



Evaluation of DNA Polymorphisms for Kinship Testing in the Population of
Saudi Arabia

Hussain Mohammed H. Alsafiah

A thesis submitted in partial fulfilment for the requirements for the degree of Doctor
of Philosophy at the University of Central Lancashire

April 2019

I. Student Declaration Form

Type of Award PhD
School Forensic and applied sciences

1. I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

2. Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

Signature of Candidate



Print name

Hussain Mohammed H. Alsafiah

I. Abstract

Short Tandem Repeats (STRs) have been the standard DNA marker employed in forensic laboratories for more than two decades. Due to the advances in the kit chemistries and separation technologies (capillary electrophoresis (CE) systems), the number of STRs that can be simultaneously typed has grown to 21-26; this provides sufficient confidence in the conclusions of most kinship cases. However, more complex cases (e.g. testing distant relatives, potential mutations, deficient cases or incest cases) or when the target population shows an increased level of consanguinity, the genetic evidence may prove inconclusive. This necessitates testing additional STRs included in supplementary STR kits. Another option is by using Massively Parallel Sequencing (MPS) systems that allow simultaneous sequencing of additional DNA markers.

A total 500 samples from the population of Saudi Arabia were collected. Two CE-based STR kits were used: Globalfiler™ PCR amplification kit (AB, USA) and SureID® 23 comp Human Identification kit (Health Gene Technologies, China) that together allowed 38 aSTRs to be analysed.

In addition, as the SureID® 23 comp kit has not been validated either by an independent laboratory or by the manufacturer, the kit was validated following the minimum criteria of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDM).

Moreover, the ForenSeq™ DNA Signature Prep Kit (Verogen) was used to sequence 87 samples and to generate sequence-based data for 122 autosomal markers included in the kit. The project allowed, in total, obtaining size-based data for 136 autosomal markers (42 aSTRs and 94 iiSNPs) and sequence-based data for 122 autosomal markers

(28 aSTRs including SE33 and 94 iiSNPs). The data were evaluated for human identification and kinship testing in Saudi Arabia

Although Globalfiler™ kit provided combined match probability (CMP) of $1.42E-26$ that is much higher than the kit currently used in Saudi Arabia that has a CMP of $2.23E-18$ (Identifiler plus kit), the availability of data for 42 aSTRs allowed other commercially available kits to be evaluated (based on the loci they contain). The study suggests adopting VeriFiler™ Plus (AB) or PowerPlex Fusion 6C system (Promega Corporation, USA) as a standard STR kit that would provide the lowest CMPs ($9.26E-29$ and $1.03E-29$, respectively). Adopting any of the three kits would provide sufficient confidence in most parent-child cases (trio or duo).

The validation of the SureID® 23 comp has shown that the kit met the criteria commonly used in forensic genetics laboratories. In addition, the kit can benefit from some developments that were identified by the validation, in particular the addition of extra alleles in the allelic ladder and also to increase the amount of input DNA that can be added to an amplification. The kit can be used if any kinship cases showed inconclusive results with GlobalFiler™, VeriFiler™ Plus or PowerPlex Fusion 6C allowing 38-40 aSTRs to be analysed.

The ForenSeq™ DNA Signature Prep Kit provided CMP of $1.97E-68$ and $3.65E-77$ for the size and sequence-based data respectively, where $1.24E-37$ (size-based data) and $5.6E-41$ (sequence-based data) were provided from the iiSNPs alone. The kit can be used when two or three mismatches were suspected to be mutations or when testing distant relationships.

The study highlighted 220 syntenic pairs, 46 of which would have significant impact on LR estimation due to lower RFs (< 0.12). The case-specific impact of linkage should be included in the estimation of LRs by using the RFs values estimated in this project.

II. Table of Contents

1	Chapter One: Literature Review	1
1.1	History of using STRs for human identification	1
1.2	National DNA Databases (NDNADs)	5
1.3	Capillary electrophoresis (CE) and labelled primers	9
1.4	Internal validation of new multiplex kits	10
1.5	Applications of STR-based systems	13
1.5.1	Criminal investigation	13
1.5.2	Kinship testing	14
1.5.3	Ancestry testing	18
1.6	Hardy Weinberg equilibrium (HWE)	20
1.7	Linkage disequilibrium (LD)	22
1.8	Common forensic statistical parameters	23
1.8.1	Match probability (MP) and power of discrimination (PoD)	23
1.8.2	Power of exclusion (PoE)	23
1.8.3	Polymorphic information content (PIC)	23
1.8.4	Analysis of molecular variance (AMOVA)	23
1.9	Limitations of STR-CE systems	24
1.10	Single Nucleotide polymorphisms (SNPs)	24
1.11	Massively Parallel Sequencing (MPS)	29
1.11.1	Library preparation	32
1.11.2	Cluster generation	34
1.11.3	Sequencing	36
1.11.4	Data analysis	38
1.11.5	Advantages and disadvantages of MPS systems	39
1.12	Project Background	42
1.13	Project Aims	47
1.14	Objectives	47
2	Chapter Two: Materials and Methods	49
2.1	Background	49
2.2	Samples collection and preparation	50
2.2.1	Ethical approval	50
2.2.2	Samples collection	50
2.2.3	DNA extraction	52
2.2.4	Quantification of the extracted DNA	53
2.3	GlobalFiler™ PCR kit.	54
2.3.1	DNA amplification	54
2.3.2	DNA separation, detection and analysis	55
2.4	Characterisation of six unusual alleles at SE33 and D1S1656 STR loci.	55
2.4.1	Sequencing the SE33 alleles	55
2.4.2	Sequencing the D1S1656 alleles	57
2.5	An evaluation of the SureID®23comp Human Identification kit.	57
2.5.1	Preparation ABI 3500 DNA Genetic Analyser.	57
2.5.2	DNA Samples	58
2.5.3	DNA amplification	59

2.5.4	DNA separation, detection and analysis	60
2.6	ForenSeq™ DNA Signature Prep kit.	61
2.6.1	Library preparation and sequencing for the D1S1656 variants.	61
2.6.2	Library preparation and sequencing for the Saudi population data.	61
2.6.3	Universal analysis software	62
2.6.4	Concordance study	62
2.6.5	Further analysis using the STRait Razor (SR)	63
2.6.6	SE33 sequence-based data	63
2.6.7	Novelty assessment	63
2.7	Evaluation of DNA markers	64
2.7.1	Forensic parameters	64
2.7.2	Hardy-Weinberg equilibrium	64
2.7.3	Linkage disequilibrium test	64
2.7.4	Population differentiation test, FST calculation, and inbreeding coefficient (FIS)	65
2.7.5	RStudio platform and packages used in the project.	66
2.8	Gel electrophoresis	66
2.8.1	Assessment of extraction procedure and DNA yield	66
2.8.2	Preparation of the 20-cm-long 3% agarose	66
2.9	An evaluation of 136 DNA markers for kinship testing	67
2.9.1	Setting up the mutation rates in the Familias3 software	68
2.9.2	Simulation study	69
2.9.3	Estimating the genetic distance between syntenic pairs	72
2.9.4	Calculation of recombination fraction (RF) using Kosambi mapping function.	73
3	Chapter Three: An evaluation of 21 autosomal STRs for the population of Saudi Arabia using the Globalfiler™ PCR Amplification Kit.	75
3.1	Overview of experiment	75
3.2	Aims of the study	75
3.3	Objectives	76
3.4	Materials and Methods	76
3.5	Results and discussion	77
3.5.1	Ethical approval	77
3.5.2	Sample collection	77
3.5.3	DNA extraction	79
3.5.4	Quantification of the extracted DNA	80
3.5.5	Validation of half volume reaction and the 50 cm capillary with POP6.	81
3.5.6	Allelic ladder variants	83
3.5.7	Population genetics	86
3.5.8	Consanguinity in the population of Saudi Arabia.	93
3.5.9	Population comparison	93
3.6	Conclusion	99
4	Chapter Four: Characterisation of STR allele variants detected in Saudi population.	101
4.1	Overview of experiment	101
4.2	Aims of the study	103
4.3	Objectives	104
4.4	Materials and Methods	104

4.5	Results and discussion	104
4.5.1	SE33 variants	104
4.5.2	D1S1656 variants	108
4.6	Conclusion	111
5	Chapter Five: An evaluation of 17 non-CODIS STRs for the population of Saudi Arabia using the SureID® 23comp Human Identification Kit.	113
5.1	Overview of experiment	113
5.2	Aims of the study	117
5.3	Objectives	118
5.4	Materials and Methods	119
5.5	Results and discussion	119
5.5.1	Preparation ABI 3500 DNA Genetic Analyser	119
5.5.2	D5 locus confirmation	120
5.5.3	Repeatability and Reproducibility.	121
5.5.4	Sensitivity stochastic effect.	121
5.5.1	Performance against common PCR inhibitors	124
5.5.2	Further performance assessment	129
5.5.3	Heterozygote peak balances.	131
5.5.4	Stutter/corresponding allele ratios.	135
5.5.5	Precision and accuracy.	135
5.5.6	Concordance study	139
5.5.7	Allelic ladder and rare alleles.	139
5.5.8	Population study and excess of homozygosity.	142
5.5.9	Population comparison	148
5.5.10	STRidER quality control	150
5.6	Conclusion	150
6	Chapter Six: Population Genetic Data For 122 DNA Markers for The Saudi Arabian Population Using the ForenSeq™ DNA Signature Prep Kit.	152
6.1	Overview of experiment	152
6.2	Aims of the study	153
6.3	Objectives	154
6.4	Materials and Methods	155
6.5	Results	156
6.5.1	Run metrics, sequencing results, and depth of coverage (DoC)	156
6.5.2	Sequence variations	163
6.5.3	The impact of sequence variations on discrimination power and heterozygosity	169
6.5.7	Sequence-Based Saudi Population Data for The SE33 Locus	176
6.6	Discussion	179
6.7	Conclusion	188
7	Chapter Seven: Evaluation Study of 136 DNA Markers for Kinship Applications in Saudi Arabia.	190
7.1	Overview of experiment	190
7.2	Aims of the study	191
7.3	Objectives	192
7.4	Materials and Methods	192
7.5	Results and discussion	193
7.5.1	Confirmation of the parent-child relationship of the pedigree's members.	193

7.5.2	Simulation results	195
7.5.3	Performance of the marker sets	222
7.5.4	Potential linkage effect of closely located markers	223
7.5.5	Defining thresholds for kinship testing in Saudi Arabia	235
7.5.6	Defining the number of tested markers for each relationship	235
7.6	Conclusion	236
8	Chapter Eight: General Conclusion	238
8.1	Human identification application in Saudi Arabia	240
8.2	Kinship testing in Saudi Arabia	242
8.3	Evidence of consanguinity in the population of Saudi Arabia	244
8.4	Future work	245
9	Chapter Nine: References	246
10	Chapter Ten: Appendixes	270
10.1	Appendix 1	270
10.1.1	Scenarios specific equations that used for calculating the RI.	270
10.1.2	Scenarios specific equations that used for calculating the RI when the child is missing	271
10.1.3	Scenarios specific equations that used for calculating the SI and HSI.	272
10.1.4	Including the mutation event into the RI-LR	273
10.1.5	Including the prior probability (Pr) to the posterior probability (Po).	276
10.1.6	RMNE calculation.	276
10.2	Appendix 2	279
10.2.1	Participant Information Sheet	279
10.2.2	consent form	282
10.3	Appendix 3	283
10.3.1	Sample collection approval from the Security Forces Hospitals Programme.	283
10.3.2	STEMH 557 ethical approval for the project	284
10.4	Appendix 4	285
10.4.1	STRidER final report for the data of the 17 non-CODIS loci.	285
10.5	Appendix 5	288
10.6	Appendix 6	330
10.6.1	Combined exceedance probability Figures	330
10.6.2	Cumulative genetic map distances (cM) of 95 SNPs.	334
11	Chapter Eleven: Publications and Participations	336
11.1	Publications	336
11.2	Participations	336

