from the SNP site. The LSO also contains a sequence of "address" bases which is unique for each SNP assay and allows the assay signal (described later) to be assigned to a particular SNP. All ASO1 oligos share a common primer site, as do all ASO2 and LSO oligos. The hybridization processes were the ASO1 is complementary at the SNP site is shown in figure 6.

Figure 6. The figure displays the hybridization step in the Illumina GoldenGate assay. The oligonucleotides ASO1 and the LSO are hybridized to the DNA at the SNP site (figure from illumnia, Illumina GoldenGate Assay Workflow).

The fact that the hybridization takes place prior to any amplification step prevents amplification bias in the assay. After hybridization the ASO is extended by a polymerase with high specificity for 3´match so it will only extend from perfectly hybridized ASOs and fill the gap between the ASO and the LSO. A subsequent ligation produces a complete, amplifiable template, and means that information about the genotype (in the form of one or both ASOs) is now connected to the bead specific address sequence. The templates are then amplified with universal primers P1 and P2 (specific to ASO1 and 2) which are labeled with Cy3 and Cy5 fluorescent dyes respectively. The amplified dye labeled DNA products are hybridized on to

their complementary bead type through their unique address sequences. An iScan platform is then used to analyze the red and green fluorescent signal from each bead on the array (Shen et al. 2005).

2.3.2 Illumina data analysis

The Genotyping module within the GenomeStudio software (Illumina) assigns genotypes and a confidence score for individual genotyped SNPs. Specifically, a clustering algorithm is used to plot individual genotypes in two-dimensions where X = theta (color), and Y = intensity (signal strength). This is called a gene plot, and example of a gene plot is given in figure 7. Typically, identical genotypes will cluster together in one of three positions (theta = 0.0, 0.5, 1.0) corresponding to genotypes (AA, AB, BB). The SNPs from each sample are placed into their respective clusters and a "Gencall" score (GCS) is calculated representing the SNPs position relative to its respective cluster. The score ranges between 0 and 1 where scores below 0.2 usually is a failed genotype and scores over 0.7 indicate a genotype of good quality (Illumina GenCall Data analysis Software). In addition a call rate is calculated for each sample, the call rate represent the number of SNPs that have been assigned a genotype were 1 indicates all SNPs for that sample and 0 no SNPs.

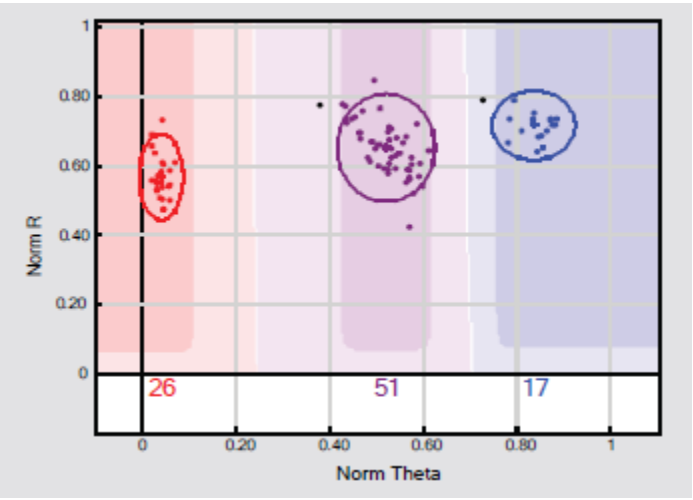


Figure 7. The figure dispays a gene plot. The y-axis represent normalized intensity (sum of intensity from the two channels, denoted Norm R) and the x-axis the normalized theta where theta values close to 0 are homozygous for the "A" allele and theta values close to 1 are homozygous for the "B" allele. The software groups the SNPs in to their respective genotype class where red is AA, purple AB and blue BB.

2.4 STR analysis

The mixtures were genotyped by the AmpFlSTR® SGM Plus® PCR Amplification kit from according to the manufactures protocol (Applied Biosystems, AmpFlSTR® SGM Plus®PCR Amplification Kit User's Manual). The amplification was performed on the 96-Well GeneAmp® PCR System 9700 (Applied Biosystems) and the electrophoresis was performed on the 3130 Genetic Analyzer. All data from the analysis were analyzed on the GeneMapper® ID-X, and the GeneMapper® Software (Applied Biosystems).

2.5 Statistical methods

Some traditional statistical methods (t-test, confidence interval) have been used. These were implemented in R (www.**r-project**.org) and excel.

The mixtures were analyzed by two statistical methods, the suggested method (Homer et al. 2008) and an alternative method based on linear regression. The alternative method was developed by Thore Egeland[4] for this assignment. For both methods all calculations were performed in R. For the STR genotypes the classical mixture interpretation approach used at the forensic institute will be followed (Clayton et al. 1997).

2.5.1 The use of raw allele intensity analysis in SNP genotyping (Statistical method 1)

When results from a SNP genotyping analysis are analyzed by the standard clustering algorithm described for the Illumina Data Analysis Software (section 2.3.2) the samples can be assigned the genotypes AA, AB and BB. The genotype for individual SNPs from a mixture are not expected to fit in the calling algorithm, since several genotypes will be present in one sample. An example is a SNP in a two person mixture where one person is homozygote AA ($\theta = 0.0$) and the other person heterozygote AB ($\theta = 0.5$). The combined genotypes (AAAB; $\theta = 0.25$) will give an uneven relationship between the A and B alleles, and in the data plot the sample will be placed between the AA and AB clusters. Depending on the SNP calling stringency, data plots falling outside the clusters will either be miss-called (as AA or AB), or no genotype will be assigned for that SNP. Thus the standard SNP calling software is inappropriate for assigning genotypes in mixed samples. However, raw fluorescence measurements for every SNP, allele frequencies may be estimated.

[4] Research scientist (statistician), Forensic Institute of Medicine, Oslo.

From the genotyping analysis, the raw fluorescence measurements for each allele in the SNPs are available. These raw data represent the amount of A and B alleles present in the sample and an allele frequency estimate for the A-allele can be calculated as

$$M_i = \frac{A_i}{A_i + k * B_i}. \quad (1)$$

Here A_i and B_i represent the raw fluorescence measurements for each SNP and k is a correction constant for the variation in the measurements. The frequency of the B allele can be calculated as 1-A.

According to Homer et al. (2008) the presence of an individual with known genotype can be determined in a mixture where the allele frequency estimates from the raw fluorescence measurements are available. The method described in Homer et al. (2008) (statistical method 1) is based on comparing the allele frequency estimates from a known individual to the allele frequencies in a reference population and in a mixture, where the reference population and the mixture are drawn from the same population.

The statistical method is illustrated in figure 8 where Y represents the allele frequency from an individual under investigation, P the allele frequency found in a reference population and M the allele frequency in the mixture calculated as (1).

For each SNP j in the analyses a genetic distance measurement (D (Y_j)) can be calculated

$$D(Y_j) = |Y_j - P_j| - |Y_j - M_j| \quad (2)$$

For all SNPs analyzed, the size and sign of D(Y) will be an indication as to whether the individual of interest is part of the mixture or not.

Because the reference population and the mixture are drawn from the same population and will be of similar genetic composition, the distance from Y to M and P is expected to be close when Y is not in the mixture. Thus under the null hypothesis the individual is not in the mixture, D(Y) is expected to be close to zero.

When the individual is present in the mixture the contribution from Y in M will make the distance small and D(Y) will be positive. For the alternative when Y is closer to P D(Y) will be negative.

Figure 8 illustrates the tree cases: a) Y is closer to the reference population and D(Y) is negative, b) Y is equally distant to the reference population and the mixture and D(Y) is zero and c) Y is closer to the mixture and D(Y) is positive.

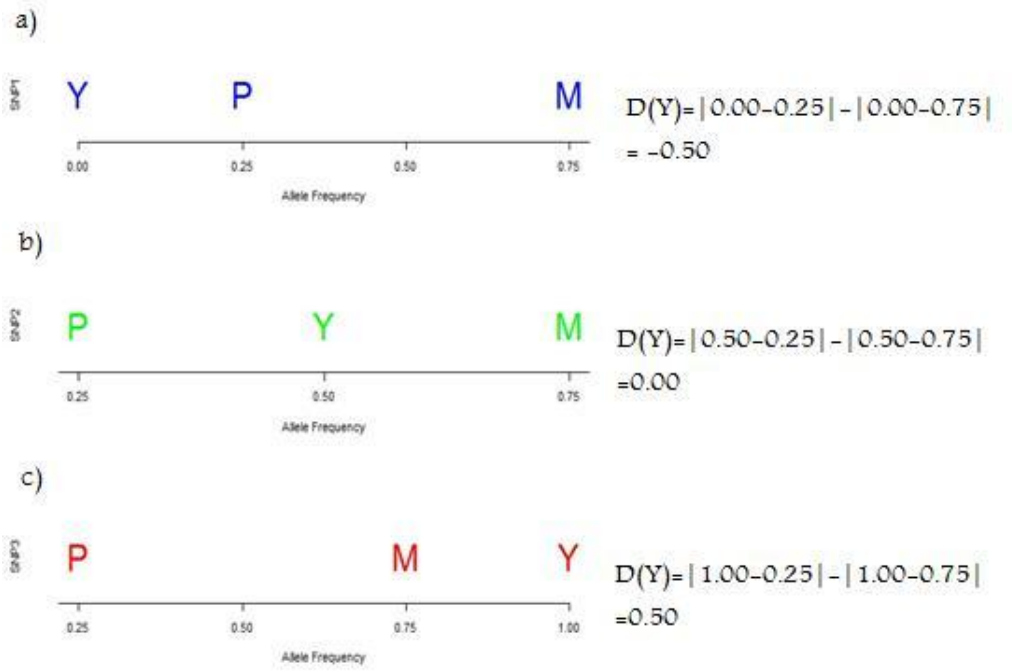


Figure 8. The figure illustrates the distance from the allele frequency estimate for an individual under investigation (Y) to the allele frequencies found in a mixture (M) and a reference population (P) in the three possible situations. In the first case Y is closest to the reference population and D(Y) is negative. In the next case Y is equally distant to the mixture and the reference population and D(Y) is zero. In the last case Y is closer to the mixture and D(Y) is negative (Homer et al. 2008).

By sampling a large amount of SNPs D(Y) is expected to follow a normal distribution due to the central limit theorem. A one sample t-test is conducted for the individual across all SNPs to obtain the test statistics:

$$T(Y_i) = \frac{\bar{D}}{SD(\bar{D})/\sqrt{s}} \qquad (3)$$

Under the null hypothesis T is expected to be close to zero, and in the alternative hypothesis T is expected to be positive. For the third case the sample is closer to the reference population and T will be negative.

Homer et al. (2008) assume that T is approximately normally distributed with expectation 0 under the null hypothesis. We reject the null hypothesis if T <- 2 or T > 2. This corresponds to a significance level of roughly 5%.

A small example of the calculations is: a reference sample has the genotypes BB, BB and AB for the SNPs 1-3. The allele frequency estimate for the B allele of reference will be 1, 1 and 0.5 for the 3 SNPs respectively. The corresponding allele frequency estimate from a mixture for the 3 SNPs was calculated to be 0.6, 0.9 and 0.7, and the allele frequency for the B allele in the population is 0.5, 0.7 and 0.3.The calculated distance measurement D(Y) for each SNP will be 0.1, 0.2 and 0.T (Y) for this sample will be 2.5, and the reference is expected to be in the mixture.

The reference samples were analyzed using the Illumina Genotyping Module and the genotypes were tabulated in a single report. The genotypes for each reference sample were transformed to B allele frequencies, 0.0, 0.5 and 1.0 for AA, AB and BB respectively.

The genotype population frequencies are allele frequencies collected from the HapMap database from American citizens with European ethnicity. The B allele frequencies were used for all SNPs. Allele frequencies for all SNPs are listed in the Appendix.

The allele frequency estimate for the mixtures were calculated from the raw data measurements in collected from the output file in the software analyses. The frequencies were calculated for the B allele.

D(Y) and the test statistics were calculated for the possibilities for all references being part of all mixtures.

2.5.2 Alternative statistics (Statistical method 2)

A statistical method 2 was developed to determine if an individual was present in a mixture, and to estimate the contribution from the individual in the total mixture. The calculations will be made for two cases; the simple data were the intensity measured is equal for all SNPs and the real data were the intensity measurements will differ between SNPs.

Case 1; Simple data

From the raw fluorescence data collected in a SNP microarray genotyping analysis, the signal measured for allele A at SNP_i ($Y_{i,A}$) will be proportional to the number of A alleles present over all individuals in the sample. This can be written as

$$Y_{i,A} = \beta x_{i,A} + (1-\beta)\mu_{i,A} + noise \quad (4)$$

were β is the fraction of the mixture originating from the individual of interest (I), $x_{i,A}$ is the number of A alleles originating from I, $\mu_{A,i}$ is the expected contribution of A alleles from an unknown individual. If p_i is the frequency of the A allele for the i'th SNP in the population

$\mu_{A,i} = 2p_i$

When removing the noise, equation 1 can be written

$$Y_{i,A} - \mu_{i,A} = \beta(x_{i,A} - \mu_{i,A}) \quad (5)$$

or

$$Z_{i,A} = \beta(x_{i,A} - \mu_{i,A}) \quad (6)$$

which is a simple linear regression with no constant.

Equally, we find for the B allele

$$Y_{i,B} = \beta x_{i,B} + (1-\beta)\mu_{i,B} + noise \quad (7)$$

To test if the individual is present in the mixture is equal to testing the null hypothesis H0: β = 0 can be performed. β reflects the contribution from the individual in the mixture.

Case 2; Real data

In the analysis of mixtures from raw data the differences in the two color channels must be considered and a correction constant C_1 is added to the formula:

$$Y_{i,A} = C_1[\beta x_{i,A} + (1-\beta)\mu_{i,A}] \quad (8)$$

(For $Y_{i,B}$ a similar equation applies with C_1 replaced by C_2)

The estimation of C_1 is explained below.

From the above equation follows:

$$\sum_{i=1}^{n} Y_{i,A} = C_1 \sum_{i=1}^{n} [\beta x_{i,A} + (1-\beta)\mu_{i,A}] \quad (9)$$

$$\sum_{i=1}^{n} Y_{i,A} = C_1 \beta \sum_{i=1}^{n} x_{i,A} + C_1(1-\beta) \sum_{i=1}^{n} \mu_{i,A} \quad (10)$$

Because

$$\sum_{i=1}^{n} x_{i,A} = n\frac{1}{n}\sum_{i=1}^{n} x_{i,A} = n\bar{x} \quad (11)$$

the equation can be written like

$$n\frac{1}{n}\sum_{i=1}^{n} Y_{i,A} = C_1 \beta n \bar{x}_A + C_1(1-\beta)2n\bar{p} \quad (12)$$

$$\bar{Y}_A = C_1 \beta \bar{x}_A + C_1(1-\beta)2\bar{p} \quad (13)$$

if we replace $\bar{x}_A$ by the expected value $\bar{p}$

$$\bar{Y}_A = C_1 \beta 2\bar{p} + C_1 2\bar{p} - C_1 \beta 2\bar{p} \quad (14)$$

from which follows

$$C_1 = \frac{\bar{Y}_A}{2\bar{p}} \quad (15)$$

a similar argument gives

$$C_2 = \frac{\bar{Y}_B}{2(1-\bar{p})} \quad (16)$$

2.5.3 Interpreting a mixed sample in STR analyses

To analyze a mixture it is necessary to decide which alleles that origins from the same individual. The mixture proportion is a measure of the contribution of DNA from each individual present in a mixture. When the individuals have contributed different amounts of DNA to the mixture they can be separated by the heights of the peaks in the epg.

The mixture proportion can be calculated as in (17) where ϕ is the peak height. A and B represent the major contributor and C and D represent the minor

$$M_X = \frac{\phi_A + \phi_B}{\phi_A + \phi_B + \phi_C + \phi_D} \quad (17)$$

The mixture proportions of the contributors are expected to be preserved throughout the mixture at each locus, but can be imprecise .The M_X will be affected by factors like degradation, stutter, stochastic variation and low DNA concentrations. The variability of M_X can in some cases be as high as ±0, 35 (Buckleton et al. 2005).

The heterozygote balance (H_B) is a measure of the difference in height between the two allele peaks of a heterozygote, and can be calculated as

$$H_B = \frac{\phi_A}{\phi_B} \quad (18)$$

Where ϕ_A is the size of the smallest peak (Bill et al. 2004).

From experimental data H_B for two alleles from the same individual will be > 0.6 for samples that are not degraded and have concentrations of more than 500pg (Gill et al. 2006).

When interpreting a mixture all possible combinations of genotypes shall be considered in relation to the mixture proportion and the heterozygote balance across all loci. Those combinations that are not supported by the guidelines formulated by these two parameters are considered to have a low posterior probability and are removed.

If the genotype of interest is the minor components the interpretation is more complex since other considerations like drop out, stutter and masking by major alleles are necessary.

The software Genmapper ID-X from AppliedBiosystems provides M_x and H_b calculations for mixtures and ranks the most probable combinations of genotypes.

When the most probable genotypes are decided the last step of the interpretation is to compare the mixture to the reference samples (Gill et al. 2006).

2.5.4 Simulation

Simulation is a general tool of great importance, where we can test theories and methods on data sets were the answers are already known. In this way the reliability of a method can be evaluated theoretically. The drawback is that simulation is under perfect conditions and experimental noise will not be accounted for.

A simulation routine was created in R to investigate which contributors that were expected to be identified in mixtures of different composite. The variables in the routine where the number of contributors, the proportions given to the mixture from the different contributors, the allele frequencies and standard deviation of signals and the number of SNPs to simulate.

The first part of the routine was to create genotypes for each individual that was in the mixture. The genotype at each SNP could be 11, 12 or 22, and were drawn by the probability given in the allele frequencies for the 1 and 2 alleles.

To resemble the raw intensity measurements used to calculate allele frequencies in the data from microarray SNP analyses, the genotypes were converted to heights for each allele. The heights where equal to the proportion of each sample to be in the total mixture. In the end the mixture was created as the sum of heights over all individuals present for each allele for each SNP.

To simulate possible noise from the analysis a standard deviation parameter could be set and the variables were calculated from the normal distribution to adjust the frequencies at each individual SNP.

3 RESULTS AND DISCUSSION

This section can be divided in two parts, DNA quality analysis and Mixture interpretation. The DNA quality samples were first analyzed using the iPLEX® assay to determine which samples were excessively degraded, diluted or amplified. The remaining samples were subsequently analyzed with the Illumina GoldenGate® assay. The results from the two technologies are divided into separate sections. In the mixture interpretation section the simulation studies from the suggested and statistical method 2 will be reviewed first. Then the results from the interpretation of the twenty-five mixtures analyzed by both statistical methods will be presented. The results from the STR analysis of the mixtures will be presented last. The true mixture compositions were revealed when all mixtures were interpreted; recall that this was a blinded experiment. In the end of this section all mixture interpretation results will be compared to the true compositions.

3.1 Sequenom iPLEX Analysis

The samples analyzed by the Sequenom iPLEX assay were dilution series, samples degraded by DNAse, samples degraded by UV-exposure and WGA samples. This section will give an overview and a short discussion of the results from the degradation procedures, WGA procedure and the genotyping results.

The SNPs that didn't give a genotyping results across all samples in a run where excluded from the analysis.

3.1.1 Dilutions

The dilutions analyzed had concentrations in the range of 0.003-28.8ng/µL, the recommended concentration for the Sequenom iPLEX analysis is 5-10ng/µL. The success of the genotyping analysis was measured in percent of genotyped markers assigned for each sample. For the 10-plex assay the lowest DNA concentration generating a 100% genotyping result was 0.05ng/µL. For the 29-plex assay the lowest DNA concentration that gave a 100% genotyping result was 0.15ng/µL. The combined genotyping success for both multiplexes is described in figure 9.

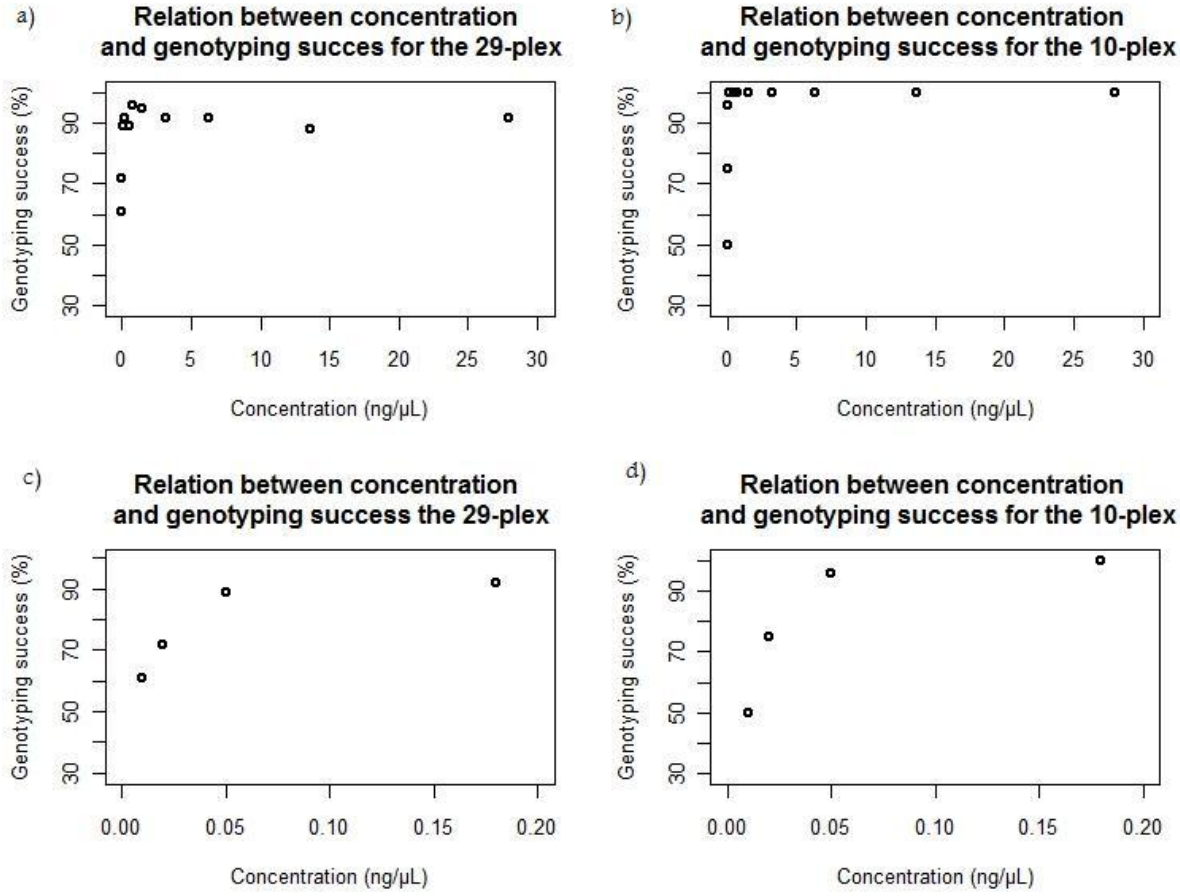


Figure 9. The figure illustrates the genotyping success in the dilution series analyzed by the iPLEX (sequenom) assay by two genotyping multiplexes (10 and 29 SNPs). The genotyping success is measured in percent of genotyped SNPs. The genotyping success of the two multiplexes is shown in separate figures where a) represents the 29-plex and b) represents the 10-plex. c) and d) give a view of the samples with the lowest DNA concentration were the genotyping success is starting to decrease.

In the samples with concentration 0.05ng/µL or less there were observed several incidences of genotyping artifacts in the form of allele drop out at one of the alleles in a heterozygous individual. These artifacts will make the interpretation more difficult and are not desirable in a forensic context (section 1.2).

Both SNP multiplexes achieved 100% genotyping success from DNA quantities much lower than what is recommended for the analysis. The results are very similar to results achieved in STR genotyping, which could be expected since the reactions both are based on PCR. However the samples analyzed are good quality blood samples and are not directly comparable to samples collected from a crime scene.

A SNP multiplex of 52 SNPs has previously been designed for forensic identification purposes. In this study a multiplex PCR amplification step provided successful amplification from as little as 0.5ng DNA. The optimal amount of DNA in the PCR was 1–10ng (Sanchez et

al. 2006). In this study a successful amplification has been achieved from DNA concentrations below that reported by Sanchez et al. (2006), however the multiplexes are composed of fewer SNPs (10 and 29). Larger DNA quantities are required when the SNP multiplexes get higher. Since there is 360 SNPs in the Illumnia GoldenGate analysis it is not expected to get results from as low DNA quantities mentioned above. 1ng/µL was decided as the minimum DNA input.

3.1.2. Degraded by DNAse

The DNAse degraded samples were treated with DNAse and collected immediately and subsequently after 5, 10, 15 and 20 min. The length of the fragments after degradation was controlled by electrophoresis on an agarose gel shown in figure 10. The samples degraded for 5, 10 and 15 min were visibly degraded, and these samples were subsequently genotyped.

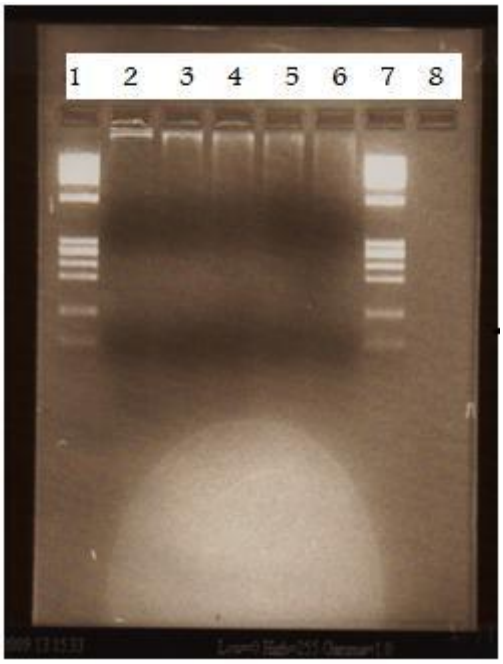
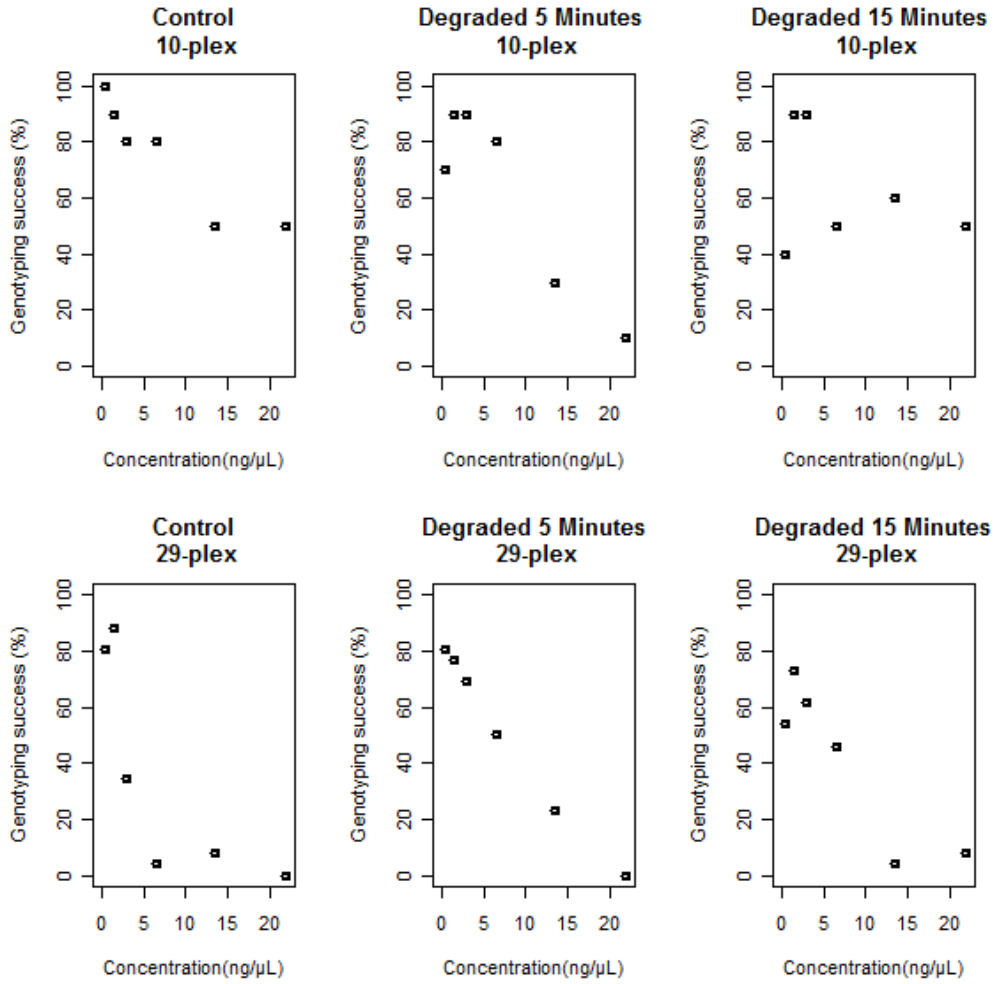


Figure 10 Electrophoresis gel after degradation of Genomic DNA samples by DNAse treatment. Well 2 contains the positive control, well 3 DNA degraded for 5 min, well 4 10 min, well 5 15 min and well 6 20 min.

The success of the genotyping analysis was measured in percent of genotyped markers assigned for each sample. The genotyping success for each sample in the dilution series was variable. It is observable that the genotyping success is higher for the more diluted samples (figure 11). This is also the case for the control sample that only was added the degradation buffer and no DNAse. This indicated that the EDTA present in the samples could be a

contamination preventing the genotyping analysis. There were also observed genotyping artifacts in the form of drop out at one allele in several SNPs.

The genotyping success rates for the samples in both multiplexes are shown in figure 11.



Figur 11. The figure illustrates the genotyping success for the samples degraded by DNAse and analyzed by the iPLEX (sequenom) assay by two genotyping multiplexes (10 and 29 SNPs). The genotyping success is measured in percent of genotyped SNPs. The genotyping success of the two multiplexes is shown in separate figures for the control sample and the samples degraded 5 and 15min in dilution series. It is observable that the genotyping success is increased by the level of dilution in the samples.

Asari et al. (2008) report successful genotyping of DNAse degraded DNA by SNP genotyping (target sequences in the range 40-67 bp). This indicates that DNAse could be an appropriate way to degrade DNA.

In this study the degraded samples that were diluted performed best (90 and 80% genotyping success for the 10 and 29-plex respectively), but there were observed several incidences of genotyping errors. It was therefore decided to degrade the samples by a UV approach.

3.1.3 Degraded by UV exposure

The samples were degraded by UV-exposure for 30, 60, 90 and 120 minutes. The level of degradation was controlled in the samples by gel electrophoresis. The result is shown in figure 12. The samples degraded for 60, 90 and 120 min were visibly degraded. These samples were subsequently genotyped.

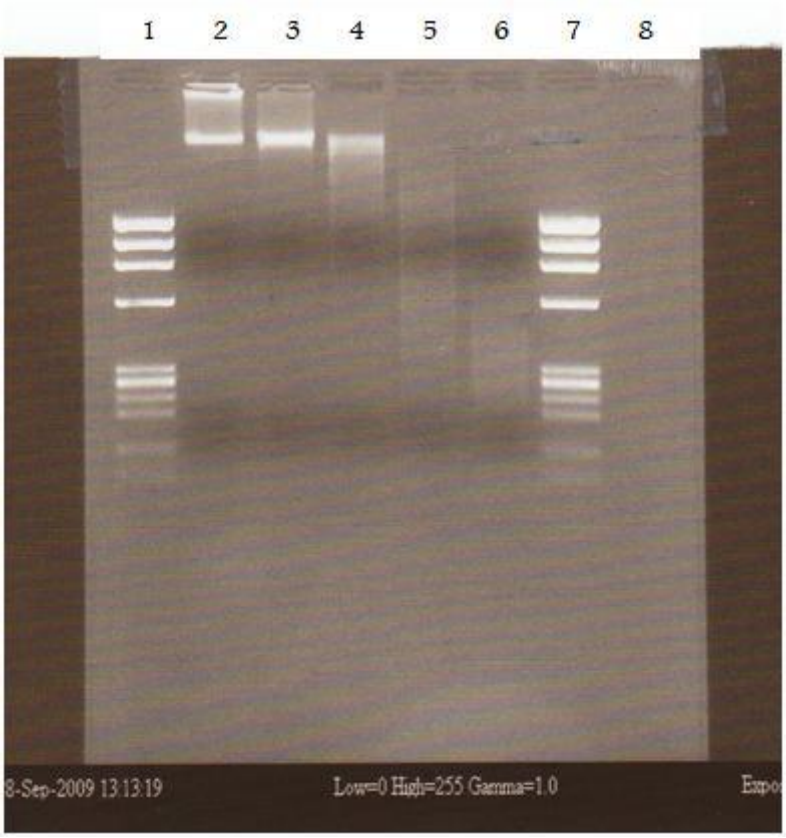


Figure 12. Electrophoresis gel after degradation of Genomic DNA samples by UV-light treatment. Well 2 contains the positive control, well 3 DNA degraded 30 min, well 4 60 min, well 5 90 min and well 6 120 min.

The success of the genotyping analysis was measured in percent of genotyped markers assigned for each sample. All the degraded samples, regardless of concentration have very low genotyping results (on average less than 30%). The genotyping result for all degraded samples and a positive control not exposed to UV light, are illustrated for both genotyping multiplexes in figure 13.