III. List of Tables

Table 1.1. Currently available autosomal STR kit that provided by AB, Promega Corporation and Qiagen. Data from https://www.thermofisher.com , https://www.promega.co.uk/ , and https://www.qiagen.com . (RM) rapidly mutating Y-STR.	4
Table 1.2. A list of the developed aiSNPs panels. This table summarizes a list of recently developed aiSNPs that have been adopted in forensic laboratories (an original table based on information from (Fondevila <i>et al.</i> 2013, Gettings <i>et al.</i> 2014, Rogalla <i>et al.</i> 2014).	26
Table 1.3. iiSNPs panels developed for human identification. The table shows the iiSNPs panels developed for human identification and the progress in match probability (MP) from 2006 – 2016. This table summarizes effort by research groups to select informative iiSNPs that can be applied globally leading to the 54 SNPs with 1.3E-22 match probability (MP).	27
Table 1.4: the maximum possible heterozygosity calculation based on maximum allele frequencies of SNP types.	28
Table 1.5. DNA markers included in three STR-MPS kits commercially available (Faith and Scheible 2016, Applied Biosystems 2017, Verogen 2018a).	31
Table 1.6. An example of how the ForenSeq™ Universal Analysis Software (UAS) reports the sequences in the Flanking Region Report. Here, the sequences of the D5S818 STR and of the rs560681 SNP were used for illustration. For STRs, the Flanking Region Report highlights variants within the repeat region in black colour (enlarged) and highlights variants in flanking region in blue colour (enlarged). In SNPs, the Flanking Region Report highlights the target SNP in black colour (enlarged) and highlights variants in flanking region in blue colour (enlarged). All highlighted variants in the flanking region are predefined and variants that were not in the predefined list are not highlighted by the software but still reported.	38
Table 1.7. Different sequences of allele 20 at D2S1338 locus using MPS technologies. STR-CE systems distinguish alleles by their sizes (Gettings <i>et al.</i> 2016).	39
Table 1.8. Comparison of the power of discrimination of four loci by using STR-CE systems, MPS systems for variants in the repeat region and for variants in both repeat and flanking regions. This study was conducted to examine STR loci variations for the Koreans population (Kim <i>et al.</i> 2017).	39
Table 1.9. Comparison between the number of alleles obtained by size-based systems and by MPS systems. This table showing data that compares the number of alleles obtained by size-based systems and by MPS systems for 23 of the most common used autosomal STRs (Gettings <i>et al.</i> 2016), SE33 data (Gettings <i>et al.</i> 2015).	40
Table 2.1. The reagents and suppliers used in the experimental work.	49
Table 2.2. The components of amplification tubes of the GlobalFiler™ PCR kit. The table shows the components of amplification tubes using the half volume reactions of the GlobalFiler™ PCR kit used in the project.	54

Table 2.3. Bone samples used in the evaluation tests of the SureID® 23 comp kit. Nine bone samples, collected from a mass grave in Iraq, were extracted using PrepFiler™ BTA Forensic DNA Extraction Kit (AB), and were quantified using Quantifiler™ Trio DNA Quantification Kit (AB). This table shows Quantifiler™ Trio small fragment concentrations (ng/μl) and degradation indexes (DI) of the samples (Alsafiah <i>et al.</i> 2019a).	59
Table 2.4. Mutation rates for aSTRs that were reviewed from literatures. The mutation rates of 38 aSTRs were reviewed from (Butler 2015, Lan <i>et al.</i> 2018, Jin <i>et al.</i> 2016) and were used in the simulation study. No mutation rates were available for D3S1744, D4S2366, D19S253 and D21S2055.	68
Table 2.5. The hypotheses 1 and 2 that were used in the simulation study. The simulation study was conducted using Familias3 software v3.2.7 (Kling <i>et al.</i> 2014). A total of 8 scenarios for five different relationships were tested. The table also shows members who were simulated (genotyped) in each run (orange colour).	71
Table 3.1. The number of participants per each city. This table is showing samples numbers collected from each city where Riyadh and Dammam cities had the highest number of samples.	78
Table 3.2. Results of expected heterozygosity calculation and of Hardy-Weinberg equilibrium exact test, conducted by Arlequin v3.5.2.1 software for the 21 STR loci. The <i>P</i> values after Bonferroni correction is significant if $P < 0.002$. The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/21 \text{ STRs} = 0.002$.	87
Table 3.3. Results of linkage disequilibrium tests of syntenic loci included the GlobalFiler® PCR amplification kit. The results showed that the tested samples did not show linkage disequilibrium (<i>P</i> value > 0.05). (p-q) indicates syntenic loci located in different arms.	87
Table 3.4. Allele frequency data and forensic statistical parameters of 21 aSTRs included in GlobalFiler® PCR amplification kit for the population of Saudi Arabia. The parameters included: matching probability, power of discrimination, polymorphism information content, power of exclusion, observed homozygosity and observed heterozygosity that were generated using the PowerStat v 1.2 (Alsafiah <i>et al.</i> 2017).	89
Table 3.5. An assessment of the 21 loci included in the GlobalFiler™ kit for kinship testing. This table shows the paternity probabilities for a typical paternity case by using combined typical paternity index for different prior probabilities ($Pr = 0.90, 0.50$ and 0.10). The GlobalFiler™ kit showed much higher (~300-fold) probabilities comparing to those probabilities calculated when using the currently used kit in Saudi Arabia (Identifiler® Plus).	91
Table 3.6. The maximum matching loci within the 500 samples. In the 500 samples, only two pairs of samples showed full matching in 9 loci (i.e. both alleles); this was the maximum number of matched loci (shaded row). One pair showed partial matching (i.e. one of the two alleles) at 19 out of 22 loci (shaded column). This table was generated by the R studio using the package of DNA tools.	92

Table 3.7. Population differentiation exact test results using the Arlequin v3.5.2.1 software. Shaded cells indicate significant differences (P value < 0.002). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/21$ STRs = 0.002. N/A values indicate data that was not collected during each of the previous studies (Alsafiah <i>et al.</i> 2017).	96
Table 4.1. Sequence structure of the SE33 and D1S1656 loci. (A) Shows the sequence structure of an SE33 allele that comprised of the 5'-local flank (15 bp), repeat region, and 3'-local flank (24 bp). (B) A typical sequence structure of a D1S1656 allele is shown. The sequence structure of reference alleles (SE33, allele 26.2 GenBank: V00481.1) and (D1S1656, allele 15.3 GenBank: G07820.1) are given for illustration. Based on the published guidelines of the International Society for Forensic Genetics (ISFG) (Parson <i>et al.</i> 2016), the local flank regions showed in A and B (greyed out sequences) are not counted in allele calling system (Alsafiah <i>et al.</i> 2018).	102
Table 4.2. Classification of the SE33 motifs. The table shows the eleven motif patterns that had >1% frequency in the tested populations. Most sequence-based alleles show A0 and A1 patterns (Borsuk <i>et al.</i> 2018).	102
Table 4.3. Sequence data for the forward strand of 4 previously uncharacterized SE33 alleles: 2, 14.3, 20.3, and 38. The 5'uncounted sequence (15 bp) and 3'uncounted sequence (24 bp) of the local flank region; and the extended 3'-flank are shown. The amplicons sizes of the GlobalFiler kit and of primer pair (SE33-1 and SE33-2) used in this study are shown. It also shows allele names based on their sizes, based on the sequence data, and the motif pattern based on the classification of Borsuk <i>et al.</i> , (2018). Allele 2 had B1 motif and showed 17 repeats on the repeat region, but a 60 bp deletion in the extended 3'-flank led to the observation of the allele 2 based on the size. Allele 14.3 had 18.1 repeats on the repeat region, and the 14.3 size-based allele resulted from a 14 bp deletion in the extended 3'-flank. In addition, the motif pattern of this allele is novel (has not been reported in the classification of Borsuk <i>et al.</i> (2018)). The Allele 20.3 had D1 motif and showed a TTT within the repeat region. Allele 38 contained two hexanucleotide repeats (A7 motif) within the repeat region. ^(a) Represents rs1045867314 SNP at Location 6:88277260 in the allele 14.3 (Alsafiah <i>et al.</i> 2018).	107
Table 4.4. The results of hair and eye colour prediction. Both samples showed high probabilities of having brown eyes and brown or black hair. These features are more likely in the population of Saudi Arabia.	111
Table 5.1. Supplementary autosomal STRs included in 3 supplementary autosomal STR kits. The kits are Microreader™ 23sp ID (Li, J. <i>et al.</i> 2017) (Suzhou Microread Genetics), Goldeneye™ DNA ID 22NC (Fu <i>et al.</i> 2018) (Goldeneye® Technology Ltd.), AGCU 21+1 (Zhu <i>et al.</i> 2015) (AGCU ScienTech Incorporation). These kits are only commercially available in China (Phillips 2017). The table also shows a set of 25 supplementary STRs and amelogenin (26plex) recommended by the NIST, but no multiplex combining these STRs is commercially available.	114
Table 5.2. STR Markers included in the SureID® 23comp kit. This table shows the locations (GRCh38) and repeat structures of the 22 STRs included in the SureID® 23comp kit. Five loci are common with the CODIS and the ESS. Twelve loci are not included in other available supplementary kits (Investigator® HDplex and PowerPlex® CS7). All information was adapted from (Qiagen 2012a, Promega Corporation 2016, Phillips <i>et al.</i> 2018b) (Alsafiah <i>et al.</i> 2019a).	116

Table 5.3. The results of the bone samples used in the validation tests of the SureID® 23 comp kit. Nine samples, collected from a mass grave in Iraq, were extracted using PrepFiler™ BTA Forensic DNA Extraction Kit (AB), and were quantified using Quantifiler™ Trio DNA Quantification Kit (AB). This table shows Quantifiler™ Trio small fragment concentrations (ng/μl), total DNA quantities added to the PCRs of SureID® 23 kit and other kits. The percentages of detected alleles of autosomal STRs (aSTRs) when using different STR kits, are also shown. The two samples that showed lower detection rate are shaded (Alsafiah <i>et al.</i> 2019a).	130
Table 5.4. PCRs contents for the SureID® 23comp, PowerPlex® 21. GlobalFiler™, PowerPlex® Fusion 6C. The table shows the contents of the 25 μl volume PCRs for four kits used to genotype the bone samples. The SureID® 23comp has less space (6.25 μl) for DNA input compared to the other three kits (15 μl). Increasing the concentration of the master and primer mixes will increase the space for the DNA input (Alsafiah <i>et al.</i> 2019a).	131
Table 5.5. Peak balance ratios study for the SureID® 23comp kit. The table shows the average of peak balance ratios calculated for the amelogenin (AMEL) and 22 STRs included in the SureID® 23comp kit. A total of 90/500 samples were used to study balance ratios. The 10 μl reaction volume was evaluated using three DNA quantities 0.5, 0.35, and 0.25 ng. The 0.5 ng showed the highest peak ratios average. The D21S2055 showed the lowest ratio at all DNA quantities (shaded row) (Alsafiah <i>et al.</i> 2019a).	132
Table 5.6. Alleles not represented by the allelic ladder of SureID® 23comp kit detected in the population of Saudi Arabia; 34 alleles were detected at 15 STRs. It shows also the frequency of these alleles in Ningbo population (data provided by the Health Gene Technologies). The frequencies of detected alleles ranged from 0.001 (one observation) to 0.066 (66 observations). Shaded rows indicate alleles observed ≥ 40 times (Alsafiah <i>et al.</i> 2019a).	140
Table 5.7. Results of the expected heterozygosity calculation and of Hardy-Weinberg equilibrium exact test, conducted by Arlequin v3.5.2.1 Software for the 17 non-CODIS loci included in the SureID® 23comp kit. The <i>P</i> values is significant if < 0.05. The five common loci with the GlobalFiler™ kit were not included in this table as they had the same results in Table 3.2.	142
Table 5.8. Allele frequency of the 17 non-CODIS loci. The table shows the allele frequency and statistical parameters for the 17 non-CODIS included in the SureID® 23comp kit. Allele frequencies for D1S1656, D2S441, D10S1248, D12S391 and D16S539 are not shown as they are presented in Table 3.4.	144
Table 5.9. The maximum of matched loci per any sample pair within the 500 samples. In the 500 samples, only two pairs of samples showed full matching in 9 loci (i.e. both alleles). This was the maximum number of matched loci (shaded row). One pair of sequences showed partial matching (i.e. one of the two alleles) at 20 out of 22 loci (shaded column). This table was generated by the R studio using the package of DNA tools (Alsafiah <i>et al.</i> 2019a).	147
Table 5.10. Population differentiation exact test results using the Arlequin v3.5.2 Software. Shaded data indicates significant differences (<i>P</i> value < 0.05). European and South Asian populations showed lower number of STRs with significant difference.	149