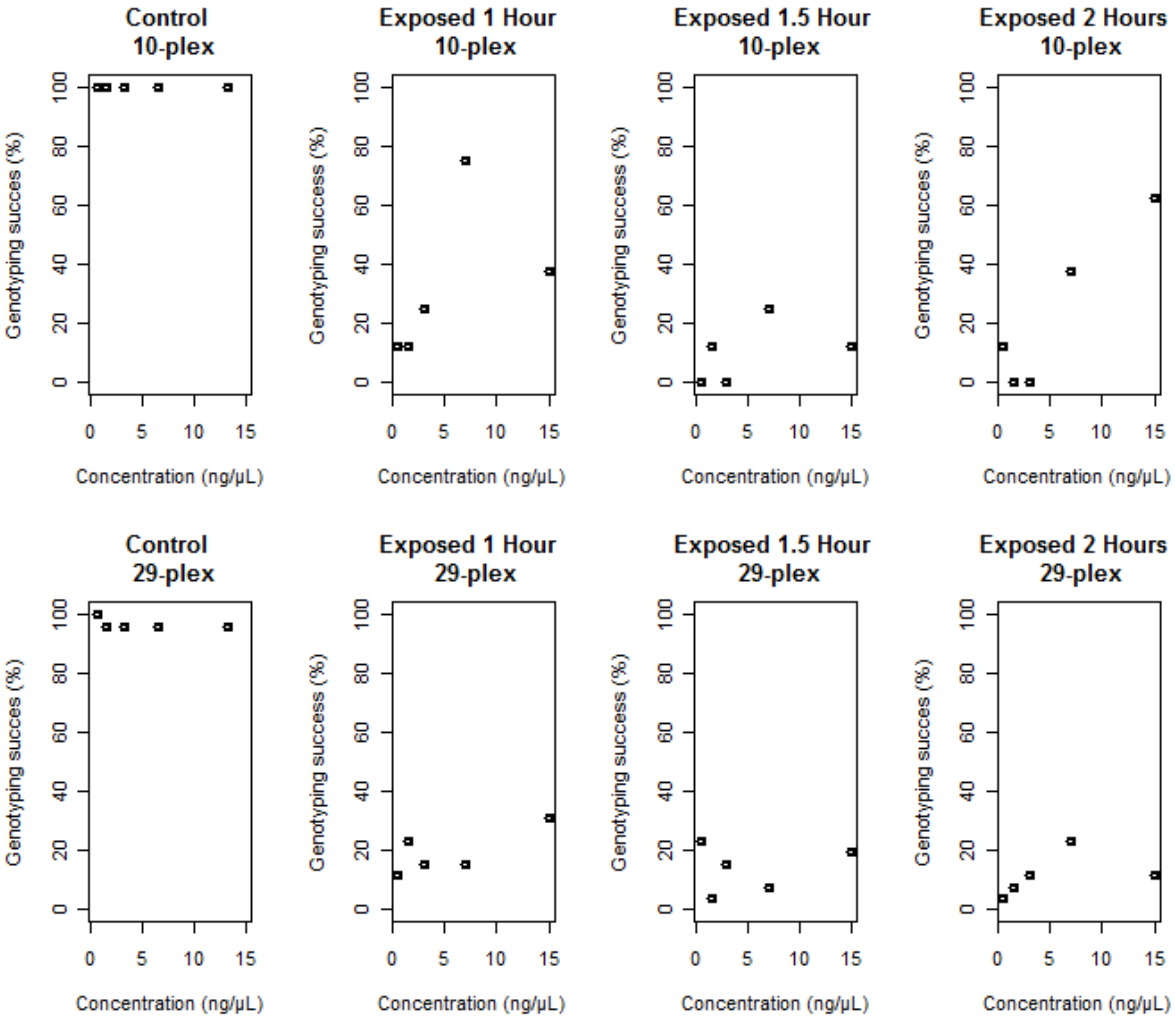


Figure 13. The figure illustrates the genotyping success for the samples degraded by UV light and analyzed by the iPLEX (sequenom) assay by two genotyping multiplexes (10 and 29 SNPs). The genotyping success is measured in percent of genotyped SNPs. The genotyping success of the two multiplexes is shown in separate figures for the control sample and the samples degraded 1, 1.5 and 2 hours in dilution series.

3.1.4 Whole Genome Amplified samples

Six samples were amplified. Two samples had concentrations that were expected to be appropriate for the procedure (5ng/µL). Four samples had concentration less than 0.05ng/µL, which represent the threshold for 100% genotyping success without prior amplification (section 3.1.1).

The concentration before and after the WGA procedure is listed in table 2. The samples with adequate concentrations were diluted and purified. All samples were genotyped.

Table 2 The concentrations of the samples before and after WGA as measured by Qubit™Flutometer (Invitrogen)

Sample	Concentration prior to WGA (ng/µL)	Concentration after WGA (ng/µL)
1	4.890	282.000
2	4.920	289.000
3	0.020	13.200
4	0.002	12.500
5	0.005	10.600
6	0.003	3.060

The success of the genotyping analysis was measured in percent of genotyped markers assigned for each sample. The results for all samples are visualized in figure 14 for the 10-plex and figure 15 for the 29-plex. Separate bars in the histogram will represent the purified and the diluted samples. Sample 5 and 6 were not purified and sample 6 was not diluted. It is apparent that the genotyping result is improved for the diluted and purified samples and that the genotyping success increases with the concentration prior to WGA.

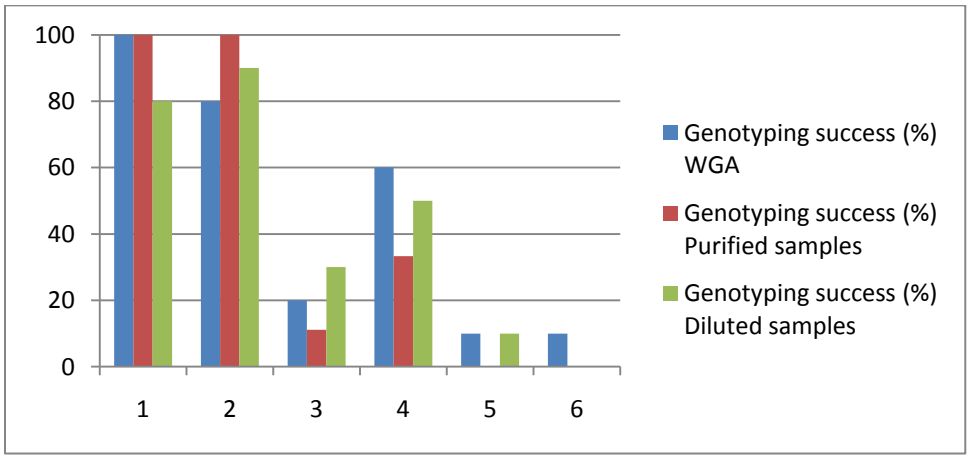


Figure 14. The genotyping success measured in percent of genotyped SNPs for the WGA samples analyzed by the 10-plex. Samples were as described in Table 2. The blue pole represents the WGA samples, the red poles are the samples purified after WGA and the green poles are the samples diluted after WGA. Sample 5 and 6 were not purified and sample 6 was not diluted. The concentrations of the samples can be found in table 2.

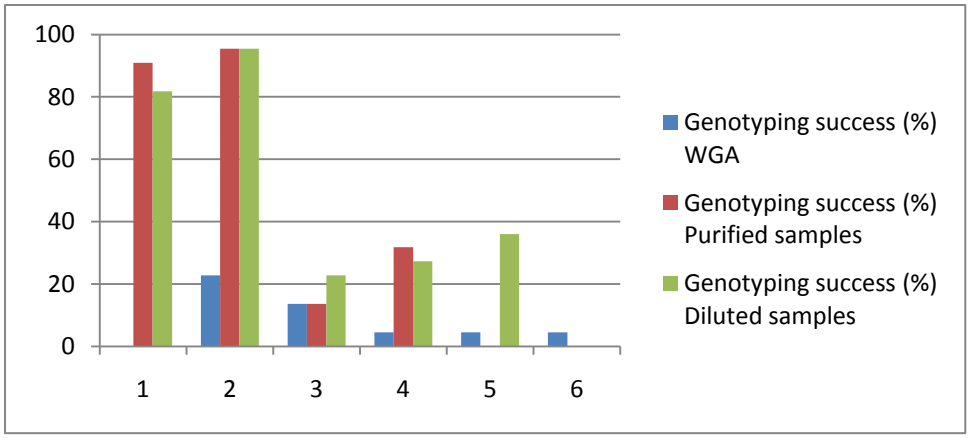


Figure 15 the genotyping success measured in percent of genotyped SNPs for the WGA samples analyzed by the 29-plex. Samples were as described in Table 2. The blue pole represents the WGA samples, the red poles are the samples purified after WGA and the green poles are the samples diluted after WGA. Sample 5 and 6 were not purified and sample 6 was not diluted. The concentrations of the samples can be found in table 2.

The occurrence of allele drop out was observed in several cases in the genotypes of the samples, there were also several observations of drop in of alleles (occurrence of alleles not present in the positive control for the samples). A discussion of drop out and drop in and the relevance for forensic case work can be found e.g. in Buckleton et al. (2005).

The amplification procedure seems to be inappropriate for the samples with concentrations less than 0.05ng/µL. Although the amount of measured DNA in the sample has increased in a matter of 600-1000fold, the samples have very low genotyping success (figure 14 and 15). This could imply that for this low concentration the amplified DNA will not represent the whole genome. It was decided to amplify samples with higher concentrations (starting at 0.05ng/µL) for the Ilumina GoldenGate analysis.

3.3 Illumina GoldenGate analysis

A dilution series, a dilution series of degraded samples, amplified samples of different concentrations and twenty-five mixtures were analyzed on the Illumina GoldenGate assay. The raw fluorescence data were directly extracted for the mixtures and will be used in the mixtures section of the results.

The presentation of the results will be described and the limitations of the Illumina GoldenGate assay experienced in the DNA quality analyses will be discussed. This will be measured by the average call rate and the average GCS for each sample (section 2.3.2).

The samples in the analysis were not of appropriate quality to create a cluster profile for the genotype calling. A standard cluster file was therefore downloaded from the Illumina home page (See appendix F for a link).

A list over all samples analyzed, measured concentrations and genotyping results can be found in appendix B.

3.3.1 Dilutions

The concentrations of the samples in the dilution series analyzed were 20, 15, 10, 5, 2 and 1ng/µL. The result of the genotyping is presented in figure 16 as the relation between concentration and call rate and the relation between concentration and GSC. It is observable that more than 95% of the SNP will be called for a sample until the concentration drops to 5ng/µL, then the call rate starts to decrease. The GSC is very similar for all samples except the sample with concentration 1ng/µL were it drops to 0.45.

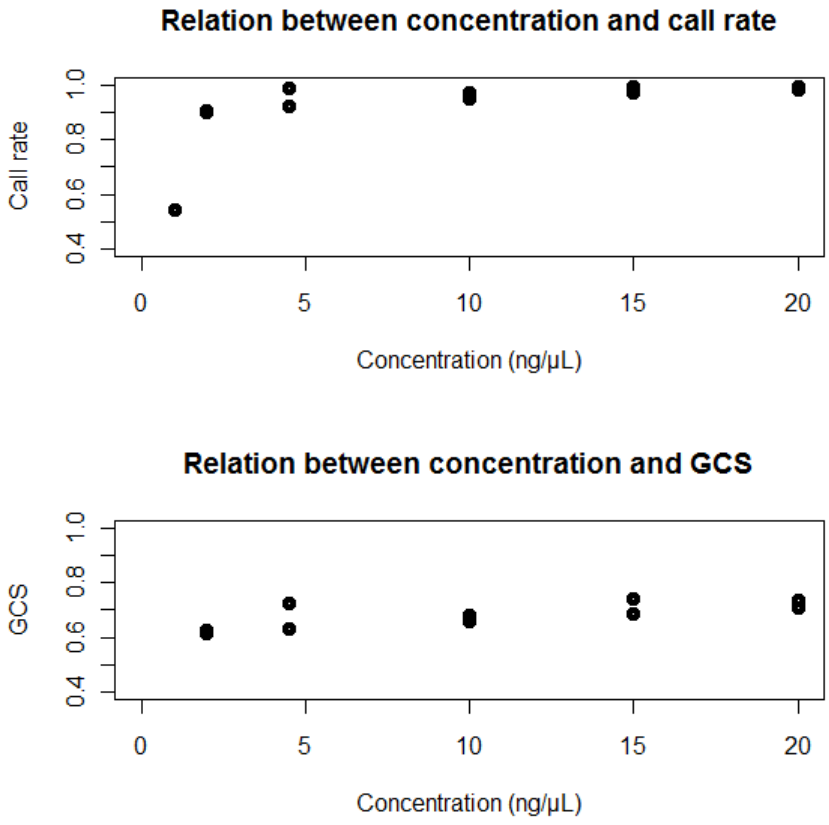


Figure 16. The figure illustrates the result of the dilution series analyzed by the Illumina GoldenGate assay. The results are presented as call rate (top) and the quality of the genotype call (GCS). The success of the analysis starts to decrease at 5ng/µL.

For the Illumina GoldenGate 360 SNP test panel the total genotyping results were not of suitable quality to form an appropriate cluster for the gene calling. A standard cluster file was downloaded from the Illumina home page[5]. None of the high quality reference samples received optimal Call rate or GCS from the analysis (call rate and GCS is explained in section 2.3.2). A reason for this could be that the clusters did not give an optimal fit for these samples. This could also affect the results of the dilution series.

In the dilutions the genotyping success started to decrease in one of the parallels at 4.5ng/µL (call rate 0.91, see table in appendix B) and kept on decreasing for the two next concentrations (figure 10). This indicates that samples with concentration less than 4.5ng/µL will not be appropriate when it comes to gene calling by cluster analysis. Krutskov et al. (2009) analyzed DNA in a series of 100, 50, 25 12.5 1.0 and 0.5ng per reaction by a 124 SNP microarray and found that 50ng per reaction was necessary to achieve a 99% call rate. For samples of low DNA quality (less than 2ng/µL) the cluster based high-density genome-wide SNP arrays appears not to be a suitable genotyping technique. This could be because the analyses are based on higher DNA concentrations and were not made for forensic use. To be appropriate for such samples the analysis must be adjusted to be more suitable for the low DNA concentration samples. The technique is not available today but with the rapid improvement and development in techniques that is happening it is not impossible that hundred or thousand of SNPs can be analyzed from small DNA concentrations.

3.3.2 Degraded Samples

The degraded samples are samples exposed to a UV-light source for 1.5h in a dilution series. All the degraded samples, regardless of concentration have very low call rate (less than 0.35) and GSC below 0.2. The result of the genotyping is presented in figure 17 as the relation between concentration and call rate and the relation between concentration and GSC. There were observed genotyping artifacts in the form of drop in and drop out of alleles in all degraded samples.

[5] http://www.illumina.com/

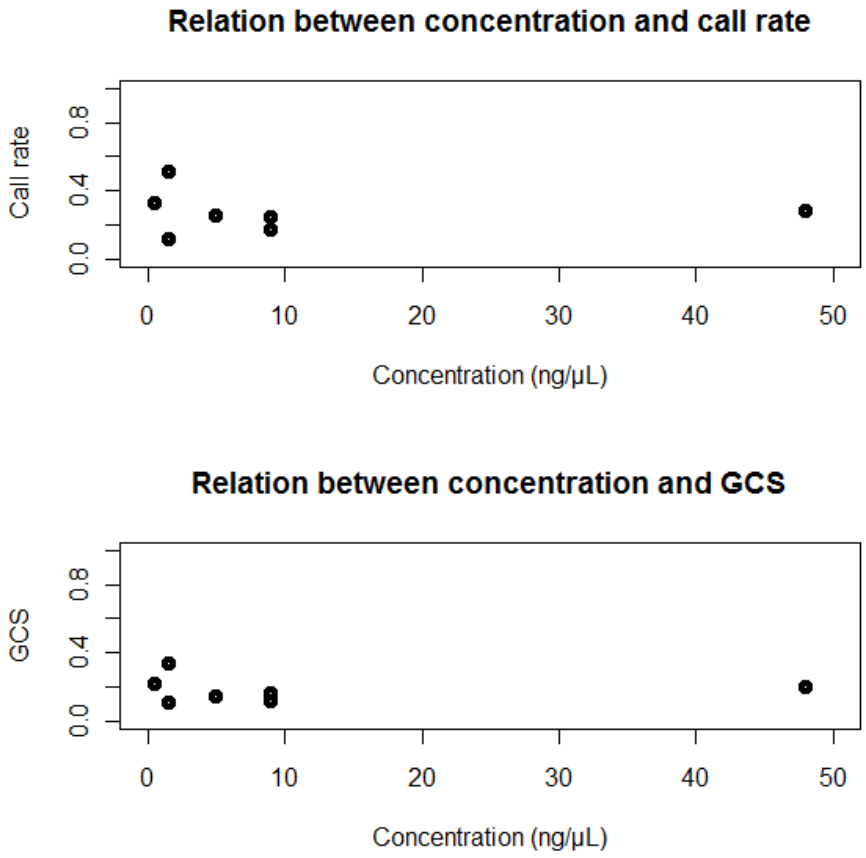


Figure 17.The figure illustrates the results of the degraded samples in dilution series analyzed by the Illumina GoldenGate assay. The results are presented as call rate (top) and the quality of the genotype call (GCS).

The sample that gave best results was degraded 4 (2ng/µL) parallel 2 (see table in appendix B). The sample had a call rate of 0.52 and GCS of 0.34. All other samples had a call rate under 0.33 and GCS under 0.22. These values are similar to the negative (no DNA template) control run for these analyses (Table in appendix B). Samples with results that are similar to a negative control cannot be considered as a contribution from the samples and is an indication that the DNA could have been too degraded for the genotyping analyses. Preparation of suitable degraded samples was a challenge and required several attempts. Several methods have been tested to prepare such samples and an alternative approach could be tested (Bender et al. 2003, Dixon et al. 2005). More tests should be done to be able to conclude about the performance of the Illumina GoldenGate assay on degraded samples.

3.3.3 Whole Genome Amplified samples

Seven samples with concentrations 6.0, 3.0, 1.6, 0.7, 0.5, 0.2 and 0.05ng/µL, a positive and a negative control were amplified. The result of the WGA is illustrated in figure 18 as the relation between concentration before and after the amplification.

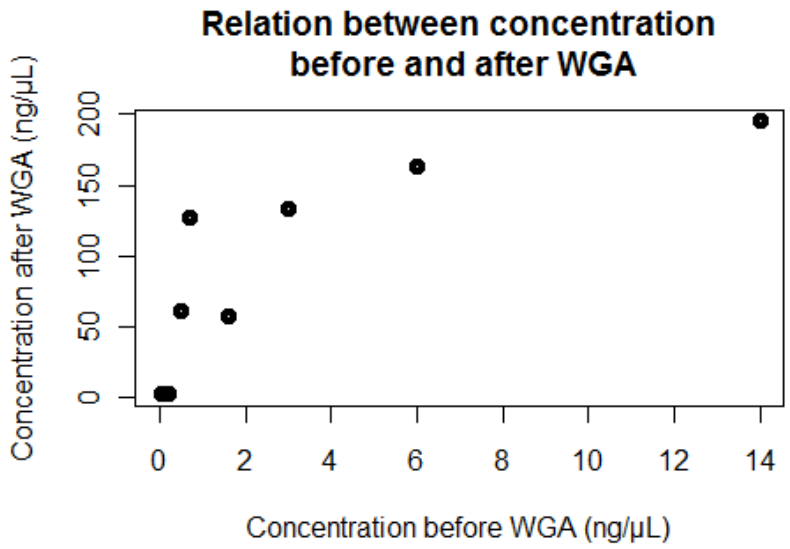


Figure 18. Illustration of the relation between the concentration of the samples prior to WGA and the concentration measured by Qubit™ flurometer after WGA.

All the WGA samples were purified by the DNA Clean & Concentrator™-25 (Zymo Research) before they were genotyped, and the concentrations of the samples were reduced in some degree by the purification. It is the new concentration that is considered in the results. All the samples except the positive control had a call rate of less than 0.8 which was deceasing by the concentration of the sample. The result of the genotyping is presented in figure 19 as well as the relation between concentration and call rate and the relation between concentration and GSC.

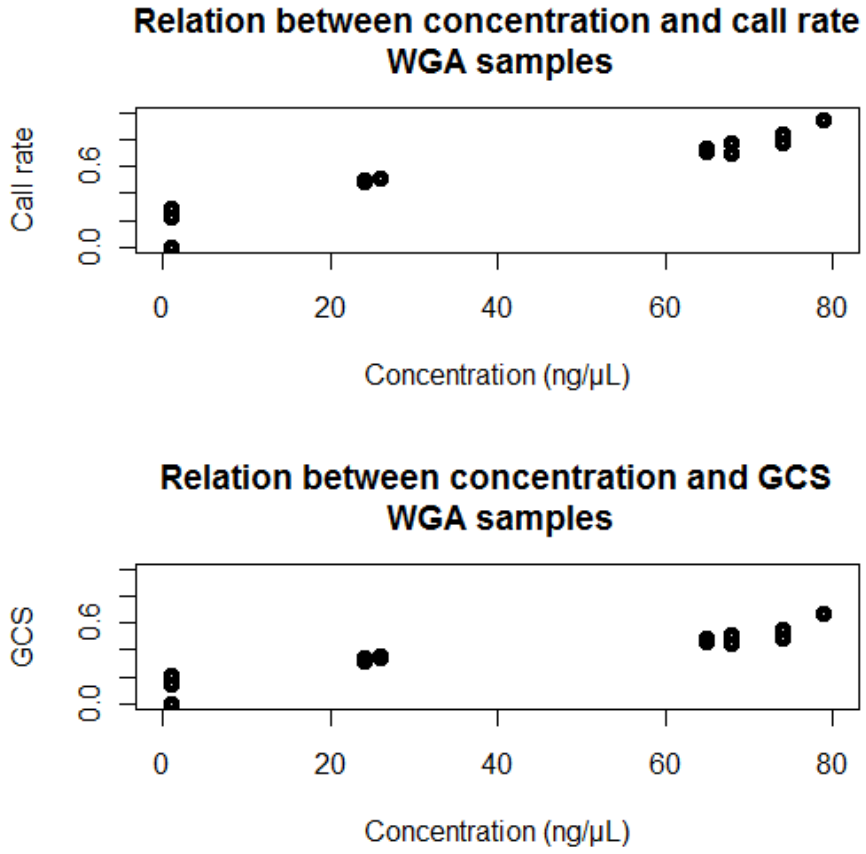


Figure 19. Illustates the result for the WGA samples analyzed by the Illumina GoldenGate assay. The results are presented as call rate (top) and the quality of the genotype call (GCS). The concentration is the concentration measured by Qubit™Flutometer (Invitrogen) after the samples were purified.

The recommended DNA concentration in a sample is 50ng/µL for Illumina GoldenGate analysis (Shen et al. 2005). In seven of the samples (included the positive control) the DNA concentration was increased to 50ng/µL or more by the amplification process. An example is the sample with a concentration of 0.5ng/µL prior to WGA which was amplified to a concentration of 61ng/µL (26 after purification). From this it is expected that the samples will achieve high call rate and GCS (section 2.3.2) from an Illumina GoldenGate analysis. However, the genotyping result did not reflect the concentration. The positive control for the reaction only received a call rate of 0.95 and the genotyping success for the rest of the samples is decreasing as the concentration decreases (figure 19). All WGA samples had a call rate and GCS that was lower than for the samples with corresponding (prior to WGA) concentration in the dilution series.

One reason for the poor results could be that the cluster used for the genotyping did not give a good fit for the WGA samples. It is recommended to form separate clusters based on only WGA samples. In this study that was not a possibility because there were not enough samples to form a cluster.

Xing et al (2008) amplified 10ng DNA and achieved successful genotyping and high concordance compared to genomic DNA when analyzed by the Affimetrix 250K array. There have been successful attempts at amplifying low template DNA of 100pg. However, at these levels there was detected some degree of bias like preferential amplification and allelic dropout (Ballantyne et al 2007). Lasken and Egholm (2003) report that for DNA quantities of less than 10ng the risk of drop out will increase. Thus the genome coverage may not be complete for amplified DNA from concentrations off less than 10ng. In this study we have seen that although the DNA has been amplified to reach the desired concentrations it seems that the whole genome has not been amplified. This conclusion is derived from the low call rates found in samples of adequate concentrations (figure 19). The results from this study imply that amplification of samples with a prior concentration of less than 10ng/µL will not be representative for the whole genome after the WGA procedure. Furthermore these samples have lower call rate and GCS than samples directly genotyped with the same (prior to amplification) concentration (Appendix B).

3.4 Statistical analysis

3.4.1 Determining the coefficient k

To examine if there was a difference in intensity between the two color channels the coefficient k was calculated as

$$k = \frac{Yraw}{Yraw + Xraw} \qquad (19)$$

for the two well performing reference samples H and C.

A plot was created to illustrate the distribution of k for both references and the plot for H is shown in figure 20.

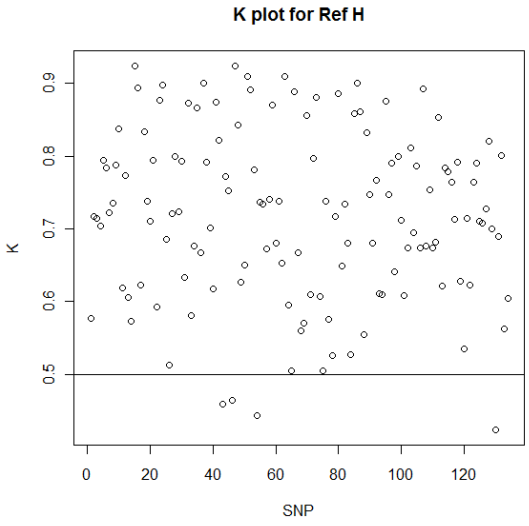


Figure 20. The plot illustrates the similarity of the fluorescence measurements from the two color channels in the Illumina GoldenGate analysis (k) calculated from the genotyping results for the reference sample H. If the signals from the two channels are equal K is expected to be 0.5.

It is visible from the plot that k is not equal to the expected value 0.5. The mean value of k was 0.72 with a 95% confidence interval of 0.70-0.74 for ref H, and for ref C 0.70 and 0.68-0.72 respectively. The confidence interval does not contain the value corresponding to Yraw=Xraw and therefore k differs significantly from 0.5.

3.4.2 Simulation

25 mixtures were simulated with the number of contributors ranging from one to five at different proportions (as in table 1). The two statistical approaches for mixture interpretation from SNP genotype raw data previously explained were used to analyze the simulated mixtures. The results from the statistical method 1 (Homer et al. 2008) are explained in the first section.The results from the statistical method 2 (section 2.5.2) are explained in the last section.

3.4.2.1 Analysis of simulated data by Statistical method 1

D(Y) and T(Y) (section 2.5.1) were calculated to identify the individuals present in the simulated mixtures. 75% of the individuals present in a mixture were identified. The probability of identifying an individual in the mixture seems to be dependent of the size of the contribution from that individual to the total mixture. The individuals that contributed less than 20% to the mixtures were not identified. An individual contributing 15% to a mixture could be identified if the other contributors were close in concentration. The relationship

between contribution to the mixture and the probability of being recognized as a part of that mixture is described in figure 21. The figure shows that all small contributions have a T value under the rejection limit, 2.

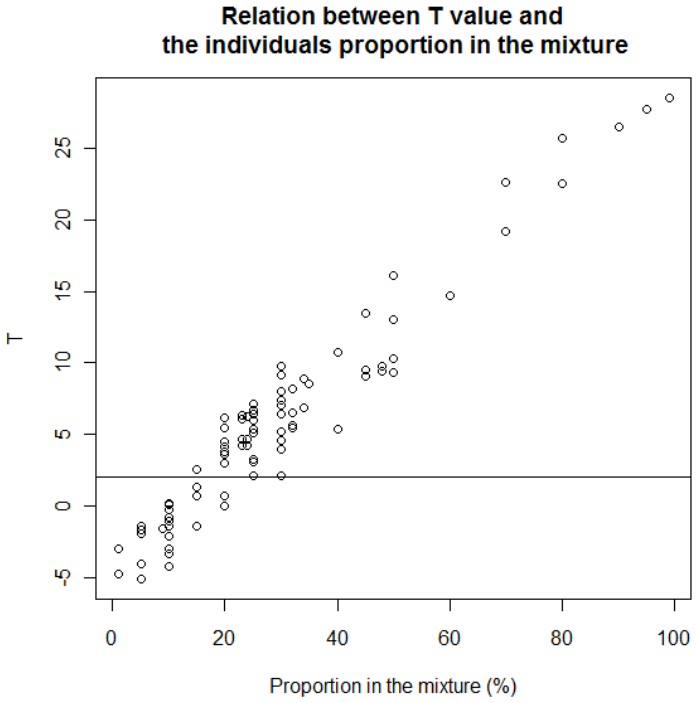


Figure 21. The T values calculated for each individual present in a mixture in the simulation analysis of the statistical method 1 (Homer et al. 2008). The plot represents the relationship between the T value calculated for the probability of an individual being present in the mixture and the proportion from that individual present in the mixture. It is visible that a person that has contributed less than 20% to the total mixture is not likely to be identified (has a T value less than 2, illustrated by the black line).

The simulated mixture of five individuals where individual 1 contributes 10% and individual 2-5 contribute 25% each, the individual contributing with a part of 0.1 was not found to be in the mixture. To test the effect of the number of SNPs in the analysis, the mixture was re analyzed with 6000, 7000, 8000, 9000, 10000, 11000 and 15000 SNPs. From table 3 we can see that individual 1 is not detected as a contributor since $|T| < 2$. The number of SNPs does not help significantly in regards to identifying the minor contributor. For practical reasons 500 000 SNPs have not been tried. It is on the other hand possible to identify 4 and 5 contributors if the contribution from each individual is the same. To be identified the variation between the sizes of the individual can be larger in a two or three person mixture (1:2.5) than in mixtures of four or five persons (1:2).

Table 3 T statistics calculated for the simulated mixture of five individuals. In the mixture individual 1 contributes 10% and individual 2-5 contributes 25%. The rows represent the number of SNPs in the simulation (6000- 15000) and the columns represent the calculated T value for each individual present in the mixture.

Number of SNPs	Individual 1	Individual 2	Individual 3	Individual 4	Individual 5
6000	-0.29	18.99	21.38	20.88	22.47
7000	-0.34	20.37	23.3	22.2	24.09
8000	-0.39	22.72	25.17	23.68	25.03
9000	-0.91	24.05	26.53	25.54	26.82
10000	-0.49	25.03	28.07	27.03	28.58
11000	-0.68	26.41	29.52	27.97	30.31
15000	-0.26	32.1	34.38	32.28	35.83

Homer et al. (2008) claims that they have recognized all contributors in complex mixtures in a simulation procedure. From these simulation studies they state that is possible to identify an individual contributing 0.1% of the total mixture by analyzing 10 000 to 50 000 SNPs. These results do not correspond to the findings in this assignment.

The paper by Homer et al. (2008) has attracted great international attention and is considered quite controversial. The reason for this is that the publishing lead to the removal of a large amount of genetic data from public databases, where the individual contributions were thought to be anonymous (http://grants.nih.gov/grants/gwas/data_sharing_policy_modifications_20080828.pdf). In addition criticism has been directed at the statistical methods (Clayton 2010, Braun et al. 2009) and alternative statistical approaches have been suggested (Clayton 2010, Visscher and Hill 2009, Jacobs et al 2009). There is reason to believe that the statistical approach might not function as well as described in the paper.

3.4.2.2 Analysis of simulated data by Statistical method 2

The regression analysis was performed to find the contribution β from each individual expected to be in the mixture. The null hypothesis H0: β=0 was tested for each β estimate and a p value vas calculated. In 94% of the cases the β corresponded to the true contribution from the individual of interest and H0 could be rejected. In total 8 individuals could not be identified, the individuals were all contributing a part of 10% or less to the mixtures. When increasing the number of SNPs in the simulated mixture to 5000, only the individuals

contributing 1% to the mixtures were unidentified. If more SNPs are analyzed it is expected to improve the results further, this has not been tested.

To find the possibilities for false positives, a random genotype was composed and the probability for a random person to be present in the mixtures was calculated. In 230 calculations for a random person being part of the mixture 17 incidences (7%) of false positives occurred. All the false contributions were of approximately 5% in the mixture and had a p value that was close to the rejection limit 0.05. To increase the number of SNPs or to decrease the rejection limit could account for this problem. There is also a problem with multiple testing and chance findings. This topic has not been discussed in this thesis.

3.4.3 Mixture analysis

In the genotyping cluster analyses the mixtures gave poor genotyping results. The average call rate was 0.58, the average GSC 0.38 and the average fluorescence signal 26000. This implies that DNA has hybridized to the beads on the genotyping chips but the results will not be appropriate for cluster analysis.

3.4.3.1 Determining the individuals present in the mixture – Statistical method 1

The allele frequency estimates for all mixtures were calculated from the raw data measurements in the file collected from the Illumina GoldenGate analysis, with respect to the B allele. The allele frequencies for the reference population were B allele frequencies from a HapMap data set. The population frequencies used in the calculations can be found in appendix D. The genotypes for the individual of interest were transformed to allele frequencies (0, 0.5 and 1 for AA, AB and BB respectively).

Twenty-five mixtures consisting of 2-5 individuals with different contributions were analyzed. The five individuals contributing to the mixtures were analyzed and genotyped. Since the true mixture compositions were unknown the presence of all five individuals was tested in all twenty -five mixtures.

In seven of the mixtures no contributors were identified, in sixteen mixtures one contributor was identified and in two mixtures two individuals were identified. Which reference sample that was identified in which mixture is given in table 4. The results do not use the information that that the mixtures are made from 2-5 contributors.