Table 6.1. Identity informative SNPs included in the ForenSeq™ DNA signature prep kit. The table shows the amplicon sizes and the chromosomes of 94 iiSNPs included in this study (Verogen 2018a).	154
Table 6.2. Samples with partial DNA profiles for the 27 aSTRs and the 94 iiSNPs. Shaded cells represent sequences below the default thresholds. All samples presented here had lower average reads count comparing to other sample. This has led to allele drop out in PentaE, rs1357617, rs2920816, and rs1736442. The D22S1045 was previously genotyped in Chapter 3 and all samples presented in the table had heterozygous genotypes. Due to the lower coverage of samples presented here and the lower allele count ratio (ACR) feature of D22S1045, the absence of the second allele in samples 4,7 and 10 was considered as alleles drop out not discordance.	157
Table 6.3. The four samples that showed lower ACRs at D22S1045. The table shows the CE data, ForeSeq data (including the true alleles, coverage and the ACRs) and the n-4 stutter of allele 1 (including coverage of the -4 stutter and stutter ratios). The four samples showed relatively lower ACRs, two of which (shaded rows) had stutter ratios of the n-4 stutter of allele 1 greater than the ACR of the second true allele (allele 2).	160
Table 6.4. Perfect association between the target iiSNPs and variants in the flanking region. This table shows association that was noticed between the target iiSNPs and variants in the flanking region. Black colour indicates the target iiSNPs and the blue colour indicates variants within the flanking region. SNPs that showed perfect association are underlined.	166
Table 6.5. Variants at the flanking region of two aSTRs and of 11 iiSNPs. The table shows 14 variants identified in this study which were reported by the UAS but were not highlighted in blue. The table presents the marker's name, allele call (CE), rs identifiers if exist, GRCh37 location reported by the UAS, number of observation (Obs. #), and the comprehensive nomenclature as recommended by the ISFG (Parson <i>et al.</i> 2016). It also indicates if a variant was previously observed in the Saudi population (Khubrani <i>et al.</i> 2019b) or not. Variants in black are the target iiSNPs, in blue variants that were highlighted by the UAS and in red variants that were reported by the UAS in the Flanking Region Report but were not highlighted in blue (see Table 1.6). N/A: no rs identifier were found for the correspondence variant at the dbSNP database. None of these variants was observed in the data of the Qatari population (Almohammed and Hadi 2019).	168
Table 6.6. Novel alleles observed in the population of Saudi Arabia. The table show 33 novel alleles assessed based on the SR database. Shaded alleles are novel and have not been observed in (Phillips <i>et al.</i> 2018a, Khubrani <i>et al.</i> 2019b, Almohammed and Hadi 2019) or in the GenBank. The reason of the novelty types is also shown, repeat sequence (RS) and flanking region sequence (FS).	173
Table 6.7. Motif patterns of the SE33 locus observed in the samples from the population of Saudi Arabia. A total of 66 allele sequences were within motif patterns classified by Borsuk <i>et al.</i> (2018), 53 of which, as expected, had the A0 and A1 motif patterns. Two unreported motif patterns were observed in three alleles and were classified as D4 and D5 motif IDs. Rows in red indicates novel motifs observed in the Saudi population and shaded rows indicates novel alleles that were not reported before in the GenBank database (Alsafiah <i>et al.</i> 2019b).	178

Table 6.8. Perfect association between the target iiSNP and variant in the flanking region observed in Khubrani *et al.* (2019b) but not in this study. In Khubrani *et al.* (2019b), perfect association was observed between rs279844 and rs279845 but was not observed in this study due to the presence of the allele TT at rs279844_rs279845 (shaded). Black colour indicates the target iiSNPs and the blue colour indicates variants within the flanking region. SNPs that showed perfect association are underlined. 182

Table 6.9. Associations between five SNP pairs observed in the Saudi population and in five major populations (African, Ad Mixed American, East Asian, European and South Asian) generated by the 1000 Genomes Project (Phase 3) using LDlink v3.7.2. (Machiela and Chanock 2015). (A) is for pairs rs6955448- rs6950322, (B) rs430046-rs409820, (C) rs409820-rs430044, (D) rs430046-rs430044, (E) rs4606077-rs1869434, (F) rs445251-rs369438, (G) rs279844-rs279845 (all populations) and (H) is for rs279844-rs279845 (Africans). Each table shows the haplotypes frequencies across 5008 samples per all population (A-G)/African population (H), D' (an indicator of allelic segregation for two genetic variants. A D' value of 0 presents no linkage of alleles and a D' value of 1 indicates at least one expected haplotype combination is not observed), R^2 value, Chi-sq. and p-value (High chi-square statistics and low p-values are evidence that haplotype counts deviate from expected values and suggest linkage disequilibrium may be present). Each table shows a statement for the correlation between the variants of interest and if ($R^2 > 0.1$), the variants are correlated. 184

Table 7.1. LR medians for eight scenarios simulated using seven different markers sets for related and unrelated simulations. The table shows the improvement on LRs when more markers were used for the tested relationships. It also shows the case pedigrees (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members (i.e. genotyped members) and crossed member were assumed as not available for testing. As expected, LRs improved when more loci were added. The improvement varied and was impacted by the type relationship tested and by the number of relatives included in the simulation. 202

Table 7.2. The results of the LD test for 14 syntenic STR-STR pairs (at the same arm) resulted from using SureID 23 kit in conjunction with GlobalFiler (12 syntenic pairs, P value = 0.004) or with Fusion 6C (14 syntenic pairs, P value = 0.0035) and their RF values. The RFs were calculated using Kosambi mapping function using genetic map distance in cM estimated using cumulative genetic map distance in cM which were reviewed from (Phillips 2017). None of the syntenic pairs showed LD after Bonferroni correction. The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/12$ STRs = 0.004 and $0.05/14$ = 0.0035. Shaded rows show all syntenic pairs with RFs < 0.12. Cautions should be considered when including D18S51-D18S1364 and PentaD-D21S2055 pairs in the calculation of LRs due to low RFs. The pair vWA-D12S391 will not have significant impact for most pedigrees as RF is ~ 0.12 (Gill *et al.* 2012). 226

Table 7.3. The results of the LD test for 166 syntenic (STR-STR, STR-SNP and SNP-SNP) pairs (at the same arm) resulted from using ForenSeq DNA Signature Prep kit alone and the RF values. The RFs were calculated using Kosambi mapping function using genetic map distance in cM estimated using high-density multi-point SNP data of HapMap as described by Phillips <i>et al.</i> (2012). The cumulative genetic map distance in cM of 27 aSTRs were reviewed from (Phillips 2017) and of the 94 iiSNPs were estimated as described by Phillips <i>et al.</i> (2012) (Appendix 6, Section 10.6.2). None of the syntenic pairs showed LD after Bonferroni correction (P value = 0.0003). The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/166$ pairs = 0.0003. Shaded rows present pairs with RFs < 0.12 (43 pairs). This table assumed that SE33 was typed as shown in Chapter 6. The data of the 87 samples were used in the test of LD.	227
Table 7.4. The results of the LD test for additional 50 syntenic (STR-STR and STR-SNP) pairs (at the same arm) resulted from combining GlobalFiler, SureID23 and ForenSeq DNA Signature Prep kits. The cumulative genetic map distance in cM of 12 STRs were reviewed from (Phillips 2017) and of D16S539 with the 94 iiSNPs were estimated as described by Phillips <i>et al.</i> (2012) (Appendix 6, Section 10.6.2). The RFs were calculated by Kosambi mapping function using genetic map distance in cM that was estimated using high-density multi-point SNP data of HapMap as described by Phillips <i>et al.</i> (2012). None of the syntenic pairs showed LD after Bonferroni correction (P value = 0.00023). The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/216$ pairs (166 pairs from Table 7.3 and 50 from this table) = 0.00023. Shaded rows present pairs with RFs < 0.12) (49 pairs in total when using the 136 loci).	232
Table 8.1. The order of 42 aSTRs studied in this project based on their MP. The table also shoes the CMP that can be obtained when using any of the latest four developed CE-based aSTR kits.	241
Table 10.1. Scenarios specific equations that used for calculating the RI. The table shows the equations that are used in calculating RI based on the genotypes of the tested individuals. The numerator (X) represents the probability of that the alleged father has passed the common allele with the disputed child. The denominator (Y) represent the probability of that random male from the same population is the source of shared allele. The RI equations are the result of numerator (X)/denominator (Y). The last five rows, show specific scenarios' equations that used for calculating the RI when the mother's genotype is not available (Stephenson 2010).	270
Table 10.2. Scenarios specific equations that used for calculating the RI when the child is missing, and the genotypes of the parent are is available (AABB 2010b)	271
Table 10.3. Scenarios specific equations that used for calculating the Sibling Index (SI) and the Half-Sibling Index (HSI) (AABB 2010b)	272
Table 10.4. Sequence-based data for 27 aSTRs generated from Chapter 6.	288
Table 10.5. HWE test for aSTRs for the data generated from Chapter 6. None of the analysed markers showed significant deviation from HWE after Bonferroni correction (P value>0.0004). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/121$ loci = 0.0004.	305

Table 10.6. SE33 data Generated from Chapter 6.	306
Table 10.7. iiSNPs sequence-based data generated from Chapter 6.	308
Table 10.8. HWE test for iiSNPs data generated from Chapter 6. None of the analysed markers showed significant deviation from HWE after Bonferroni correction (P value > 0.0004). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/121$ loci = 0.0004.	323
Table 10.9. LD test for 122 autosomal markers. A total of 292 pairs (STR-STR, STR-SNP and SNP-SNP) of syntenic markers (q-q, p-p, and p-q) were tested and no LD was detected after Bonferroni correction (P value > Bonferroni-corrected P value 0.0001). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/292$ pairs = 0.0001.	325
Table 10.10. The cumulative genetic map distances of 95 SNPs estimated in this study. The 95 SNPs includes 94 iiSNPs and (rs925658351) for D16S539 STR (shaded row). The cumulative genetic map distances were estimated as described by (Phillips <i>et al.</i> 2012). The SNP position (bp) on 1000 Genome Browser was used to find the approximate HAP MAP Position (bp) and then to give the cumulative genetic map distance estimation that were eventually used to calculate the RFs.	334