Table 4. The contributors identified to be present in the mixtures calculated from the statistical method 1 (Homer et al. 2008), the samples where no contributors were identified are not listed.

Sample name	Contributor s
Blind 1	H
Blind 4	D
Blind 8	F +D
Blind 9	F
Blind 10	F
Blind 11	F
Blind 12	F
Blind 13	F
Blind 14	H
Blind 16	D
Blind 17	F +D
Blind 18	H
Blind 19	H
Blind 20	B
Blind 21	D
Blind 23	H
Blind 24	B
Blind 25	D

As an attempt to improve the results the calculations were redone with k=3 (section 3.4.1). Some of the T values became larger but the individuals recognized in the mixture did not change very much.

3.4.3.2 Determining the individuals present in the mixtures – Statistical method 2

The regression analysis was performed for all reference samples in all mixtures. The null hypothesis; the reference is not part of the mixture was tested in all cases. For all cases where the null hypothesis could be rejected the proportion of the references found to be in the mixtures (β) are given in table 5. A full table of all β and corresponding p values is found in appendix E.

Table 5. The results from the mixture analysis calculated by the statistical method 2. For each mixture (rows) the proportion (β) found to be present from each individual (columns) is listed. If no β value is listed the individual was not identified in that mixture.

	β F	β D	β B	β H	β C
Blind 1	0.0793	0.2792	0.3642	0.3619	
Blind 2	0.0709	0.2344	0.3222	0.2917	0.2247
Blind 3	0.1007	0.4349	0.5370		
Blind 4	0.1451	0.8801			
Blind 5	0.1449	0.2203	0.2990	0.2567	0.2000
Blind 6	0.2947	0.3495	0.4164		
Blind 7	0.1997	0.2545	0.3402	0.3261	
Blind 8	0.4833	0.5275			
Blind 9	0.3872	0.2789	0.4016		
Blind 10	0.3929	0.1449	0.2531	0.3398	
Blind 11	0.6054	0.1608	0.0864	0.1467	0.1188
Blind 12	0.7463	0.2710			
Blind 13	0.7077	0.1335	0.2295		
Blind 14	0.2225	0.1346	0.3429	0.4284	
Blind 15	0.1576	0.2561	0.3975	0.2221	0.1269
Blind 16		0.9527			
Blind 17	0.3989	0.6103			
Blind 18	0.1051	0.2044	0.3288	0.3575	0.1121
Blind 19	0.0737	0.1946	0.3632	0.4532	
Blind 20	0.1727	0.3120	0.5835	0.0563	
Blind 21	0.3067	0.6888			
Blind 22		0.2348	0.3379	0.3001	0.2409
Blind 23		0.2921	0.3666	0.3964	
Blind 24		0.4489	0.5661		
Blind 25		0.8937			

3.4.3.3 Determining the individuals present in the mixtures - STR analysis

The samples Blind 1, Blind 2, Blind 5, Blind 7, Blind 10, Blind 11, Blind 14, Blind 15, Blind 18, Blind 19, Blind 22 and Blind 23 were by the maximum allele count method (described in section 1.2) found to be from four or more contributors. These mixtures were not analyzed, because although the alleles from each of the individuals could be recognized in all mixtures the probability that the alleles from a random person also be recognized is high. The epg from the sample Blind 2 is shown in figure 22. The maximum number of alleles presents in any loci is 8 which leads to the conclusion that four or more individuals must have contributed to the mixture.

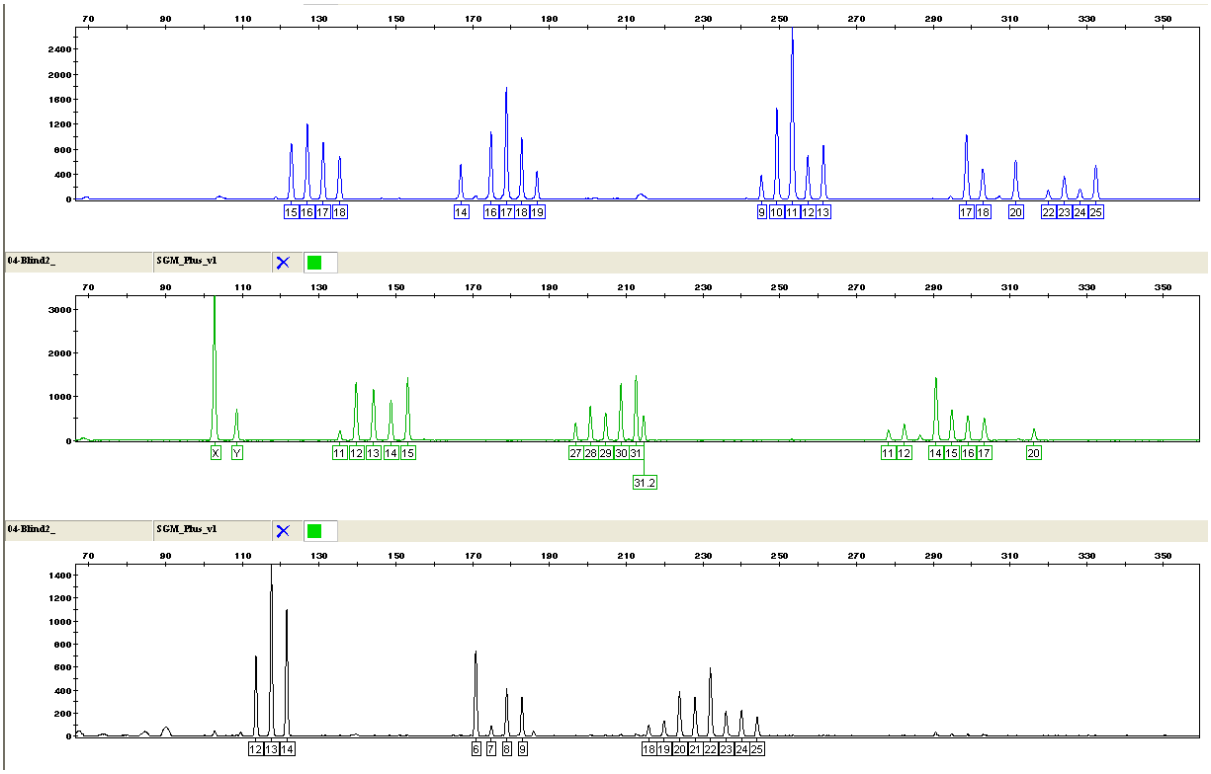


Figure 22 The epg shows all alleles present in the mixture Blind 2. To estimate the number of contributors in the mixture the maximum number of alleles at any loci must be counted (section 1.2). The maximum number of alleles present I n any loci is 8 which leads to the conclusion that four or more individuals must have contributed to the mixture.

The samples Blind 4, Blind 8, Blind 12, Blind 17, Blind 21, Blind 24 and Blind 25 were found to be mixtures of two persons. Blind 8 and blind 12 were analyzed in Genmapper ID-X prior to comparing with reference samples to find the most likely major and minor contributor. The most likely combinations in the samples Blind 8 and Blind 12 are listed in table 6 were the most likely allele combinations are listed in the top row. In Blind 8 the mixture proportion was close to 0.5, this makes it harder to distinguish the minor from the major contributors; hence more alternatives will be probable.

Table 6. The mixture analysis result for the STRs analyzed by the Genmapper ID-X software for the two person mixtures Blind 8 and Blind 12 are listed. The rows represent the probability of the genotype combinations where the most probable combinations are listed first.

Blind 8 major	D3S1358	vWA	D16S539	D2S1338	AMEL	D8S1179	D21S11	D18S51	D19S433	TH01	FGA
	15,15	17,18	9,11	17,17	X,X	12,13	27,28	14,14	13,14	6,7	22,23
		17,19	11,11		X,Y	11,13	28,30	11,14	12,13	7,8	19,22
		18,19				11,12		14,15	13,13	7,9.3	18,22
		18,18				12,12		11,15	12,14	6,8	19,23
Blind 8 minor	15,17	18,19	11,11	22,24	X,Y	11,12	28,30	11,15	12,13	8,9.3	18,19
		18,18	9,11		X,X	12,12	27,28	14,15	13,14	6,9.3	18,23
		17,18				12,13		11,14	12,14	6,8	19,23
		17,19				11,13		14,14	13,13	7,9.3	18,22
Blind 12 major	15,17	18,18	9,11	22,24	X,Y	11,12	27,28	11,15	13,14	7,9.3	18,19
			9,11	17,22		11,12			13,14		
Blind 12 minor	15,15	17,19	11,11	17,17	X,X	12,13	28,30	14,14	12,13	6,8	22,23
			11,F1	24,24		13,13			12,12		

The samples Blind 3, Blind 4, Blind 6 Blind 9, Blind 12, Blind 13 Blind 16, Blind 17, Blind 20, Blind 24, Blind 25 were typed to find the probable allele combinations and remove artifacts mistaken to be alleles (Clayton et al 1998). The samples were then compared to the genotypes of the references to find the expected composition of the mixture. All individuals found to be part of the mixture are listed in table 7.

Table 7. For all mixtures analyzed by STRs individuals found to be present in each mixture (rows) are listed. The samples where no individuals are listed were not analyzed because there were four or more contributors present.

Sample name	Individuals recognized in the mixture
Blind 1	-
Blind 2	-
Blind 3	D F B
Blind 4	D F
Blind 5	-
Blind 6	D F B
Blind 7	-
Blind 8	D F
Blind 9	D F B
Blind 10	-
Blind 11	-
Blind 12	D F
Blind 13	(D) F B
Blind 14	-
Blind 15	-
Blind 16	D
Blind 17	D F
Blind 18	-
Blind 19	-
Blind 20	D F B
Blind 21	D F
Blind 22	-
Blind 23	-
Blind 24	D B
Blind 25	D

3.4.4 Comparing the blinded samples with the true mixture components

When the mixtures had been interpreted by all methods the true mixture composition was revealed. The translation from blinded to mixture of known composition for all blinded samples are given in table 8.

Table 8. The mixture composite in the blinded samples with composition of the mixtures being described in table 1.

Blind	**Mixture**
Blind 1	MixtureG
Blind 2	MixtureH
Blind 3	MixtureF
Blind 4	MixtureE
Blind 5	MixtureD
Blind 6	MixtureB
Blind 7	MixtureC
Blind 8	MixtureA
Blind 9	MixtureR
Blind 10	MixtureS
Blind 11	MixtureT
Blind 12	MixtureU
Blind 13	MixtureV
Blind 14	MixtureW
Blind 15	MixtureX
Blind 16	MixtureY
Blind 17	MixtureQ
Blind 18	MixtureP
Blind 19	MixtureO
Blind 20	MixtureN
Blind 21	MixtureM
Blind 22	MixtureL
Blind 23	MixtureK
Blind 24	MixtureJ
Blind 25	MixtureI

3.4.4.1 Are the contributors in a mixture identified - Statistical method 1

In table 9 the individuals that were identified by the statistics from Homer et al (2008) are marked in yellow. The proportion of the individual in the mixture is listed in the table.

Table 9. The true mixture proportions from all individuals present in the mixture are listed in the table. The individuals identified by the mixtures by statistics from Homer et al. (2008) are marked in yellow.

Sample	Individual F	Individual D	Individual B	Individual H	Individual C
MixtureG	0.1	0.3	0.3	0.3	
MixtureH	0.1	0.225	0.225	0.225	0.225
MixtureF	0.1	0.45	0.45		
MixtureE	0.1	0.9			
MixtureD	0.2	0.2	0.2	0.2	0.2
MixtureB	0.33	0.33	0.33		
MixtureC	0.25	0.25	0.25	0.25	
MixtureA	0.5	0.5			
MixtureR	0.45	0.25	0.3		
MixtureS	0.5	0.1	0.15	0.25	
MixtureT	0.7	0.1	0.01	0.09	0.1
MixtureU	0.8	0.2			
MixtureV	0.8	0.05	0.15		
MixtureW	0.3	0.1	0.25	0.35	
MixtureX	0.2	0.25	0.3	0.15	0.1
MixtureY	0.01	0.99			
MixtureQ	0.4	0.6			
MixtureP	0.15	0.2	0.25	0.3	0.1
MixtureO	0.1	0.2	0.3	0.4	
MixtureN	0.2	0.3	0.5		
MixtureM	0.3	0.7			
MixtureL	0.05	0.238	0.238	0.238	0.238
MixtureK	0.05	0.317	0.317	0.317	
MixtureJ	0.05	0.475	0.475		
MixtureI	0.05	0.95			

The individuals that were identified in the mixtures were always the major (one of the majors if several) contributor. This corresponds with the observation from the simulation studies, that the mixture proportions could be influencing the results. There were seven mixtures where no contributors were identified. These mixtures were composed of three, four and five contributors. Moreover in the mixtures of two individuals at least one contributor was always identified. It seems that the number of contributors may have a greater impact on the result than what expected from the simulations. Although great attention has been directed towards the paper by Homer et al. (2008) the main focus has been on the statistical approach (Clayton 2010, Visscher and Hill 2009, Jacobs et al 2009) and not towards the mixture proportions, the sample concentration or the possibility to reduce the number of SNPs in the analysis. This is a relatively small experiment and the number of SNPs in the analysis (360) is not comparable to

the 500 000 analyzed by Homer et al. (2008). The choice of SNPs and the allele frequencies could also have an impact on the results. Furthermore there is a chance that the SNPs used in this analysis were not appropriate for this method. The noise from the Illumina GoldenGate data could in some way be inhibiting for the calculations. To improve the results the statistics for some samples were recalculated with k=3, which did not improve the number of individuals identified significantly (1 of 20 individuals were identified). It is possible that k was not properly calculated. There is need for further experience with the Illumina GoldenGate data to exclude such factors.

The selection of SNPs for the Illumina GoldenGate 360 SNP test panel includes SNPs on the X and Y chromosome. The SNPs located on the X and Y chromosomes may not be suitable for these analyses and a different result could have been obtained if they were removed. This has not been tested.

3.4.4.2 Are the contributors in a mixture identified - Statistical method 2

The β estimated in the regression analysis were compared to the true mixture proportions from each reference sample. The β and the true proportion are listed in table 10. In the mixtures Y, L, K, J and I the minor contributor in the mixture F is not recognized. In the mixture N H is a false positive. The individuals contributing 5% or less were not recognized except for in two cases individual B in mixture T (1%) and individual D in mixture V. The β for these samples were higher than the true contribution.

Table 10. The individual contributions (β) calculated from the statistical approach 2, and the true mixture proportions are listed for each reference sample (rows) in each mixture (columns). The samples which were a part of a mixture but not were identified are marked red, the false positive is marked green. There were two individuals contributing a part of less than 10% that were identified. These are marked in yellow.

Mixture	F	β F	D	β D	B	β B	H	β H	C	β C
MixtureG	0.100	0.079	0.300	0.279	0.300	0.364	0.300	0.362		
MixtureH	0.100	0.071	0.225	0.234	0.225	0.322	0.225	0.292	0.225	0.225
MixtureF	0.100	0.101	0.450	0.435	0.450	0.537				
MixtureE	0.100	0.145	0.900	0.880						
MixtureD	0.200	0.145	0.200	0.220	0.200	0.299	0.200	0.257	0.200	0.200
MixtureB	0.330	0.295	0.330	0.349	0.330	0.416				
MixtureC	0.250	0.200	0.250	0.254	0.250	0.340	0.250	0.326		
MixtureA	0.500	0.483	0.500	0.527						
MixtureR	0.450	0.387	0.250	0.279	0.300	0.402				
MixtureS	0.500	0.393	0.100	0.145	0.150	0.253	0.250	0.340		
MixtureT	0.700	0.605	0.100	0.161	0.010	0.086	0.090	0.147	0.100	0.119
MixtureU	0.800	0.746	0.200	0.271						
MixtureV	0.800	0.708	0.050	0.134	0.150	0.230				
MixtureW	0.300	0.223	0.100	0.135	0.250	0.343	0.350	0.428		
MixtureX	0.200	0.158	0.250	0.256	0.300	0.397	0.150	0.222	0.100	0.127
MixtureY	0.010		0.990	0.953						
MixtureQ	0.400	0.399	0.600	0.610						
MixtureP	0.150	0.105	0.200	0.204	0.250	0.329	0.300	0.358	0.100	0.112
MixtureO	0.100	0.074	0.200	0.195	0.300	0.363	0.400	0.453		
MixtureN	0.200	0.173	0.300	0.312	0.500	0.583		0.056		
MixtureM	0.300	0.307	0.700	0.689						
MixtureL	0.050		0.238	0.235	0.238	0.338	0.238	0.300	0.238	0.241
MixtureK	0.050		0.317	0.292	0.317	0.367	0.317	0.396		
MixtureJ	0.050		0.475	0.449	0.475	0.566				
MixtureI	0.050		0.950	0.894						

The result from the mixture analysis corresponds to the limitations found in the simulation study. The individuals contributing less than 10% to the mixtures were not identified except in mixture T where the individual B (1%) was identified and in mixture V where individual D (5%) was recognized. The β estimates for these samples were higher than the true contribution. It is not possible to exclude the possibility that an error occurred in the laboratory and the samples have a higher concentration than expected or that it could be a statistical error.