IV. List of figures

Figure 1.1. Number of individual's DNA profiles on the UK's NDNAD. This figure is showing the number of individuals' DNA profiles stored on the NDNAD of the UK starting from 2008 to 2016 (National Police Chief's Council 2017, 2015). 5

Figure 1.2. Match percentage when loading a DNA profile from a crime scene. This figure is showing the efficiency of the UK's NDNAD when adding a DNA profile from a crime scene (National Police Chief's Council 2017). 6

Figure 1.3. The efficiency of the NDNAD before and after applying the Protection of Freedoms Act (May 2012). This figure is showing that the efficiency of the NDNAD was not affected after applying the PoFA (May 2012) (National Police Chief's Council 2015, 2017). 7

Figure 1.4. Number of individual's DNA profiles on the NDIS. This figure is showing the total number of individual's DNA profiles (offenders and arrestee) on the NDIS in the period of 2000-2017. It can be noticed that starting from 2004, the increase trend in the individual's DNA profiles, was higher than before, which was in response to new regulations that allow NDIS to collect samples even from arrestees (James 2012), 2017 data (Federal Bureau of Investigation 2017). 8

Figure 1.5. This figure is showing the percentage of crime scene samples that match to a sample on the database in the period of 2000 -2016 (James 2012) , 2017 data (Federal Bureau of Investigation 2017). 8

Figure 1.6. An example of discordance in allele calling between two kits. The figure shows the genotype of the same sample at the SE33 locus using genRES MPX-2 (A) and genRES MPX-2sp (B) (Lederer *et al.* 2008). 10

Figure 1.7. An explanation of one of the possible reasons of discordance between different kits. This figure explains the cause of the discordance of the same sample when using genRES MPX-2 and genRES MPX-2sp (Serac, Bad Homburg). The annealing region of the 5' primer of the genRES MPX-2 includes the 60 bp deletion present in one allele. The 5' primer of the genRES MPX-2sp is closer to the repeat region and the 60 bp deletion will not be reflected in the allele size (An original figure). 10

Figure 1.8. Inheritance pattern of alleles through generations. The pedigree shows 15
three generations of a family and the portion of DNA shared between the family
members. In the parent-child relationships, each of the offspring No. 4,6,9 and 11 has
four expected genotypes (A,C),(A,D),(B,C),(B,D) each of which has 100% chance to have
one shared allele with the father (No. 2) and 100% chance to have another shared allele
with the mother (No. 1). In full sibling relationships (e.g. 4,6,9 and 11) there is 25%
chance of having zero shared allele, 50% chance of having one shared allele and 25%
chance of having two shared alleles. In half-sibling relationships (No. 12 and 13), there
is 50% chance of having one shared allele, and 50% chance to have zero shared allele
between them. There is 50% chance of having one shared allele and 50% of having zero
shared allele between uncles (No. 4,6,9 and 11) and nephews (No. 12-17). There is 50%
chance of having one shared allele and 50% of having zero shared allele between
grandchild (NO. 12-17) with any of their grand grandparent (No. 1 and 2). Finally, there
is a 25% (4/16) chance of having one shared allele and 75% (12/16) of having zero
shared allele between the first cousins 13, (14 or 15),16 and 17. The shared alleles are
termed as identical by descent (IBD) as they have originated from common ancestors.
An original figure, and the table was adopted from (Butler 2015).

Figure 1.9. Inheritance pattern of the maternal and the paternal DNA component to 16
offspring. 1) shows the maternal alleles (A, B), paternal (C, D), and the possible alleles
combinations of the offspring. Each of the maternal and the paternal allele has 50%
chance to be passed to the offspring. 2) shows a typical case of that the alleged father
cannot be excluded from being the true father of the male offspring. 3) shows a typical
exclusion case where the alleged father did not share any of his alleles with the disputed
offspring with the assumption of no mutation event is suspected (an original figure).

Figure 1.10. HW frequencies resulted from two alleles A and a and their frequencies p 20
and q respectively, where $p + q = 1$. When a population is within the expectations of
HWE, the equations of homozygous and heterozygous genotypes can be used to
estimate the genotypes frequencies. Adopted from (Brooker 2012).

Figure 1.11. The expected genotype frequencies based on the alleles frequencies in a 21
population that met the HWE expectations. It can be seen that the highest percentage
of heterozygosity can be obtained when the two alleles (assuming only two alleles can
be seen at a marker) have a frequency of 0.5 (Figure from (Butler 2015)).

Figure 1.12. The steps of PCR1 and PCR2 during the library preparation using the 33
ForenSeq™ DNA Signature Prep kit. PCR1 is for amplifying the target regions with
tagged primers (green colours). In PCR2, the tags on the amplicons are used to attach
adapters and to attach sequencing primer in the sequencing stage. The adapters
contain a part used as a unique index for each library (light red and light yellow) and a
part complimentary to primers attached to the flow cell (dark red and dark yellow) (An
original figure based on information from (Verogen 2018a)).

Figure 1.13. Cluster generation of the ForenSeq™ DNA Signature Prep kit. The figure 35
is showing a detailed process for cluster generation. By the end of this stage, millions
of the forward strands are ready for sequencing (An original figure based on
information from (Verogen 2018a)).

Figure 1.14. Reversible termination strategy used by Illumina systems. The figure shows 36 two types of termination strategies (Irreversible and reversible). The Irreversible strategy blocks the 3' by hydrogen atom. The reversible strategy caps the hydroxy group at the 3' position by a removable cap (O-azidomethyl) (an original figure adopted from (Chen *et al.* 2013)).

Figure 1.15. Sequencing by synthesis used by the illumina systems. Once the sequencing 37 primer is annealed to the forward strand, four labelled nucleotides A, G, C, T with reversible terminating groups are added simultaneously and are competing to be incorporated to the target base. The complimentary nucleotide is annealed and the TCEP is added to remove the dye, the reversible terminating groups, and to generate hydroxy group at the 3' position simultaneously. This allows second base annealing in the second read. The fluorescent of the cleaved dye is imaged and recorded (an original figure adopted from (Chen *et al.* 2013)).

Figure 1.16. The minimum requirements for STR nomenclature system. This figure 41 showing an example of the minimum requirements for STR nomenclature when using MPS systems that were recommended by the ISFG (Parson *et al.* 2016).

Figure 1.17. Saudi Arabia administrative divisions. This map is showing the 13 43 administrative provinces: Makkah, Al-Madinah, Riyadh, Eastern Province, Al-Qassim, Asir, Hail, Tabuk, Northern Borders, Jizan, Al-Baha, Al-Jouf, and Najran. It also shows the eight Arab countries (image from <https://www.123rf.com/>).

Figure 2.1. This map is showing distribution of collection sites. Collection started from 51 Riyadh to Dammam, Riyadh, Abha, Makkah (Mecca), Al-Madinah, Tabuk, and then back to Riyadh.

Figure 2.2. A hypothetical pedigree created by an in-house Excel sheet. The hypothetical 67 pedigree comprised of three generations and 13 members. Circles represent female members and squares represent male members. The profiles of the members were generated by the in-house Excel sheet based on the allele frequencies of the 136 loci.

Figure 2.3. Mutation rate settings in Familias3 software. For aSTRs, extended stepwise 69 model was used by applying the mutation rate in Table 2.4 in the rate box, 0.1 in the range box, 0.000001 in the rate 2 box for microvariants alleles. For iSNPs, equal probability was used by applying 2.5E-8 in the rate box.

Figure 2.4. An example of how the genetic distance (cM) between syntenic pairs was 74 calculated. This figure shows how the genetic distance (cM) between syntenic pairs was calculated as described by Phillips *et al.* (2012).

Figure 3.1. An example of an FTA card used for sample collection. Each card contained 78 the sample number and the donor sex (F/M). The same number was printed in the consent form. A series of 2 mm punches can be seen in the upper blood spot.

Figure 3.2. Extracted DNA run on agarose gels (1%) (A, B, and C). A) Shows results of the 79 5 samples following manufacturer's protocol. B) Shows results of the same samples after applying the 6 hours/overnight incubation in the ATL buffer before starting the manufacturer's protocol. C) Shows results of the 5 samples when using a volume of 100 µl of the AE buffer for the elution stage rather than 150 µl, in addition to the first modification.

Figure 3.3. The average concentration of each DNA samples extracted. The average of 81
the samples concentrations was 1.5 ng/μl that ranged from 0.07 - 13.5 ng/μl. Each dot
represents the average of the two reading/sample.

Figure 3.4. Internal validation of half volume reaction and 50 cm capillary/ POP6. The 82
figure shows one of the replicates of two Globalfiler™ profiles for the positive control
using the manufacturer's protocol (A) and the half volume reaction (B). Due to the use
of 50 cm capillary, the run time was increased to 3800 s which was sufficient to cover
the designated area of the largest locus SE33 that allowed the local Southern method
to be used.

Figure 3.5. Allele variants of 7 and 8 at D1S1656. The figure shows allele 7 (A) and 8 (B) 83
at the D1S1656 locus and the allelic ladder (C). Allele 7 is located outside the designated
area of the locus. Both alleles are not represented in the allelic ladder of the
Globalfiler™ PCR amplification kit. The alleles were reported in STRBase (Ruitberg *et al.*
2001) but no sequence data was available (Alsafiah *et al.* 2017).

Figure 3.6. Non-reported Allele variants at SE33. The figure shows alleles 14.3 (A), 20.3 84
(B), and 38 (C) at SE33 and the allelic ladder (D). The alleles have not been reported
before in the STRBase (Ruitberg *et al.* 2001) and are not represented in the allelic ladder
of the Globalfiler™ PCR amplification kit (Alsafiah *et al.* 2017).

Figure 3.7. Two electropherograms (A & B) for the same sample using two different STR 85
kits. (A) shows the genotype of D7S820 locus (9, 12, OL) using the GlobalFiler® PCR
amplification kit. (B) shows the genotype of the same locus (9,12) using PowerPlex® 21
System. This confirmed that the OL allele belonged to SE33 and the OL allele appeared
within the allelic window of D7S820 because of the adjacent locations of the D7S820
and SE33 in the GlobalFiler® PCR amplification kit (Alsafiah *et al.* 2017).

Figure 3.8. The OL allele at the SE33 and the allelic ladder of the GlobalFiler® PCR 86
amplification kit. The figure shows the size of the OL allele comparing to the allelic
ladder. As it had been confirmed that the OL allele belonged to the SE33 locus, it was
possible to calculate the repeat numbers based on the sizes of the OL allele and the
nearest allele in the allelic ladder (4.2); the OL was called as allele 2 (size-based call).
The black arrow points to stutter artefact of the OL allele (Alsafiah *et al.* 2017).

Figure 3.9. A multi-dimensional scale (MDS) for the average FST values of 13 common 98
loci. Fourteen populations were included in the comparison: Saudi Arabian (this study),
Saudi Arabian (Khubrani *et al.* 2019a), Saudi Arabian population in Riyadh (Osman *et al.*
2015), Saudi Arabian in Dubai (Alshamali *et al.* 2005), Saudi Arabian in Kuwait (Al-Enizi
et al. 2013), Qatari (Perez-Miranda *et al.* 2006), UAE population (Jones *et al.* 2017),
Kuwaiti (Al-Enizi *et al.* 2013), Omani-Dubai (Alshamali *et al.* 2005), Yemeni-Dubai
(Alshamali *et al.* 2005), Iraqi-Kuwait (Al-Enizi *et al.* 2013), Egyptians-Kuwait (Al-Enizi
et al. 2013), Iranian-Kuwait (Al-Enizi *et al.* 2013), and Indian-Kuwait (Al-Enizi *et al.* 2013).
Note: the data of Saudi population in (Sinha *et al.* 1999) was not included in the FST test
due to the limited number of common loci included in the study (four loci). SA: Saudi
Arabian and UAE: United Arab Emirates. The cmdscale function was used in R software
to generate a multi-dimensional scale (MDS).

Figure 4.1. A 20-cm-long 3% agarose gel for the novel SE33 alleles; from the left side, 106
alleles 20.3, 14.3, 38, 2 and a 100 bp ladder. It shows the separation of alleles 20.3 and
29.2 (35 bp) that could not be achieved with a shorter (10 cm) gel.

- Figure 4.2. An example of the quality of sequencing results. This figure shows an electropherogram for the forward strand of allele 2 at the SE33 locus. 106
- Figure 4.3. Sequencing data of the reverse strand of the alleles 7 and 8 at the D1S1656 locus. This data was generated using ForenSeq™ DNA Signature Prep (Primer Mix B) and MiSeq FGx System (Verogen). (A) Shows the sequence data of allele 7; (B) Shows the sequence data of allele 8. Due to the presence of the A variant of rs78443572 SNP (TAGG, G: 73%, A: 27%) in the alleles 8 and 13, these alleles ended with TAGA rather than TAGG (Alsafiah *et al.* 2018). 109
- Figure 4.4. An estimation of the biogeographical ancestry. The figure shows the result of the biogeographical ancestry estimation of the two samples using the piSNPs and aiSNPs included in the Primer Mix B of the ForenSeq™ DNA Signature Prep kit. Both samples were classified as ad-Mixed Americans, one of which was more like the European main population. 110
- Figure 5.1. Allelic ladder of the SureID® 23comp kit. This figure shows the allelic ladder provided with the SureID® 23 comp kit. It represents 232 alleles that are supported by 53 additional bins for variant alleles (pink bins). It also shows the successful calibration and optimisation of the ABI 3500 DNA Genetic Analyser (Alsafiah *et al.* 2019a). 120
- Figure 5.2. The genotype of the 9947A control DNA at the D5 locus included in the SureID® 23 comp kit. The locus had 14, 23, which is the genotype of D5S2800, confirming the correct name. The locus name is now updated by the manufacturer from D5S2500 (as shown in the locus name) to D5S2800. 121
- Figure 5.3. Sensitivity and stochastic tests for the SureID® 23comp kit. Serial dilutions (500, 250, 125, 62, and 31) pg were prepared from the 2800M control DNA (Promega Corporation). Each test was done in five replicates and the highest number of detected alleles are shown. Each cell represents an allele and merged cells represent homozygote loci in 2800M. Green cells identify detected alleles with $\geq 60\%$ peak balance ratios. Yellow cells identify detected alleles with $< 60\%$ peak balance ratios. Red cells represent non-detected alleles with threshold of 50 RFU/150 RFU for heterozygotes/homozygotes (Alsafiah *et al.* 2019a). 123
- Figure 5.4. Testing of the SureID® 23comp kit with tannic acid. Three different concentrations of 100 ng/ μ l, 120 ng/ μ l and 150 ng/ μ l were tested. This figure shows the results of the control sample (no inhibition) and of the 100 ng/ μ l (tannic acid) sample. Figure 5.5 shows the results of 120 and 150 ng/ μ l of tannic acid. 125
- Figure 5.5. Testing of the SureID® 23comp kit with tannic acid. Three different concentrations of 100 ng/ μ l, 120 ng/ μ l and 150 ng/ μ l were tested. This figure shows the results of the 120 ng/ μ l (tannic acid) sample and of the 150 ng/ μ l (tannic acid) sample. Full profiles were achieved with ≤ 120 ng/ μ l of tannic acid. 126
- Figure 5.6. Testing of SureID® 23comp kit with humic acid. Three different concentrations of 50 ng/ μ l, 75 ng/ μ l and 100 ng/ μ l were tested. This figure shows the results of the control sample (no inhibition) and of the 50 ng/ μ l (humic acid) sample. Figure 5.7 shows the results of 75 and 100 ng/ μ l of humic acid. 127

Figure 5.7. Testing of SureID® 23comp kit with humic acid. Three different concentrations of 50 ng/μl, 75 ng/μl and 100 ng/μl were tested. This figure shows the results of the 75 ng/μl (humic acid) sample and of the 100 ng/μl (humic acid) sample. Full profiles were achieved with ≤75 ng/μl of humic acid. 128

Figure 5.8. SureID® 23comp kit performance with two common PCR inhibitors. Full profiles were generated in the presence of 75 ng/μl of humic Acid and 120 ng/μl of tannic acid. These figures are similar to those reported for the SureID®PanGlobal (Liu *et al.* 2017). However, the kit was not as robust with inhibitors as PowerPlex® Fusion 6C, GlobalFiler™, and Investigator® 24plex (Lin *et al.* 2017) (Figure from (Alsafiah *et al.* 2019a). 129

Figure 5.9. Peak balance ratios study for the D21S2055 locus. This figure shows a study of the correlation between the size difference between heterozygous alleles and the peak balance ratio for the D21S2055 locus using data of 500 samples. The peak ratios of all genotypes that have the same size difference (nt) (e.g. the genotypes 13, 17; 14, 18; and 15, 19 have the same size difference of 4 nt) were averaged and are represented by the black dots. The blue line shows the smoothed mean of the peak ratios. Heterozygote alleles with >50 nucleotides difference showed peak ratios <45% (Alsafiah *et al.* 2019a). 134

Figure 5.10. Stutter ratios study for the SureID 23comp Kit. The figure shows the average of - 4 stutter ratios for STRs included in the SureID 23comp kit. Each box represents the stutter ratios of an STR. The x-axis represents alleles and the y-axis represents the stutter ratios. The line was drawn based on the average ratios of observed stutters. Alleles of x.1, x.2 and x.3 are plotted at x.25, x.5, and x.75 respectively. The average of stutter ratios ranged from 3.8% for D2S441 to 16.15% for D12S391 (Alsafiah *et al.* 2019a). 136

Figure 5.11. Precision study of the SureID 23comp Kit. The figure shows standard deviation (s.d) values of the fragment sizes of 22,981 alleles generated from 500 samples tested by the SureID 23comp. The highest s.d. was observed in allele 21 at D7S3048 (0.1048 nt) (Alsafiah *et al.* 2019a). In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar). 137

Figure 5.12. Accuracy study of the SureID 23comp Kit. The average of the size values of each allele in the data of the 500 samples and in 21 allelic ladders were compared to the actual sizes of the corresponding allele (actual sizes provided by the manufacturer). The size differences per nucleotides were calculated and are represented by the coloured dots. All alleles fell within the range of ±0.41 nt of the allelic window; the largest differences were seen at D6S474 allele 17 (0.4096 nt) and D7S3048 allele 26 (-0.4084 nt) (Alsafiah *et al.* 2019a). 138

- Figure 5.13. Alleles outside the windows of the allelic ladder of the SureID® 23comp kit. 141
This figure shows ten alleles observed in the population of Saudi Arabia that are not represented and were situated outside the designated window of their loci. a) Alleles 7 and 8 at D1S1656. b) Alleles 26.3 and 27.3 at D13S325. c) Allele 30 at D7S3048. d) Allele 16 at D4S2366. e) Allele 12 at D3S1744. f) Allele 10 at D6S474. g) Alleles 6 and 7 at D15S659. Allele 7 at D1S1656 (a) was situated under the designated area of D18S1364 (Alsafiah *et al.* 2019a).
- Figure 5.14. Multi-dimensional scaling for the average FST values. Five populations were 148
included in the comparison and each number represent a population, Saudi Arabia (this study), European (Iyavoo *et al.* 2019), African (Iyavoo *et al.* 2019), South Asian (Iyavoo *et al.* 2019) and Ningbo population (data provided by the Health Gene Technologies). The European and South Asian populations were more similar to Saudi population than the African and Ningbo populations. The cmdscale function was used in R software to generate a multi-dimensional scale (MDS).
- Figure 6.1. Run metric indicators of the sequencing results. The indicators of the 156
sequencing showed that the average quality of the generated reads is within the optimal ranges.
- Figure 6.2. Depth of coverage for 27 aSTRs analysed in this study. The average reads 158
count was 673 for all aSTRs that ranged from 173 reads for D5S818 to 2936 reads for TH01. In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).
- Figure 6.3. Depth of coverage for 94 iiSNPs analysed in this study. The average was 120 159
reads for all iiSNPs that ranged from 36 for rs1736442 to 1320 reads for rs1109037. In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).
- Figure 6.4. Average ACRs of 27 aSTRs. The ACRs of all aSTRs were >60% and ranged from 161
92.5% for D17S1301 to 65.5% for D22S1045.
- Figure 6.5. Average ACRs of 94 iiSNPs. All iiSNPs showed >60% ACRs except rs6955448 162
SNPs that showed an average of 40% ACR.
- Figure 6.6. Averages of stutter ratios for the 27 aSTRs. Each STR is represented by a plot 164
and the x-axis represents alleles and the y-axis represent stutter ratios. Stutter ratios ranged from 0.6% for allele 6 in TPOX to 31.4% for allele 30 in FGA. Allele variants of x.1, x.2 and x.3 were plotted as x.25, x.50 and x.75.
- Figure 6.7. The number of observed alleles by sequencing. Nineteen aSTRs presented a 165
greater number of observed alleles, 13 of which had more alleles based on the repeat region sequences (green), 8 aSTRs had more alleles based on the flanking region sequences (red) (two aSTRs had variants in both regions), and 8 aSTRs did not show difference in the number of observed alleles.

Figure 6.8. Improvements in the discrimination power of the 27 aSTRs.	170
Figure 6.9. Improvements in the discrimination power of the 94 iiSNPs.	171
Figure 6.10. Allele of interest at D19S433. A) shows the genotype of the sample using the GlobalFiler™ kit, the sequencing results using the ForenSeq™ kit, and typical sequences of the alleles 14 and 14.2 comparing to the allele of interest. B) shows the repeat region (blue) and the location of AG deletion (green) and the 5' and the 3' anchors used by the SR (yellow).	175
Figure 6.11. The number of observed size and sequence-based SE33 alleles.	176
Figure 6.12. The number of SE33 sequence variants observed per allele.	177
Figure 6.13. Sanger sequencing results for the discordance event. (A) the reference sequence of nucleotides 88277350 – 88277381 (GRCh38) at the 3' flanking region of the SE33 locus. (B) the sequence of the sample showed the discordance event. It shows a TTTT deletion at 88277355_88277358 (GRCh38) that explains the discordance between sequence and CE data.	179
Figure 7.1. A hypothetical pedigree created by an in-house Excel sheet. The hypothetical pedigree comprised of three generations and 13 members. Circles represent female members and squares represent male members (This figure is a copy of Figure 2.2).	194
Figure 7.2. A confirmation of the parent-child relationship assumed between the pedigree's members. The figure shows a screen shot from the Familias3 software for the results of the blind search (parent-child). Each parent-child relationship was validated for the 136 loci before starting the simulation tests.	194
Figure 7.3. The impact of adding more DNA markers to the simulation tests on the LR. The figure shows how testing more DNA markers improves the LR and thus reduces uncertainty. The blue line represents LR distribution for related simulations, the red line represents LR distribution for unrelated simulations, the light blue area represents the true positive (TP), the light red area represents the true negative (TN), the yellow area represents the false positive (FP), and the green area represents false negative (FN). The marker sets A, B and C are examples of different marker sets where the number of markers in set A lower than in set B, which is lower than in set C. The green and yellow areas are the uncertainty areas. When more markers are used (e.g. set B) LR distribution of related moves to the right (LR increased) and LR distribution of unrelated moves left (LR decreased). The uncertainty areas are decreased when more markers are added (e.g. set C) (\log_{10} of LR 1 = 0) (an original figure).	197
Figure 7.4. LR distributions of the simulation study for parents-child relationship (trio pedigree) using 15 aSTRs included in the Identifiler kit, which was plotted based on data generated by Familias3 software. The green histogram represents LR distributions for the true positive simulations (parents-child relationship), the orange histogram represents the LR distributions of true negative simulations (unrelated). The 15 aSTRs showed 100% TP and 0% FP up to the 100,000 LR threshold.	198

Figure 7.5. LR improvements (increment) for different relationships using the seven 200
marker sets for related simulations. The figure shows LR improvements when more loci
used and shows the impact of type of the relationship simulated and impact of including
relatives in the simulation tests, on the LRs. In the box plots, the lower whisker
represents 25% of the lowest data, the upper whisker represents 25% of the highest
data. The rectangle shows that 75% of the data are below the upper line, 25% of the
data are below the lower line, and the centre bar represents the median of the data
(50% of the data above this bar and 50% of the data below the bar).