The regression based statistics works well for identifying individuals in a mixture and by analyzing only 360 SNPs all contributors that are 10% or less of the mixture can be recognized in complex mixtures. In addition the analysis will also provide the proportion of the individuals present in the mixture, which is not possible to find from the statistical method

1 (Homer et al. 2008). There is no need to know the number of contributors to perform these analyses.

Theoretically β can be estimated as negative, and this does not make sense. However this is not a practical problem. An alternative model which secures positive β estimates would be too complex.

There is also dependence between the two signals from a heterozygote marker. This dependence is not accounted for in our model. There may be some bias in the estimate.

While 360 SNPs does not seem like a suitable number of markers for the statistical method 1, the statistical method 2 performs well with regards to identification. In section 3.4.2.2 it is shown that more SNPs are expected to improve the regression based approach. However the number of SNPs necessary for improvement is not expected to be as high as for the statistical method 1 (see section 3.4.2.2). A method for mixture interpretation should be reduced to as few SNPs as possible, because more SNPs lead to the need for higher DNA concentrations in the analyzed sample. The low DNA concentrations might be one of the limiting factors for these methods to become useful in a forensic context.

There are mathematical and economical considerations that need to be taken into account before deciding on using SNPs for forensic purposes. These are not discussed here.

3.4.4.3 STR analysis – comparing the statistical approaches

The correct major and minor genotypes in the mixtures blind 8 and blind 12 are marked in table 13.

Table 18. The mixture analysis results for the STRs analyzed by the Genmapper ID-X software for the two person mixtures Blind 8 and Blind 12 are listed. The true genotype combinations found from the reference samples are marked in yellow.

Blind 8 major	D3S1358	vWA	D16S539	D2S1338	AMEL	D8S1179	D21S11	D18S51	D19S433	TH01	FGA
	15,15	17,18	9,11	17,17	X,X	12,13	27,28	14,14	13,14	6,7	22,23
		17,19	11,11		X,Y	11,13	28,30	11,14	12,13	7,8	19,22
		18,19				11,12		14,15	13,13	7,9.3	18,22
		18,18				12,12		11,15	12,14	6,8	19,23
Blind 8 minor	15,17	18,19	11,11	22,24	X,Y	11,12	28,30	11,15	12,13	8,9.3	18,19
		18,18	9,11		X,X	12,12	27,28	14,15	13,14	6,9.3	18,23
		17,18				12,13		11,14	12,14	6,8	19,23
		17,19				11,13		14,14	13,13	7,9.3	18,22
Blind 12 major	15,17	18,18	9,11	22,24	X,Y	11,12	27,28	11,15	13,14	7,9.3	18,19
			9,11	17,22		11,12			13,14		
Blind 12 minor	15,15	17,19	11,11	17,17	X,X	12,13	28,30	14,14	12,13	6,8	22,23
			11,F1	24,24		13,13			12,12		

For the mixture Blind 12 the genotype combination listed as most likely in the genemapper ID X software (Applied Biosystems) was the same combination as the one found in the reference samples. In the mixture Blind 8 the correct combination was found in the first, second third and fourth alternative. In this mixture the contributors were close in regards to concentration and it seems that the most probable alternative will be more difficult to estimate for such mixtures.

In Blind 13 individual D is not observed in at least two alleles, in Blind 16, 24 and 25 the alleles from the minor contributor F is not observable. In these samples Fs contribution is 1, 5 and 5% respectively. For all other analyzed samples the contributors were correctly identified.

For mixtures consisting of two and three individuals the success for the mixture interpretation results for the STRs and the results for the statistical method 2 in the previous section are corresponding.

Budowle et al. (2009) argues that identifying an individual in a mixture is dependent of the presence of clear minor and major contributors to separate the individual genotypes. Due to shared alleles and masking this is not likely to be the case for complex mixtures. They also

propose that depending on the complexity of the mixture they may only be informative for exclusion purposes. The samples with four or more contributors were not analyzed because it was not possible to distinguish the individual contributors. To test this statement the genotypes of two random individuals not present in any mixtures were compared to the alleles present in the mixture Blind 5. The alleles from the random individuals were recognized in 19 and 20 of 22 possible alleles respectively.

Although the STR mixture interpretation is comparable to the statistical method 2 for two or three person mixtures, the new method provides a clear improvement when it comes to determining the presence of individuals in mixtures composed of four or more individuals. Budowle et al. (2009) and Gill et al (2006) address and propose international guidance for mixture interpretation to eliminate the possibility that the interpretation results are dependent on the person interpreting the mixtures. The statistical method 2 provides a mixture interpretation method that can be automated and has the possibility to exclude such factors.

In the STR mixture interpretation it was observed that the mixtures which are created from blood samples, are of very good quality. Mixtures like these where all alleles for all contributors are found to be present are seldom seen in the forensic samples with more than two contributors. So although the statistical method 2 provides good mixture interpretation results, these results cannot be compared to real mixtures. Further testing on real samples will be necessary to come to a conclusion in regards to the usability for these analyses in a forensic context.

There are some unique challenges to analyzing forensic samples when it comes to the low DNA concentration and degraded quality. There may also be the stricter quality requirements for forensic applications because genotyping errors can lead to a wrong conclusion in regards to contenting a criminal to a crime scene by the findings of a DNA profile. In a normal genotyping scenario a failed genotype for a SNP can often be discarded without problems. In a forensic context discarding data is much more problematic. However, the results from the mixture interpretation by the statistical method 2 compared to the STR analysis tell us that there is possibility for improvement when it comes to identifying individuals in complex mixture. Thus high-density genome-wide SNP arrays can be a suitable tool for forensic analyses.

6 REFERENCES

Asari M, Watanabe S, Matsubara K, Shiono H and Shimizu K. Single nucleotide polymorphism genotyping by mini-primer allele-specific amplification with universal reporter primers for identification of degraded DNA. Analytical Biochemistry 2009;**386**:85-90

Ballantyne KN, van Oorschot RAH. and Mitchell RJ. Increasing amplification success of forensic DNA samples using multiple displacement amplification. Forensic Sci Med Pathol 2007; **3**:182-187.

Bender K, Farfán MJ and Schneider PM. Preparation of degraded human DNA under controlled conditions. Forensic Science International 2004;**139:**135-140

Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, Buckleton J. PENDULUM—a guideline-based approach to the interpretation of STR mixtures
Forensic Science International 2005;**148**:181-189
Buckleton J, Triggs C, Walsh S. 2005 Forensic DNA evidence interpretation.CRC PRESS

Butler JM. Forensic DNA typing, biology, technology, and genetics of STR markers. Second edition. ELSEVIER 2005.

Butler JM, Coble MD, Vallone MP. STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Science, Medicine, and Pathology 2007;**3**:200-205

Budowle B and van Daal A. Extracting evidence from forensic DNA analyses: future molecular biology directions. BioTechniques 2009;**46**:339-350

Braun R, Rowe W, Schaefer C, Zhang J and Buetow K. Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet 2009;**5**(10):e1000668. doi:10.1371/journal.pgen.1000668

Børsting C, Rockenbauer E, Morling N. Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. Forensic Science International: Genetics. December 2009;**4**:34-42

Coskun S and Asmaldi O. Whole genome amplification from a single cell: a new era for preimplantation genetic diagnosis. Prenat Diagn 2007;**27**:297-302.

Clayton D. On inferring presence of an individual in a mixture: a Bayesian approach. Biostatistics (To appear 2010)

Clayton TM, Whitaker JP, Sparkes R, Gill P. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Science International. 1998;**91**:55-70.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, and Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. PNAS . 2002;**8**:5261–5266

Dixon LA, Dobbins AE, Pulker HK, Butler JM, Vallone PM, Coble MD, Parson W, Berger B, Grubweiser P, Mogensen HS, Morling N, Nielsen K, Sanchez JJ, Petkovski E, Carracedo A, Sanchez-Diz P, Ramos-Luis E, Brion M, Irwin JA, Just RS, Loreille O, Parsons TJ, Syndercombe-Court D, Schmitter H, Stradmann-Bellinghausen B, Bender K and Gill P. Analysis of artificially degraded DNA using STRs and SNPs- results of a collaborative European (EDAP) exercise. Forensic Science International 2006;**164**:33-44

Gill P. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes Int J Legal Med 2001;**114**:204-210

Gill P, Fereday L, Morling N, Schneider PM. The evolution of DNA databases-Recommendations for new European STR loci. Forensic Science International 2006;**156**:242-244

Gill, P, Brenner, CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz, M, Schneider PM, Weir BS. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International. 2006;**160**:90-101

Gunderson KL, Kruglyak S, Graige MS, et al. Decoding randomly ordered DNA arrays. Genome research 2004;**14**:870-877

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J, Stephan D A, Nelson SF and Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;**4**(8): e1000167. doi:10.1371/journal.pgen.1000167

Illumina 2006. GoldenGate Assay Workfolw. (http://www.illumina.com/documents/products/workflows/workflow_goldengate_assay.pdf)

Illumina 2005. Technical Note Illumina GenCall Data Analysis Software (http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf)

Jacobs K B, Yeager M, Wacholder S, Craig D, Kraft P, Hunter JD, Paschal J, Manolio TA, Tucker M, Hoover RN, Thomas GD, Chanock SJ and Chatterjee N. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nature genetics 2009;**11**(41):1253-1259

Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SLB, Karoma NJ, Kungulilo S, Kim J, Lu R,Odunsi A,Okonofua, F, Parnas J, Schulz LO, Zhukova OV, Kidd JR. Developing a SNP panel for forensic identification of individuals. Forensic Science International. 2006;**164**:20-32.

Krjutskov K, Viltrop T, Palta P, Metspalu E, Tamm E, Suvi S, Sak K, Merilo A, Sork H, Teek R, Nikopensius T, Kivisild T, Metspalu A. Evaluation of the 124-plex SNP typing microarray for forensic testing. Forensic Science International: Genetics 2009. DOI: 10.1016/j.fsigen.2009.04.007.

Lasken RS and Egholm M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. TRENDS in biotechnology 2003;**12**(21):531-535

Pusch W, Wurmbach JH, Thiele H, Kostrzewa M. MALDI-TOF mass spectrometry-based SNP genotyping. Pharmacogenomics. 2002;**4**:537-548

Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006;**27**:1713 – 1724

Sequenom 2009. iPLEX Gold application guide (http://www.sequenom.com/Files/Genetic-Analysis---Graphics/iPLEX-Application-PDFs/iPLEX-Gold-Application-Guide-v2r1)

Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A. High-throughput SNP genotyping on universal bead arrays. Mutation Research 2005;**573**:70-82

Sobrino B, Brion M, Carracedo A. SNPS in forensic genetics: a review on SNP typing methodologies. Forensic Science International 2005;**154**:181-194.

Visscher PM. and Hill GW. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet 2009 5(10): e1000628. doi:10.1371/journal.pgen.1000628

Xing J, Watkins WS, Zhang Y, Witherspoon DJ and Jorde LB. High fidelity of whole-genome amplified DNA on high-density single nucleotide polymorphism arrays. Genomics 2008;**92**:452-456

Appendix A- Samples analyzed by the iPLEX assay

All samples analyzed in the iPLEX assay are listed in the table. The table gives the assignment of the samples (dilution, degraded, WGA), the sample name, the measured concentration and the genotyping results measured in percent SNPs that was assigned a genotype from the analysis.

Assignment	Sample Name	Concentration (ng/µL)	Result 29-plex (%)	Result 10-plex (%)
Dilution	1b1	28.80	88.00	100.00
Dilution	1b2	13.90	77.00	100.00
Dilution	1b3	6.50	85.00	100.00
Dilution	1b4	3.50	85.00	100.00
Dilution	1b5	1.60	88.00	100.00
Dilution	1b6	0.80	92.00	100.00
Dilution	1c1	28.20	92.00	100.00
Dilution	1c2	13.70	92.00	100.00
Dilution	1c3	5.90	96.00	100.00
Dilution	1c4	2.90	96.00	100.00
Dilution	1c5	1.40	100.00	100.00
Dilution	1c6	0.70	96.00	100.00
Dilution	1f1	26.90	96.00	100.00
Dilution	1f2	13.30	96.00	100.00
Dilution	1f3	6.60	96.00	100.00
Dilution	1f4	3.30	96.00	100.00
Dilution	1f5	1.60	96.00	100.00
Dilution	1f6	0.70	100.00	100.00
Dilution	2c1	4.90	100.00	96.00
Dilution	2c2	1.56	100.00	96.00
Dilution	2c3	0.52	100.00	100.00
Dilution	2c4	0.18	100.00	84.00
Dilution	2c5	0.05	100.00	92.00
Dilution	2c6	0.02	75.00	64.00
Dilution	2c7	0.01	37.50	64.00
Dilution	2f1	4.02	100.00	96.00
Dilution	2f2	1.55	100.00	96.00
Dilution	2f3	0.42	100.00	96.00
Dilution	2f4	0.12	100.00	92.00
Dilution	2f5	0.04	87.50	92.00
Dilution	2f6	0.01	75.00	72.00
Dilution	2f7	0.00	50.00	92.00
Dilution	2g1	5.00	100.00	96.00
Dilution	2g2	1.35	100.00	96.00

Dilution	2g3	0.46	100.00	72.00
Dilution	2g4	0.15	100.00	100.00
Dilution	2g5	0.04	100.00	84.00
Dilution	2g6	0.02	75.00	80.00
Dilution	2g7	0.01	62.50	28.00
UV Degraded/control	D1	26.90	96.00	100.00
UV	D2	13.30	96.00	100.00
UV	D3	6.60	96.00	100.00
UV	D4	3.30	96.00	100.00
UV	D5	1.60	96.00	100.00
UV	D6	0.70	100.00	100.00
UV Degraded	d1t1	70.00	30.77	50.00
UV Degraded	d1t2	15.00	30.77	37.50
UV Degraded	d1t3	7.00	15.38	75.00
UV Degraded	d1t4	3.00	15.38	25.00
UV Degraded	d1t5	1.50	23.08	12.50
UV Degraded	d1t6	0.50	11.54	12.50
UV Degraded	d1_5t1	75.00	26.92	37.50
UV Degraded	d1_5t2	15.00	19.23	12.50
UV Degraded	d1_5t3	7.00	7.69	25.00
UV Degraded	d1_5t4	3.00	15.38	0.00
UV Degraded	d1_5t5	1.50	3.85	12.50
UV Degraded	d1_5t6	0.50	23.08	0.00
UV Degraded	d2t1	87.00	34.62	25.00
UV Degraded	d2t2	15.00	11.54	62.50
UV Degraded	d2t3	7.00	23.08	37.50
UV Degraded	d2t4	3.00	11.54	0.00
UV Degraded	d2t5	1.50	7.69	0.00
UV Degraded	d2t6	0.50	3.85	12.50
DNAse Degraded	d00	21.90	0.00	50.00
DNAse Degraded	d01	13.50	7.69	50.00
DNAse Degraded	d02	6.50	3.85	80.00
DNAse Degraded	d03	3.00	34.62	80.00
DNAse Degraded	d04	1.50	88.46	90.00
DNAse Degraded	d05	0.50	80.77	100.00
DNAse Degraded	d50	19.40	0.00	10.00
DNAse Degraded	d51	13.50	23.08	30.00
DNAse Degraded	d52	6.50	50.00	80.00
DNAse Degraded	d53	3.00	69.23	90.00
DNAse Degraded	d54	1.50	76.92	90.00
DNAse Degraded	d55	0.50	80.77	70.00
DNAse Degraded	d150	16.00	7.69	50.00

DNAse Degraded	d151	13.50	3.85	60.00
DNAse Degraded	d152	6.50	46.15	50.00
DNAse Degraded	d153	3.00	61.54	90.00
DNAse Degraded	d154	1.50	73.08	90.00
DNAse Degraded	d155	0.50	53.85	40.00
WGA	2c1W	282.00	0.00	100.00
WGA	2f1W	289.00	22.73	80.00
WGA	2c6W	13.20	13.64	20.00
WGA	2f6W	12.50	4.55	60.00
WGA	2c7W	10.60	4.55	10.00
WGA	2f7W	3.06	4.55	10.00
WGA purified	2c1_1R	48.70	90.91	100.00
WGA purified	2c6_1R	1.22	95.45	100.00
WGA purified	2f1_1R	47.50	13.64	11.11
WGA purified	2f6_1R	1.05	31.82	33.33
WGA Diluted	2c1D	5.00	81.82	80.00
WGA Diluted	2f1D	5.00	95.45	90.00
WGA Diluted	2c6D	5.00	22.73	30.00
WGA Diluted	2f6D	5.00	27.27	50.00
WGA Diluted	2c7D	5.00	36.36	10.00

Appendix B – Samples analyzed by the Illumina GoldenGate assay

All samples analyzed in the Illumina GoldenGate assay are listed in the table. The table lists the assignment of the samples (dilution, degraded, WGA, Blind = mixture), the measured concentration and the genotyping results as call rate for the sample, mean GCS for all SNPs analyzed in the sample and the mean intensity measured for the sample.