Figure 7.6. LR improvements (decrease) for different relationships using the seven 201
marker sets for unrelated simulations. The figure shows LR improvements when more
loci were used and shows the impact of type of the relationship simulated and impact
of including relatives in the simulation tests, on the LRs. Higher impact of the 94 iiSNPs
on parent-child relationship (when used alone or when they were included in the 121
or the 136 loci) can be seen in the bottom right (will be discussed at the end of this
study). In the box plots, the lower whisker represents 25% of the lowest data, the upper
whisker represents 25% of the highest data. The rectangle shows that 75% of the data
are below the upper line, 25% of the data are below the lower line, and the centre bar
represents the median of the data (50% of the data above this bar and 50% of the data
below the bar)..

Figure 7.7. LR distributions of the simulation study for parent-child relationship (duo 204
pedigree) using different marker combinations. The figure also shows the case pedigree
(hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and
crossed member was not available for testing.

Figure 7.8. The TP and FP at different LR thresholds generated from the simulation study 205
for parent-child relationship (duo pedigree) using different marker combinations. Each
marker set is represented by a unique colour. True positive (TP) and false positive (FP).

Figure 7.9. LR distributions of the simulation study for full-siblings/unrelated (Scenario 206
1) using different marker combinations. The figure also shows the case pedigree
(hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and
crossed members were not available for testing.

Figure 7.10. The TP and FP at different LR limits generated from the simulation study 207
for full-siblings/unrelated (Scenario 1) using different marker combinations. Each
marker set represented by a unique colour. True positive (TP) and false positive (FP).

Figure 7.11. LR distributions of the simulation study for full-siblings/unrelated (Scenario 208
2) using different marker combinations. The figure also shows the case pedigree
(hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and
crossed members were not available for testing.

Figure 7.12. The TP and FP at different LR limits generated from the simulation study 209
for full-siblings/unrelated (Scenario 2) using different marker combinations. Each
marker set represented by a unique colour. True positive (TP) and false positive (FP).

Figure 7.13. LR distributions of the simulation study for full-siblings/unrelated (Scenario 210
3) using different marker combinations. The figure also shows the case pedigree
(hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and
crossed member was not available for testing.

Figure 7.14. The TP and FP at different LR limits generated from the simulation study for full-siblings/unrelated (Scenario 3) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).	211
Figure 7.15. LR distributions of the simulation study for first-cousin/unrelated (Scenario 1) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.	212
Figure 7.16. The TP and FP at different LR limits generated from the simulation study for first-cousin/unrelated (Scenario 1) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).	213
Figure 7.17. LR distributions of the simulation study for first-cousin/unrelated (Scenario 2) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.	214
Figure 7.18. The TP and FP at different LR limits generated from the simulation study for first-cousin/unrelated (Scenario 2) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).	215
Figure 7.19. LR distributions of the simulation study for half-siblings/unrelated using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members.	216
Figure 7.20. The TP and FP at different LR limits generated from the simulation study for half-siblings/unrelated different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).	217
Figure 7.21. LR distributions of the simulation study for grand-parent or child/unrelated using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing..	218
Figure 7.22. The TP and FP at different LR limits generated from the simulation study for grand-parent or child/unrelated using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).	219
Figure 10.1. Incorporating the allele-specific mutation rates in the calculation of the RI-LR. This Figure explains the third way of incorporating the allele-specific mutation rates in the calculation of the RI-LR. The allele-specific mutation (paternal and maternal) rates for are provided in (AABB 2008). $\mu_{B \rightarrow E}$: is the mutation rate of allele B to allele E, $\mu_{A \rightarrow E}$: is the mutation rate of allele A to allele E, and p_E : the frequency of the allele E (An original figure).	274
Figure 10.2. Incorporating the mutation event into the calculation of the RI-LR. This figure describes a way of including the mutation event into the calculation of the RI-LR using a fixed probability for each type of mutation that was suggested by Brenner (2018) (An original figure).	275

Figure 10.3. An example of calculating the PI, paternity probability, RMNE and PE. This figure shows a typical parentage case and shows how the strength of evidence can be estimated. In this example, the specific equation was adopted from Table 10.1. based on the genotypes of the tested individuals and the frequencies of the D1S1656 alleles were adopted from (Alsafiah <i>et al.</i> 2017). By only one locus, the PI shows that there is 1/9.6 chance random unrelated man from the same population is the biological father. The paternity probability shows that 90.6% (posterior probability) is the chance that the AF is the source of the shared allele comparing to 50% (prior probability). Based on the RMNE, the PE is 80.3% that means 80.3% of the population is excluded from being the biological father of the disputed child (an original figure).	277
Figure 10.4. An example of calculating the PI, paternity probability, RMNE and PE in a mother-less case. This figure shows a typical parentage mother-less case and shows how the strength of evidence can be quantified and estimated. In this example, the specific equation was adopted from Table 10.1. (an original figure).	278
Figure 10.5. Exceedance probability for Parent-child relationship when using seven different marker combinations.	330
Figure 10.6. Exceedance probability for full-siblings (Scenario 1) relationship when using seven different marker combinations.	330
Figure 10.7. Exceedance probability for full-siblings (Scenario 2) relationship when using seven different marker combinations.	331
Figure 10.8. Exceedance probability for full-siblings (Scenario 3) relationship when using seven different marker combinations.	331
Figure 10.9. Exceedance probability for first-cousin (Scenario 1) relationship when using seven different marker combinations.	332
Figure 10.10. Exceedance probability for first-cousin (Scenario 2) relationship when using seven different marker combinations.	332
Figure 10.11. Exceedance probability for grand parent/child relationship when using seven different marker combinations.	333
Figure 10.12. Exceedance probability for half-sibling relationship when using seven different marker combinations.	333

V. Acknowledgements

Firstly, I would like to thank my parents Mohammed and Fatemah, who have supported and encouraged me during this journey, for their prayers and their kind supportive words. Unlimited thanks to my wife Wafa and my children Fatemah, Nader, Wasan, Dana and Kadi for their understanding, patience, and their kind support.

I would like also to express my deep thanks and to show my sincere of appreciation to the Director of Studies, Dr William H. Goodwin for his guidance, advice and patience throughout the PhD project. Also, lots of thanks to my Co-Supervisor Dr Sibte Hadi and my referee Dr Arati Iyengar, and my Research Degree Tutor Professor Jaipaul Singh, for their support and advices during the project.

Many thanks to the examiners Dr Judith Smith (School of Forensic and Applied Sciences, UCLan) and Dr Penny Haddrill (University of Strathclyde) for their time reading my thesis.

Many thanks to Ms Sarah Naif and Mr Richard Kessell (Verogen, UK) for allowing sequencing the two samples during the training course and for their technical support through my NGS work. Many thanks also to Ms Ausma Bernotaite (Health Gene Technologies, China) for the technical support on the SureID kit.

I would like also to thank Professor Mark A. Jobling and Dr Jon Wetton (University of Leicester, UK) for allowing me to carry out the lab work and analysis of the samples sequenced with ForenSeq DNA Signature Prep Kit and the MiSeq FGx instrument the in Alec Jeffreys Forensic Genomics Unit.

My deep thanks to Christopher P. Phillips (Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela), for the productive discussions

either at the ISFG2019 or through emails regarding using HapMap high density SNP data in the project. Many thanks to Daniel Kling (Department of Family Genetics, Norwegian Institute of Public Health) for his technical support during using Familias3 software.

I would like to show my appreciation to my friend Dr Ibrahim Almutairi (King Abdul-Aziz City for Science and Technology, Saudi Arabia) for inspiring me through encouraging me to continue my studies after the MSc.

Last, but not least, many thanks to my colleague Pet-Paul Wepeba for his support during this project and to Hussain Alsaleh (University of Strathclyde) for his technical support in the R platform.

VI. Abbreviations

Allele Counts Ratios	ACR
Amelogenin	AMEL
American Association of Blood Banks	AABB
Analysis of molecular variance	AMOVA
Analytical Threshold	AT
Ancestry Informative SNPs	aiSNPs
Ancestry Informative STRs	aiSTRs
Applied Biosystems	AB
Autosomal STRs	aSTRs
Capillary Electrophoresis	CE
Centimorgans	cM
Combined DNA Index System	CODIS
Combined Match Probability	CMP
Combined Paternity Index	CPI
Combined Power of Discrimination	CPD
Combined Power of Exclusion	CPE
Combined Relationship Index	CRI
Degradation Index	DI
Denatured Normalised Libraries	DNL
Depth of Coverage	DoC
Di-Deoxynucleotides Tri-Phosphate	ddNTPs
Disaster Victim Identification	DVI
European Network of Forensic Science Institutes	ENFSI
European Standard Set	ESS
False Negative	FN
False Positive	FP
Federal Bureau of Investigation	FBI
ForenSeq™ Universal Analysis Software	UAS
Forensic Science Services	FSS
Genome Project Consortium	GPC
Gulf Cooperation Council	GCC
Half-Sibling Index Likelihood Ratio	HSI-LR
Hardy Weinberg Equilibrium	HWE
Human identification STRs	idSTRs
Human Sequencing Control	HSC
Identical by Descent	IBD
Identity informative SNPs	iiSNPs
Inbreeding Coefficient	F_{IS}
International Society of Forensic Genetics	ISFG
Interpretation Threshold	IT
Laboratory Information Management Systems	LIMS
Library Normalisation Beads	LNB
Likelihood Ratio	LR
Linkage Disequilibrium	LD
Massively Parallel Sequencing	MPS
Match Probability	MP
Maternity Index	MI
Minor Groove Binder	MGB

Mitochondrial-DNA	mtDNA
Multi-Dimensional Scale	MDS
National DNA Databases	NDNADs
National DNA Index System	NDIS
National Institute of Standards and Technology	NIST
Next Generation Sequencing	NGS
Nucleotides	nt
Observed Homozygosity	Ho
Paternity Index	PI
Phenotypic Informative SNPs	piSNPs
Polymorphism Information Content	PIC
Pooled Normalised Library	PNL
Posterior Probability	Po
Power of Discrimination	PoD
Power of Exclusion	PoE
Principal Component Analysis	PCA
Prior Probability	Pr
Proteinase K	PK
Protection of Freedoms Act	PoFA
Random Man Not Excluded	RMNE
Recombination Fraction	RF
Relationship Index Likelihood Ratio	RI-LR
Sample Purification Beads	SPB
Saudi DNA Data Bank	SDDB
Scientific Working Group on DNA Analysis Methods	SWGDM
Second Generation Multiplex	SGM
Security Forces Hospitals Programme	SFHP
Sequencing by Synthesis	SBS
Shrimp Alkaline Phosphatase	SAP
Short Tandem Repeat	STR
Siblings Index Likelihood Ratio	SI-LR
Single Nucleotides Polymorphisms	SNPs
Standard Deviation	s.d.
Strait Razor software	SR
Tris-2-CarboxyEthyl Phosphine	TCEP
True Positive	TP
Typical Paternity Index	TPI

1 Chapter One: Literature Review

1.1 History of using STRs for human identification

Following the discovery and characterization of short tandem repeat (STR) polymorphisms they were rapidly applied in forensic genetics (Goodwin 2015). For over 20 years, STR markers have been the standard system for forensic genetics worldwide.

Automated DNA sequencers, modified *Taq* polymerases and fluorescently labelled primers, enabled multiplexing several STRs in a single reaction. The first multiplex used by a national forensic service provider, which included four tetranucleotide-STR loci (VWA, TH01, F13A1, and FES), was developed by the UK's Forensic Science Services (FSS) (Kimpton *et al.* 1994). This was improved with a Second Generation Multiplex (SGM) that included a sex-chromosome marker (amelogenin) and six STR loci (vWA, TH01, FGA, D8S1179, D18S51, and D21S11) (Sparkes *et al.* 1996); both the quadraplex and SGM assays were produced in-house.

In response to the demand for commercial kits, Applied Biosystems (AB, USA) developed AmpFISTR Blue (D3S1358, vWA, and FGA) as the first commercial STR kit, which was followed by AmpFISTR Green (TH01, TPOX, and CSF1PO). Then, loci in both kits were combined with D13S317, D7S820, and D5S818 in developing the AmpFISTR Profiler PCR Amplification Kit (Applied Biosystems 2004). Promega Corporation (USA) have been the other commercial company pivotal in developing commercial kits and, in 2000, Promega developed the PowerPlex 1.1 that included CSF1PO, TPOX, TH01, vWA, D16S539, D7S820, D13S317, and D5S818 (Greenspoon *et al.* 2000).

The loci first selected by the FSS in the UK, Combined DNA Index System (CODIS) of the USA and the European Standard Set (ESS) have influenced the selection and the number of STR loci included in commercial STR kits.

In 1997, the Federal Bureau of Investigation (FBI) laboratory started to evaluate the data available for a number of STRs and selected 13 to make up the CODIS (Budowle *et al.* 1998). The CODIS loci are CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11. The CODIS loci could initially be genotyped in two reactions by using AmpFISTR Profiler Plus (FGA, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, and D21S11) and AmpFISTR COfiler (CSF1PO, FGA, TH01, TPOX, D3S1358, D7S820, D16S539) developed by AB. Promega Corporation developed PowerPlex 2.1 (FGA, TH01, TPOX, vWA, D3S1358, D8S1179, D18S51, and D21S11) (Levedakou *et al.* 2002), in addition to the PowerPlex 1.1, which together covered the CODIS loci. Common loci between AmpFISTR Profiler Plus and AmpFISTR COfiler (FGA, D3S1358, and D7S820), and between PowerPlex 1.1 and PowerPlex 2.1 (TH01, TPOX and vWA) were used for quality control purposes, to minimize the potential for generating chimeric profiles. However, using two reactions for the CODIS loci was not ideal for crime scene samples. Therefore, Promega Corporation and AB developed PowerPlex 16 system (Krenke *et al.* 2002) and AmpFLSTR Identifier Kit (Collins *et al.* 2004) respectively, each of which include all of the CODIS in one reaction.

The European Network of Forensic Science Institutes (ENFSI) also evaluated number of STR loci to establish their own set of markers (European Standard Set (ESS)), which would facilitate data sharing between European countries. In 1999, the ENFSI defined the ESS as VWA, TH01, FGA, D8S1179, D18S51, and D21S11 (Schneider 2009), which

were already included in the SGM, PowerPlex 2.1, PowerPlex 16 system and AmpFLSTR Identifier Kits. In 2009, the ESS was expanded by adding six loci (Schneider 2009), three of which were characterized as mini-STRs with maximum amplicon size 123 bp (D10S1248, D22S1045, and D2S441) and three others: D12S391, D1S1656, and TPOX. In 2015, seven loci were added to the CODIS loci, these were D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 and D22S1045 (Hares 2015). Based on the latest expansions in the CODIS and the ESS, the GlobalFiler kit (AB), Investigator 24plex QS (Qiagen, Germany) and PowerPlex Fusion 6C (Promega Corporation) were developed.

In the UK, AmpFISTR SGM Plus (AB) was the first commercial used kit, which targets ten STR-markers (SGM's markers plus D3S1358, D19S433, D16S539 and D2S1338) (Cotton *et al.* 2000). In 2014, this panel was expanded by adding six loci (D10S1248, D22S1045, D2S441, D1S1656, D12S391, and SE33) (Home office 2013), which can be genotyped using AmpFISTR NGM SElect Kit (AB) (Applied Biosystems 2015), PowerPlex ESI 17 Pro System (Promega Corporation) (Promega Corporation 2017), or Investigator ESSplex SE Plus (Qiagen) (Qiagen 2012b). Table 1.1 reviews the currently available STR kits provided by AB, Promega Corporation and Qiagen.

Little known about the STR loci used in China; however, AB has designed AmpFISTR Sinofiler kit (Applied Biosystems 2012) that is only available in China (Zhang *et al.* 2015). The kit includes the Chinese population specific locus of D6S1043 and other 14 STRs (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, D5S818, D13S317, D16S539, D2S1338, D19S433, vWA, D12S391, D18S51 and FGA).

Although the selection of a certain set of STR loci was based on population evaluation studies, it is also influenced by millions of DNA profiles that already exist in National DNA Databases (NDNADs).

Table 1.1. Currently available autosomal STR kit that provided by AB, Promega Corporation and Qiagen. Data from <https://www.thermofisher.com>, <https://www.promega.co.uk/>, and <https://www.qiagen.com>. (RM) rapidly mutating Y-STR.

Autosomal STR		Applied BioSystems						Promega				Qiagen								
		Identifier™	MiniFiler™	NGM™	NGM Select™	NGM Detect™	GlobalFiler™	VeriFiler™	VeriFiler™ Plus	Fusion 6C system	Fusion system	21 system	18D system	ESX 17 & ESI 17 systems	ESX 16 & ESI 16 systems	24plex QS	ESSplex SE QS	ESSplex SE Plus	ESSplex Plus	IDplex Plus
1	CSF1PO																			
2	D5S818																			
3	D7S820																			
4	D13S317																			
5	TPOX																			
6	D3S1358																			
7	D8S1179																			
8	D16S539																			
9	D18S51																			
10	D21S11																			
11	FGA																			
12	TH01																			
13	vWA																			
14	D2S1338																			
15	D19S433																			
16	D1S1656																			
17	D2S441																			
18	D10S1248																			
19	D12S391																			
20	D22S1045																			
21	SE33																			
22	D6S1043																			
23	PentaD																			
24	PentaE																			
Sex-specific markers	Amelogenin																			
	Y-indel																			
	DYS391																			
	DYS570 (RM)																			
	DYS576 (RM)																			
Internal Quality Control																				

1.2 National DNA Databases (NDNADs)

The rationale for offender DNA databases is that criminals tend to be repeat-offenders and therefore having records of DNA profiles may help to solve future crimes rapidly. By the end of 2016, the number of the Interpol member countries with a national DNA database reached to 69 countries (Interpol 2016).

The effectiveness of DNA databases depends on the legislation that governs the collection and storage of profiles along with the coverage of the population. As of 2016, for example, the number of individual's DNA profiles in the NDNAD of the UK reached 5.86 million profiles and the probability of finding a match between a sample from a crime scene and a known person in the NDNAD, was 63.3% (National Police Chief's Council 2017) (Figure 1.1 and Figure 1.2).

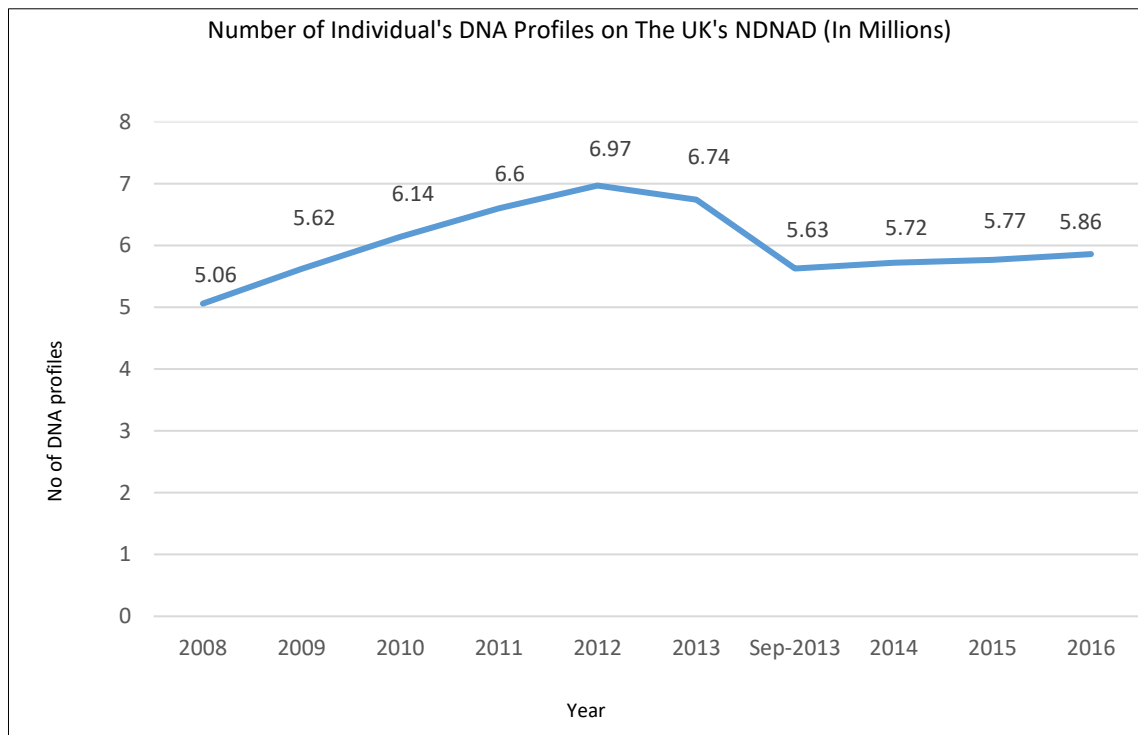


Figure 1.1. Number of individual's DNA profiles on the UK's NDNAD. This figure is showing the number of individuals' DNA profiles stored on the NDNAD of the UK starting from 2008 to 2016 (National Police Chief's Council 2017, 2015).