Sample name	Concentration (ng/µL)	Call rate	Mean GCS	Mean intensity
Negative control-1	0.00	0.17	0.11	6453.00
Negative control-2	0.00	0.34	0.21	12516.00
Negative control-3	0.00	0.23	0.15	8314.00
Positive control C	50.00	0.97	0.72	22914.00
Reference D	50.00	0.93	0.63	17970.00
Reference B	50.00	0.59	0.37	20337.00
Reference F	50.00	0.96	0.67	19422.00
Reference H	50.00	0.99	0.74	24462.00
Dilution1	20.00	0.99	0.74	23541.00
Dilution1-2	20.00	0.98	0.71	21118.00
Diluton2	15.00	0.99	0.74	26238.00
Dilution2-2	15.00	0.97	0.69	23453.00
Dilution3	10.00	0.97	0.68	18866.00
Dilution3-2	10.00	0.95	0.66	20707.00
Dilution4	4.50	0.99	0.73	27804.00
Dilution4-2	4.50	0.92	0.63	17223.00
Dilution5	2.00	0.90	0.61	19177.00
Dilution5-2	2.00	0.90	0.63	19386.00
Dilution6	1.00	0.84	0.57	21129.00
Dilution6-2	1.00	0.54	0.35	23580.00
Negative control Degraded	0.00	0.33	0.22	7941.00
Degraded 1	48.00	0.28	0.20	9795.00
Degraded 1-2	48.00	0.00	0.00	0.00
Degraded 2	9.00	0.17	0.12	6985.00
Degraded 2-2	9.00	0.25	0.17	6433.00
Degraded 3	5.00	0.00	0.00	0.00
Degraded 3-2	5.00	0.26	0.15	253.00
Degraded 4	1.50	0.12	0.11	1820.00
Degraded 4-2	1.50	0.52	0.34	18189.00
Degraded 5	0.50	0.33	0.22	164.00
Degraded 5-2	0.50	0.00	0.00	0.00
Negative control WGA	0.00	0.00	0.00	0.00
Positive control WGA C	14.00	0.95	0.67	21045.00
WGA 1	6.00	0.77	0.49	19204.00
WGA 1-2	6.00	0.84	0.55	19266.00
WGA 2	3.00	0.69	0.44	17572.00

WGA 2-2	3.00	0.78	0.51	18273.00
WGA 3	1.60	0.48	0.31	15838.00
WGA 3-2	1.60	0.50	0.34	19248.00
WGA 4	0.70	0.71	0.46	18369.00
WGA 4-2	0.70	0.74	0.49	18557.00
WGA 5	0.50	0.51	0.35	18251.00
WGA 5-2	0.50	0.51	0.34	15001.00
WGA 6	0.20	0.28	0.21	13735.00
WGA 6-2	0.20	0.22	0.15	2302.00
WGA 7	0.05	0.00	0.00	0.00
WGA 7-2	0.05	0.00	0.00	0.00
Blind 1	50.00	0.50	0,312	22179.00
Blind 1_2	50.00	0.54	0,349	26218.00
Blind 2	50.00	0.55	0,357	23803.00
Blind 2_2	50.00	0.55	0,365	22078.00
Blind 3	50.00	0.45	0,288	29087.00
Blind 3_2	50.00	0.52	0,332	25605.00
Blind 4	50.00	0.88	0,621	24785.00
Blind 4_2	50.00	0.83	0,546	19754.00
Blind 5	50.00	0.53	0,338	22642.00
Blind 5_2	50.00	0.53	0,337	19258.00
Blind 6	50.00	0.41	0,251	23102.00
Blind 6_2	50.00	0.51	0,318	27287.00
Blind 7	50.00	0.53	0,336	28658.00
Blind 7_2	50.00	0.55	0,349	28659.00
Blind 8	50.00	0.56	0,381	31663.00
Blind 8_2	50.00	0.57	0,353	23635.00
Blind 9	50.00	0.54	0,354	26225.00
Blind 9_2	50.00	0.53	0,333	25926.00
Blind 10	50.00	0.57	0,366	28547.00
Blind 10_2	50.00	0.57	0,363	23876.00
Blind 11	50.00	0.62	0,420	25938.00
Blind 11_2	50.00	0.64	0,429	26237.00
Blind 12	50.00	0.68	0,463	26327.00
Blind 12_2	50.00	0.66	0,447	26228.00
Blind 13	50.00	0.62	0,401	22155.00
Blind 13_2	50.00	0.69	0,471	27769.00
Blind 14	50.00	0.52	0,321	27971.00
Blind 14_2	50.00	0.49	0,313	25102.00
Blind 15	50.00	0.47	0,295	28340.00
Blind 15_2	50.00	0.54	0,339	29789.00
Blind 16	50.00	0.95	0,679	27611.00
Blind 16_2	50.00	0.90	0,634	33826.00
Blind 17	50.00	0.59	0,385	23375.00
Blind 17_2	50.00	0.59	0,393	26552.00

Blind 18	50.00	0.49	0,328	24357.00
Blind 18_2	50.00	0.50	0,318	22945.00
Blind 19	50.00	0.53	0,332	22301.00
Blind 19_2	50.00	0.57	0,370	26482.00
Blind 20	50.00	0.50	0,315	26665.00
Blind 20_2	50.00	0.53	0,330	25330.00
Blind 21	50.00	0.64	0,417	20731.00
Blind 21_2	50.00	0.63	0,408	27817.00
Blind 22	50.00	0.56	0,351	26081.00
Blind 22_2	50.00	0.54	0,347	25863.00
Blind 23	50.00	0.56	0,359	26557.00
Blind 23_2	50.00	0.47	0,283	28735.00
Blind 24	50.00	0.42	0,261	23597.00
Blind 24_2	50.00	0.47	0,303	32418.00
Blind 25	50.00	0.80	0,534	25314.00

Appendix C - SNPs and primer sequences in the iPLEX assay

The SNPs analyzed in the iPLEX assay and their corresponding primer sequences are found in the table

SNP_ID	2nd-PCRP	1st-PCRP
AQP4_2588_C/T	ACGTTGGATGGAGTTGGAATCTAACTGCCC	ACGTTGGATGTGAATTGCGCCCTTTTAAAC
rs1130183	ACGTTGGATGATCAGCACCAGCTCAAAGTC	ACGTTGGATGCATGTGGTAGATGAGACCAG
rs12133079	ACGTTGGATGAGCTGTTATCTGCTGCTCCC	ACGTTGGATGTTATCCGCTTCTGACTCTGG
rs11265313	ACGTTGGATGAGACCCCAGTAAGAGCATTG	ACGTTGGATGGGCCCAACAAGAAGAGTTAG
rs1839318	ACGTTGGATGTGGAAATCTTACCGCTGGTC	ACGTTGGATGTGGAATCACAGCTGGCAAAG
rs74163676	ACGTTGGATGCTAGCTTCCTTAGCCTACTG	ACGTTGGATGATGCGATCATAGGTGCTGTC
rs1130182	ACGTTGGATGCCACCATCCTGGAAATCTTC	ACGTTGGATGCTGAAACGAATGGTCTCAGC
rs1186675	ACGTTGGATGAGAGTGAGAAGGGCAGATTG	ACGTTGGATGCCTGGGTTAGTTCTGATGAG
rs4656873	ACGTTGGATGACGTTAACCATGGACACAGC	ACGTTGGATGCCTAACTGCTTCCATGGTTG
rs12968026	ACGTTGGATGGGAACATTCAGTGACATGGG	ACGTTGGATGGGTTAGGCAAGAAAGCCAAG
rs1186679	ACGTTGGATGCCACAAAAAAATTCTGGCGG	ACGTTGGATGCCATCCAATGCTACACATAG
rs72557972	ACGTTGGATGTCTCTACCTGACTCCTGTTG	ACGTTGGATGAAAGAAGCCTTCAGCAAAGC
rs61327137	ACGTTGGATGCAGAGAAGAGATCAGAGAGG	ACGTTGGATGGGAAGAGACAGTTGGCATTG
rs11661256	ACGTTGGATGGTGTGGCTGGAAGAATCAAG	ACGTTGGATGTAGAGGAGGGCTCTCATTTC
AQP4_17906-A/G	ACGTTGGATGGCAAATTCTATAGTGCTTATG	ACGTTGGATGGGCAACTGAAGATGGAAGTC
rs2820553	ACGTTGGATGTGTTGTTTCTCCTCCACCAC	ACGTTGGATGAGAGTCTAGGGCAGTGTTAG
AQP4_7751_C/T	ACGTTGGATGGGGAGATTTTCAGAATGCC	ACGTTGGATGGCATTAAAACAAGGTGTGCG
rs17375748	ACGTTGGATGGAAGGGAGGGAAAATGCAAG	ACGTTGGATGGAATCTGTGCCCTGTGAATC
rs35248760	ACGTTGGATGCTGGGATCCACCATCAACTG	ACGTTGGATGGTCCAAAGCAAAGGGAGATG
rs151244	ACGTTGGATGGAGTGGTGGGTTCATAGATG	ACGTTGGATGCATCACTCGTGCATGTATGG
rs74163677	ACGTTGGATGCAGCACCAGGGCTGATTAAA	ACGTTGGATGGAATTTCTTGAGAGCCTTGTG
rs74163681	ACGTTGGATGTGTGGGATTTGTAGCCTGTG	ACGTTGGATGAGAGCCCTCCTCTAAATGTC
rs17853258	ACGTTGGATGGTGCCACAGCTACCAGATAC	ACGTTGGATGCAAGCTTCTGCTCTTCTCTG
AQP4_19867_G/T	ACGTTGGATGGTCCAAAACAATGTCACCTG	ACGTTGGATGATTGAAGAACTCAACTCAGC
rs9961118	ACGTTGGATGCAGTCTTTCTTCTCTCATGT	ACGTTGGATGCTCCCTTTTCCCACTTTATA
rs12122979	ACGTTGGATGCAGTTCCAACTTGAATAGCG	ACGTTGGATGAACACTTGGAAGCCCAGAAG
rs2486253	ACGTTGGATGACTGTAGATGGACACCGAAG	ACGTTGGATGATTGATCTCTCCTCGAACTC
rs3763043	ACGTTGGATGCACGTCTATCAGCTTATTCC	ACGTTGGATGTGCATGACTGTGACATACTG
rs162009	ACGTTGGATGCCATGTCACTGAATGTTCCC	ACGTTGGATGATTCCCCTCACCCTTTTTGG
rs162008	ACGTTGGATGCCAGAGTGCAGCTCTCATTG	ACGTTGGATGAACCCAATCAGACAAGTGGC
rs74163685	ACGTTGGATGATCCATCCTCAGGCCATTTG	ACGTTGGATGAACCCAGGGAGTTAAACCAG
rs7512587	ACGTTGGATGACTGAGCACCTCATGAGAAG	ACGTTGGATGAGCAGTTCTGCAAAGTGTCC
rs1130181	ACGTTGGATGCGTGAGAATGGAGCACATTG	ACGTTGGATGCCACTGCATGTCAATGAAGG
rs1890532	ACGTTGGATGGGTTTAGACACGTGGAGAAG	ACGTTGGATGTCCACTTCTGATCCCAGTTC
rs3875089	ACGTTGGATGGGCTTTTGCAGATCTGAAAC	ACGTTGGATGGAAGAGAGGCATAGAGAAGG
rs1058427	ACGTTGGATGTTGTACCTTGTCTGTCATGC	ACGTTGGATGAAAGGCCCTGTCCCAATCTC
rs2339214	ACGTTGGATGGTCTAAGTGTAGCCTGTACT	ACGTTGGATGGTGGTCACCAAAATGTTGGC
rs72878794	ACGTTGGATGTATGAACCCACCACTCAACC	ACGTTGGATGTATAAAGTGGGAAAAGGGAG
rs74163678	ACGTTGGATGTCTCACCCCCCAAAAAGAAG	ACGTTGGATGCACAAACTCCTGTTTGGTCG

Appendix D- SNPs in the Illumina GoldenGate assay

The SNPs analyzed in the Illumina GoldenGate assay their chromosome location and the allele frequencies for each SNP are listed in the table.

			Caucasian (CEU)	
Locus_Name	Build 35 Chromosome	Build 35 Coordinate	Allele A freq	Allele B freq
rs1000821	17	71945598	0.49	0.51
rs1005488	X	131116515	0.33	0.67
rs1007321	X	21727592	0.74	0.26
rs1010172	13	108644009	0.65	0.35
rs1011526	X	65199108	0.24	0.76
rs1013087	8	56346944	0.47	0.53
rs1013758	3	43601379	0.53	0.47
rs1015117	2	86652521	0.46	0.54
rs1016461	6	69092970	0.48	0.52
rs1017507	4	135500593	0.42	0.58
rs1019837	2	83326738	0.66	0.34
rs1019977	18	17257904	0.16	0.84
rs1020382	19	218039	0.59	0.41
rs1021393	15	96209001	0.49	0.51
rs1021516	5	116572071	0.43	0.57
rs1022239	13	64322047	0.44	0.56
rs1022573	6	127096607	0.58	0.42
rs1024516	7	85925849	0.69	0.31
rs1024694	X	146883956	0.49	0.51
rs1027702	1	158444515	0.33	0.67
rs1030588	15	44513473	0.54	0.46
rs1037958	15	55310834	0.42	0.58
rs1039524	3	115494964	0.58	0.42
rs1043415	20	35378652	0.40	0.60
rs1050755	10	112043589	0.52	0.48
rs1057613	4	100862163	0.46	0.54
rs1075840	15	89602911	0.61	0.39
rs1075870	3	196162579	0.48	0.52
rs1080085	6	161682446	0.70	0.30
rs1108081	15	35893455	0.47	0.53
rs1139266	11	45789511	0.67	0.33
rs11457	15	61673432	0.65	0.35
rs1147696	3	121602169	0.45	0.55
rs1152324	X	106099634	0.18	0.82
rs1157023	4	64318692	0.23	0.78
rs1163016	12	79554821	0.69	0.31
rs11664524	18	7222892	0.82	0.18
rs1179992	12	119958152	0.34	0.66
rs11813505	10	24661886	0.59	0.41
rs1190742	X	135801250	0.48	0.52
rs1206147	6	97671894	0.48	0.53
rs1229133	1	118706429	0.64	0.36
rs1260658	6	109526918	0.65	0.35

rs12624577	20	539820	0.62	0.37
rs1264216	X	65218278	0.76	0.24
rs1266490	15	89258224	0.34	0.66
rs1268722	10	50621216	0.47	0.53
rs131020	22	47494320	0.47	0.53
rs1320131	2	241131279	0.22	0.78
rs1328273	9	16013469	0.44	0.56
rs1330225	1	106547985	0.67	0.33
rs1338248	6	93639259	0.47	0.53
rs1339737	1	235717811	0.34	0.66
rs1351631	3	43493171	0.45	0.55
rs1352695	4	158919874	0.47	0.52
rs1363157	5	163142736	0.48	0.52
rs136488	22	30918910	0.47	0.53
rs1366199	5	115349647	0.84	0.16
rs1372177	13	78469870	0.47	0.53
rs1374197	3	17369619	0.47	0.53
rs1375062	8	142034963	0.63	0.37
rs1383895	8	31823084	0.53	0.47
rs1383972	4	86741508	0.64	0.36
rs1392702	3	56809019	0.39	0.61
rs1396226	12	73586112	0.63	0.37
rs1402810	2	139217540	0.25	0.75
rs1408209	13	92805575	0.18	0.82
rs1409778	1	196044182	0.47	0.53
rs14132	12	19565448	0.41	0.59
rs1426311	18	27152565	0.49	0.51
rs1433251	12	71362298	0.78	0.22
rs1433451	15	85408023	0.34	0.66
rs1435850	2	229256426	0.25	0.75
rs1440369	8	73728570	0.39	0.61
rs1441443	3	74005900	0.85	0.15
rs1446596	2	210111034	0.47	0.53
rs1459531	4	119100475	0.66	0.34
rs1461131	3	117483362	0.47	0.53
rs1466286	1	26554872	0.24	0.76
rs1468924	3	180465671	0.35	0.65
rs1468996	7	5507698	0.46	0.54
rs1472578	3	160291340	0.28	0.72
rs1479137	3	144106496	0.56	0.44
rs1479371	3	103500624	0.31	0.69
rs1487921	X	137133858	0.49	0.51
rs1491233	4	100833238	0.46	0.54
rs1494996	4	135362040	0.71	0.29
rs1501225	1	60610101	0.40	0.60
rs1508595	12	87488484	0.17	0.83
rs1510834	2	13694789	0.75	0.25
rs1512327	4	149679984	0.43	0.58
rs1521527	2	165253332	0.51	0.49
rs1523192	X	115082685	0.62	0.38
rs1524876	15	29050564	0.49	0.51
rs1530390	18	46569899	0.21	0.79

rs1534880	22	35653611	0.51	0.49
rs1536289	13	58781783	0.53	0.47
rs1536570	1	81417598	0.43	0.57
rs1538956	6	127005719	0.43	0.58
rs1541317	6	118130009	0.43	0.57
rs1542707	12	46921444	0.64	0.36
rs1545540	15	58486900	0.67	0.33
rs1548837	12	12945584	0.71	0.29
rs1551740	4	115734909	0.36	0.64
rs1553161	13	62045071	0.50	0.50
rs1554622	2	219431723	0.61	0.39
rs1558022	X	116290201	0.51	0.49
rs1560550	5	121217395	0.58	0.43
rs1570637	9	71834932	0.28	0.72
rs1630675	11	132601535	0.61	0.39
rs163077	2	38197256	0.30	0.70
rs1631833	4	110290487	0.31	0.69
rs1648282	15	43213156	0.64	0.36
rs1648312	15	43244641	0.63	0.37
rs168206	18	3683772	0.42	0.58
rs169125	6	95623707	0.03	0.98
rs1705772	12	34066023	0.61	0.39
rs1716758	X	117241790	0.21	0.79
rs173686	5	82847256	0.64	0.36
rs1739897	1	75828083	0.69	0.31
rs1792737	18	51997365	0.34	0.66
rs179562	14	30294209	0.83	0.17
rs1796048	2	97065450	0.26	0.74
rs1807912	5	109245567	0.73	0.27
rs1826734	12	101652738	0.41	0.59
rs185493	5	177923864	0.28	0.72
rs1861577	12	16455240	0.47	0.53
rs1861809	12	108708308	0.49	0.51
rs1864003	5	141763720	0.46	0.54
rs1865680	Y	6911479	0.70	0.30
rs186659	20	55239747	0.59	0.41
rs1868092	2	46525853	0.47	0.53
rs1868280	8	141965436	0.63	0.37
rs1868660	8	20530464	0.17	0.83
rs1872923	8	28957696	0.47	0.53
rs1880863	4	123239547	0.42	0.58
rs1882719	X	150272071	0.62	0.38
rs1883906	X	126414779	0.59	0.41
rs1884688	X	37260332	0.76	0.24
rs188481	13	62710392	0.58	0.42
rs1894758	13	110841296	0.23	0.77
rs1921708	X	8105875	0.47	0.53
rs1928533	6	45617802	0.65	0.35
rs1934070	X	121103292	0.60	0.40
rs1935074	X	79983248	0.71	0.29
rs1945085	11	76414893	0.41	0.59
rs1945465	11	78034146	0.24	0.76

rs1947393	14	49052346	0.62	0.38
rs1953088	6	13033922	0.63	0.37
rs1955734	14	36208379	0.45	0.55
rs1966455	Y	8916909	0.03	0.97
rs1969888	13	105648715	0.25	0.75
rs1981431	20	43408865	0.40	0.60
rs1986601	2	70119154	0.64	0.36
rs1990023	5	129649489	0.34	0.66
rs1990637	16	51595176	0.20	0.80
rs1991315	2	17761718	0.38	0.62
rs1993104	19	56932061	0.32	0.68
rs1996818	3	70395474	0.43	0.57
rs200148	6	143387389	0.68	0.32
rs2001660	2	9564039	0.42	0.58
rs2007350	1	29466743	0.35	0.65
rs2008312	2	65366295	0.69	0.31
rs2008924	Y	12675868	0.00	1.00
rs2010962	18	52101925	0.43	0.58
rs201492	7	101347573	0.56	0.44
rs2016160	1	47283621	0.32	0.68
rs2016878	X	109740071	0.44	0.56
rs2026999	9	100219712	0.72	0.27
rs204057	1	29092673	0.22	0.78
rs2040962	X	116385288	0.49	0.51
rs2046718	3	174955029	0.31	0.69
rs2051713	12	89479599	0.36	0.64
rs2054615	2	213310258	0.41	0.59
rs2055426	3	118703034	0.59	0.41
rs2058276	Y	2711817	0.67	0.33
rs2061589	12	86796951	0.12	0.88
rs2063099	16	50018372	0.48	0.52
rs2064034	X	48580224	0.51	0.49
rs2108389	19	3542590	0.50	0.50
rs2124036	8	126717316	0.72	0.27
rs2151065	9	16235716	0.48	0.52
rs2164062	18	17233261	0.18	0.83
rs2179393	1	11340237	0.56	0.44
rs2200290	X	126713002	0.28	0.72
rs225160	4	52725167	0.60	0.40
rs2252257	22	17014854	0.14	0.86
rs226386	12	9153994	0.53	0.47
rs2273348	1	11013343	0.14	0.86
rs2290753	1	240132479	0.69	0.31
rs2291409	1	240058238	0.69	0.31
rs2296412	6	84625643	0.98	0.02
rs236919	11	116600571	0.29	0.71
rs2370409	3	134492512	0.59	0.41
rs2377473	20	30313909	0.36	0.64
rs2442567	8	6445077	0.02	0.98
rs2485729	X	138707739	0.01	0.99
rs2642995	1	243498606	0.60	0.40
rs2686085	3	198711008	0.77	0.23

rs268666	19	45610005	0.52	0.48
rs276990	16	84778717	0.62	0.38
rs2837888	21	41257856	0.46	0.54
rs2837956	21	41401386	0.39	0.61
rs2839899	9	77580553	0.71	0.29
rs2873	15	29018547	0.49	0.51
rs2873108	7	153888819	0.83	0.18
rs2878079	1	242716131	0.89	0.11
rs2961408	X	40013862	0.13	0.87
rs2967305	16	80877154	0.77	0.23
rs307195	5	6120107	0.36	0.64
rs310935	13	61931782	0.47	0.53
rs311848	14	58270833	0.35	0.65
rs31251	5	130861845	0.58	0.42
rs3294	7	98293016	0.22	0.78
rs337514	12	61538458	0.41	0.59
rs34609	5	60517165	0.50	0.50
rs352476	15	63592442	0.39	0.61
rs354731	20	58384823	0.57	0.43
rs359280	10	17359713	0.66	0.34
rs369643	6	149841491	0.50	0.50
rs3734693	6	44073143	0.83	0.18
rs374653	15	91011734	0.73	0.27
rs3750203	8	144803169	0.62	0.37
rs3780346	9	92053743	0.50	0.50
rs3845596	1	14934605	0.58	0.42
rs387812	15	56862078	0.59	0.41
rs3899	Y	7334895	0.00	1.00
rs3901	Y	20111123	1.00	0.00
rs4076107	X	39696224	0.03	0.97
rs4107736	8	29995506	0.60	0.40
rs4246828	8	144240466	0.71	0.29
rs461785	16	64366545	0.47	0.53
rs4675966	2	241964332	0.47	0.53
rs477467	12	118390712	0.37	0.63
rs4778137	15	26001430	0.73	0.27
rs4846012	1	11492192	0.35	0.65
rs4900525	14	101474494	0.70	0.30
rs4904574	14	88968503	0.28	0.72
rs4949	X	146704727	0.69	0.31
rs4957114	5	1009974	0.78	0.22
rs4969481	17	77586501	0.13	0.87
rs501110	4	99538555	0.34	0.66
rs520354	2	21171264	0.53	0.47
rs525869	X	90402304	0.63	0.37
rs530501	X	148273210	0.49	0.51
rs531577	3	139883997	0.41	0.59
rs535534	13	27107323	0.39	0.61
rs537111	X	108534073	0.47	0.53
rs540819	11	104371671	0.55	0.45
rs565973	18	61233578	0.55	0.45
rs573615	15	41401573	0.31	0.69

rs591510	10	5981959	0.67	0.33
rs591728	3	101916282	0.68	0.32
rs595241	12	132264929	0.49	0.51
rs6047134	20	2089054	0.27	0.72
rs625628	18	52354151	0.53	0.48
rs6426327	1	242415852	0.87	0.13
rs6474795	9	1389674	0.74	0.26
rs6486532	12	129261499	0.54	0.46
rs6503211	17	9333425	0.19	0.81
rs6540401	X	147024359	0.10	0.90
rs6743486	2	78435569	0.08	0.92
rs6808013	3	187432931	0.79	0.21
rs7010024	8	81940797	0.88	0.12
rs709160	3	12501402	0.57	0.43
rs711814	2	176794848	0.70	0.30
rs717239	4	77532922	0.41	0.59
rs717571	17	72132418	0.48	0.52
rs717651	13	30070594	0.60	0.40
rs718066	19	45013680	0.43	0.57
rs718206	1	11228740	0.29	0.71
rs718251	8	52877076	0.71	0.29
rs719185	13	107079922	0.62	0.38
rs722263	7	92531463	0.68	0.32
rs722269	6	42302894	0.33	0.68
rs722317	11	15880138	0.43	0.57
rs722497	1	29841137	0.40	0.60
rs726111	4	6030055	0.53	0.48
rs727056	6	170044	0.26	0.74
rs727345	10	31939079	0.38	0.62
rs727619	6	170623826	0.63	0.38
rs729639	3	13801855	0.34	0.66
rs731544	10	35697409	0.53	0.47
rs736201	5	78873312	0.63	0.37
rs736779	2	68483570	0.64	0.36
rs737516	3	43533089	0.45	0.55
rs738402	22	25007934	0.43	0.57
rs739096	22	35066240	0.54	0.46
rs740158	7	76700487	0.54	0.46
rs741418	2	75274841	0.60	0.40
rs742585	X	45834201	0.41	0.59
rs743151	X	37784743	0.60	0.40
rs747398	20	15497323	0.77	0.23
rs749477	3	10631823	0.59	0.41
rs7528979	1	10027723	0.18	0.82
rs753012	8	9019806	0.29	0.71
rs753842	19	4928744	0.20	0.80
rs755569	2	129793500	0.53	0.47
rs756658	22	17830390	0.48	0.52
rs758439	X	147770441	0.72	0.28
rs760109	X	150223090	0.47	0.53
rs760335	14	93884697	0.46	0.54
rs762318	3	38491373	0.72	0.28

rs764602	14	48550169	0.47	0.53
rs766325	1	18701764	0.55	0.45
rs767022	6	9084652	0.19	0.81
rs767778	13	79095116	0.30	0.70
rs769322	8	119738549	0.37	0.63
rs7860423	9	138231384	0.19	0.81
rs7881297	X	121338454	1.00	0.00
rs813779	2	5540855	0.14	0.86
rs816943	20	62146453	0.57	0.43
rs828869	2	74235155	0.61	0.39
rs829290	X	95972054	0.46	0.54
rs839556	6	142913571	0.58	0.42
rs841603	12	27202749	0.65	0.35
rs852223	3	132129178	0.63	0.37
rs8627	1	8347201	0.49	0.51
rs877826	5	138646696	0.32	0.68
rs878400	9	127987598	0.63	0.37
rs883434	2	233040183	0.45	0.55
rs884080	1	2058911	0.63	0.37
rs884839	8	67182542	0.88	0.13
rs885978	22	17572780	0.51	0.49
rs890459	4	29116098	0.35	0.65
rs893367	3	53884771	0.48	0.52
rs895530	Y	8006392	0.30	0.70
rs898271	13	90539922	0.48	0.52
rs901170	8	25826193	0.65	0.35
rs903770	12	115754677	0.72	0.28
rs906895	11	6236824	0.54	0.46
rs913199	1	65583083	0.43	0.57
rs913258	9	4867246	0.84	0.16
rs915180	1	152892156	0.36	0.64
rs915774	21	44400145	0.75	0.25
rs916041	12	11703597	0.47	0.53
rs916208	X	127666786	0.47	0.53
rs917711	18	52007286	0.47	0.52
rs918044	12	125402711	0.37	0.63
rs924901	2	77495618	0.60	0.40
rs933938	8	70104749	0.58	0.43
rs936013	15	58505557	0.33	0.67
rs942150	6	131744986	0.47	0.53
rs953114	X	40953796	0.56	0.44
rs959419	19	20618532	0.39	0.61
rs960345	4	105588251	0.45	0.55
rs962272	17	44333282	0.52	0.48
rs963314	7	147300280	0.43	0.58
rs963447	X	15313153	0.36	0.64
rs964176	2	160149391	0.56	0.44
rs966707	6	51736178	0.57	0.43
rs971879	X	141070074	0.52	0.48
rs972881	2	107478040	0.30	0.70
rs9740	3	123487743	0.63	0.37
rs975612	2	72300989	0.64	0.36

rs980099	X	106118132	0.82	0.18
rs981270	14	85249374	0.33	0.67
rs990949	12	3356843	0.70	0.30
rs994502	10	26540172	0.82	0.18
rs999634	4	52768041	0.64	0.36

Appendix E- β and p values calculated from the alternative statistical approach

The calculated β and corresponding p values for all mixtures from the alternative statistical approach are listed in the table.

	β F	p F	β C	p C	β B	p B	β D	p D	β H	p H
Blind 1	0.0793	0.0041572	0.0042	0.7243232	0.3642	4.91E-22	0.2792	1.22E-21	0.3619	7.84E-47
Blind 2	0.0709	0.003784	0.2247	1.39E-20	0.3222	7.00E-22	0.2344	1.31E-19	0.2917	7.58E-38
Blind 3	0.1007	0.0005034	0.0462	0.1096121	0.5370	7.28E-50	0.4349	2.06E-51	0.0550	0.0516274
Blind 4	0.1451	4.16E-05	0.0456	0.2044734	0.0644	0.2073797	0.8801	3.15E-205	0.0268	0.4423851
Blind 5	0.1449	6.75E-09	0.2000	1.64E-15	0.2990	8.68E-18	0.2203	1.80E-16	0.2567	2.84E-27
Blind 6	0.2947	3.33E-25	0.0243	0.411521	0.4164	2.20E-25	0.3495	1.16E-30	0.0301	0.29513
Blind 7	0.1997	1.20E-15	0.0193	0.4559632	0.3402	3.69E-24	0.2545	2.96E-21	0.3261	1.79E-44
Blind 8	0.4833	1.60E-69	0.0219	0.4748424	0.0656	0.1284956	0.5275	2.86E-70	0.0293	0.3230224
Blind 9	0.3872	4.36E-51	0.0279	0.316613	0.4016	1.86E-27	0.2789	1.07E-21	0.0383	0.1581302
Blind 10	0.3929	6.30E-60	0.0083	0.7568391	0.2531	2.90E-13	0.1449	3.98E-07	0.3398	6.28E-45
Blind 11	0.6054	1.47E-146	0.1188	4.96E-05	0.0864	0.0246821	0.1608	3.37E-07	0.1467	1.90E-07
Blind 12	0.7463	1.14E-180	0.0138	0.6862107	0.0545	0.2415142	0.2710	8.14E-14	0.0195	0.5545119
Blind 13	0.7077	8.86E-147	0.0035	0.918842	0.2295	7.91E-07	0.1335	0.0003055	0.0271	0.4138184
Blind 14	0.2225	1.30E-17	0.0102	0.706616	0.3429	2.09E-22	0.1346	3.48E-06	0.4284	5.06E-75
Blind 15	0.1576	1.08E-10	0.1269	3.09E-07	0.3975	5.69E-35	0.2561	4.54E-23	0.2221	2.72E-21
Blind 16	0.0675	0.0776994	0.0517	0.1747919	0.0702	0.1942861	0.9527	1.60E-220	0.0287	0.4429662
Blind 17	0.3989	1.54E-40	0.0224	0.4845286	0.0614	0.1795498	0.6103	2.89E-91	0.0291	0.3496661
Blind 18	0.1051	3.58E-05	0.1121	1.40E-05	0.3288	1.94E-21	0.2044	6.16E-14	0.3575	2.12E-54
Blind 19	0.0737	0.0080345	0.0098	0.7290863	0.3632	4.24E-22	0.1946	9.03E-11	0.4532	1.35E-78
Blind 20	0.1727	1.48E-09	0.0435	0.1321999	0.5835	2.78E-60	0.3120	1.84E-25	0.0563	0.0452358
Blind 21	0.3067	5.43E-22	0.0287	0.3847978	0.0588	0.2154093	0.6888	7.39E-118	0.0218	0.4962055
Blind 22	0.0250	0.3091922	0.2409	1.21E-23	0.3379	5.00E-24	0.2348	2.18E-19	0.3001	5.02E-40
Blind 23	0.0394	0.1434274	0.0192	0.4818159	0.3666	8.41E-25	0.2921	7.27E-25	0.3964	1.10E-61
Blind 24	0.0548	0.0845999	0.0394	0.2103007	0.5661	6.46E-43	0.4489	3.73E-45	0.0476	0.1230281
Blind 25	0.0967	0.012178	0.0416	0.2820125	0.0691	0.2234623	0.8937	4.66E-159	0.0147	0.696472

Appendix F – List of vendors

Applied Biosystems, Foster City, USA. http://www.appliedbiosystems.com

Illumina, California, USA. http://www.illumina.com/

Invitrogen, Paisley, UK. http://www.invitrogen.com

Qiagen, Hilden, Germany. http://www1.qiagen.com

Sequenom, California, USA. http://www.sequenom.com/

Thermo Fisher Sientific, Willmington, USA. http://www.nanodrop.com/

Zymo Research, California, USA. http://www.zymoresearch.com/