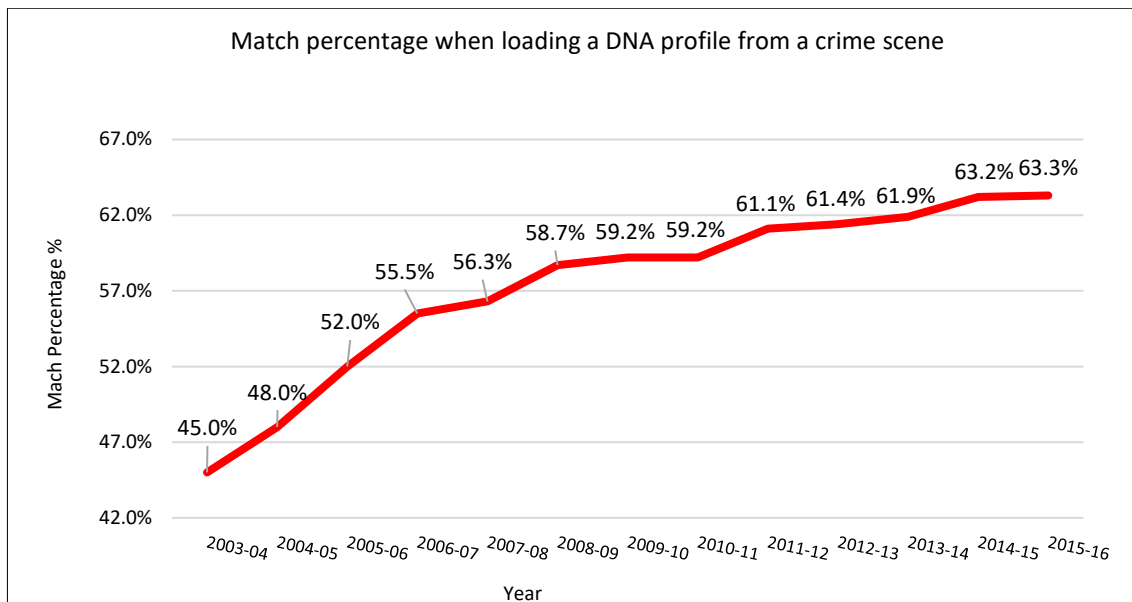


Figure 1.2. Match percentage when loading a DNA profile from a crime scene. This figure is showing the efficiency of the UK's NDNAD when adding a DNA profile from a crime scene (National Police Chief's Council 2017).

in response to the Protection of Freedoms Act (PoFA) (May 2012) (National Police Chief's Council 2017), more than 592,000 DNA profiles for innocent people have been deleted, between March 2012 to September 2013, (Figure 1.3), the efficiency of the NDNAD was not affected that illustrates that the PoFA is balanced in establishing the balance between rights of the state and individuals (Figure 1.3).

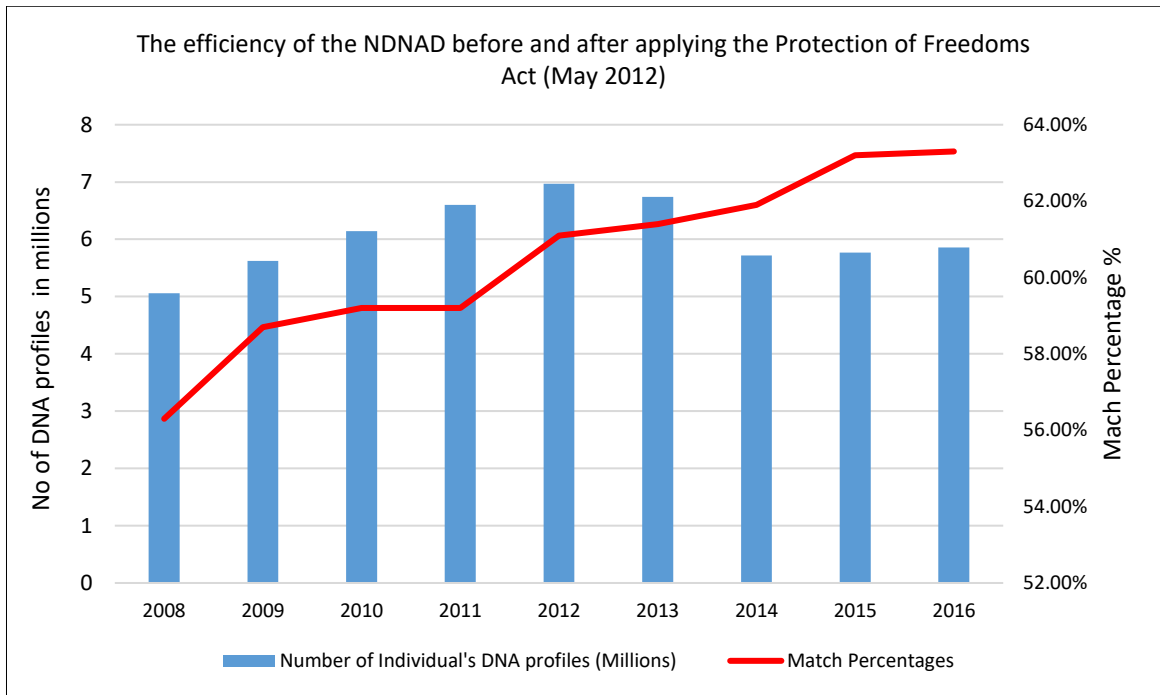


Figure 1.3. The efficiency of the NDNAD before and after applying the Protection of Freedoms Act (May 2012). This figure is showing that the efficiency of the NDNAD was not affected after applying the PoFA (May 2012) (National Police Chief's Council 2015, 2017).

Another example is the National DNA Index System (NDIS) of the USA that used CODIS loci to profile samples for the database (James 2012). Three types of samples were included in the NDIS that are 1) individuals convicted of crimes, 2) unknown human remains, and 3) samples collected from crime scenes (James 2015). Later, the NDIS was permitted to collect and to analyse samples from arrestees based on amended legislations issued in 2004, and in 2005 (James 2015). Between 2000 and up to July 2017, the total number of individual's profiles was 15,760,528; this included 2,794,862 for arrestees, and the number of matches reached 385,590 (2.44%) (Federal Bureau of Investigation 2017) (Figure 1.4 and Figure 1.5).

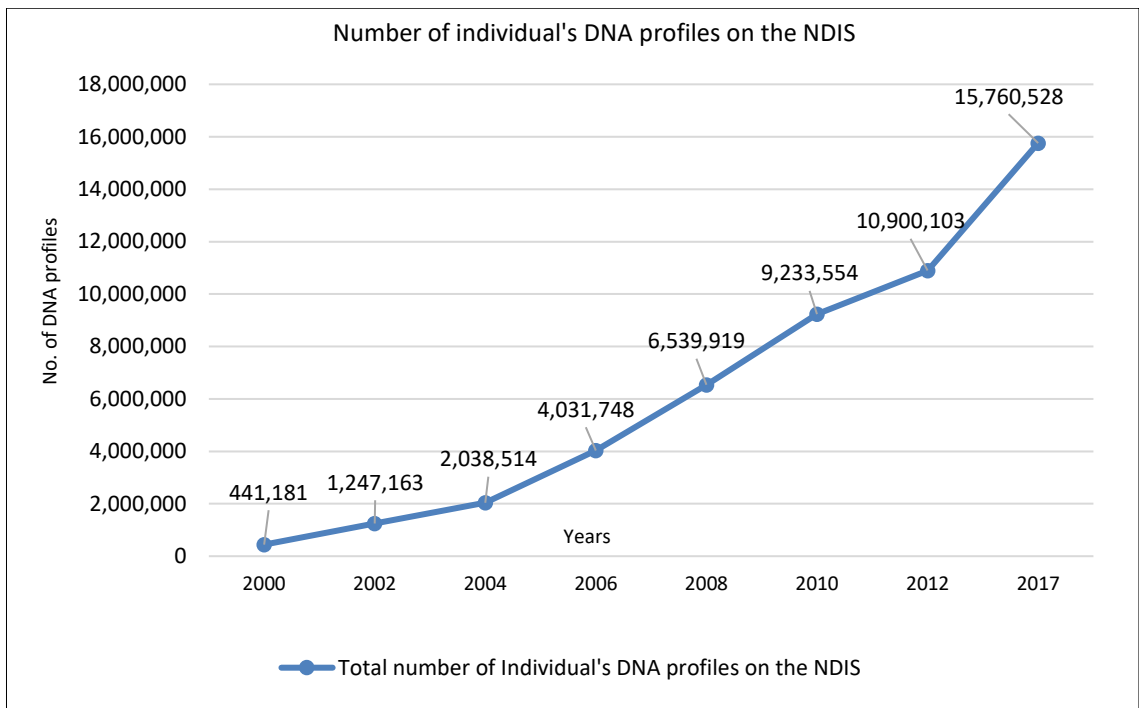


Figure 1.4. Number of individual's DNA profiles on the NDIS. This figure is showing the total number of individual's DNA profiles (offenders and arrestee) on the NDIS in the period of 2000-2017. It can be noticed that starting from 2004, the increase trend in the individual's DNA profiles, was higher than before, which was in response to new regulations that allow NDIS to collect samples even from arrestees (James 2012), 2017 data (Federal Bureau of Investigation 2017).

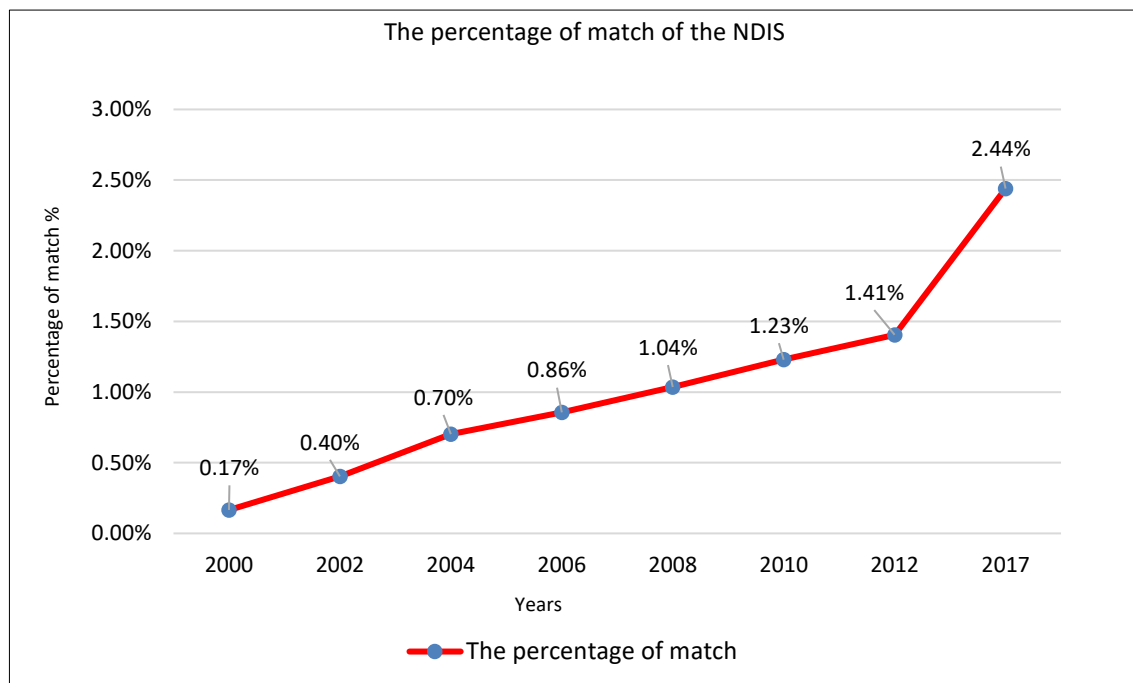


Figure 1.5. This figure is showing the percentage of crime scene samples that match to a sample on the database in the period of 2000 -2016 (James 2012) , 2017 data (Federal Bureau of Investigation 2017).

The Chinese DNA database is considered as the largest and the most rapidly growing DNA database. In 2013, the database included 20 million profiles that has led to more than 410,000 matches between known and unknown samples (Ge *et al.* 2014), and by 2017 the number had grown to 68 million DNA profiles (DNA Resources-Forensic and Policy 2019).

1.3 Capillary electrophoresis (CE) and labelled primers

The CE system is the most widely adopted system in forensic genetics laboratories. This system is based on using a primer pair for each locus, one of which is fluorescently labelled with a dye (e.g. 6-FAM, VIC, NED, TAZ, or SID). The movement of labelled amplicons through the polymer (e.g. POP4) in the capillary is based on their size (smaller size moves faster). During the movement, a CCD camera detects the fluorescence signal of the dye when it is excited by a laser and translates it to a peak on the electropherogram. When the expected sizes of two STRs amplicons overlap, they can be labelled with two different dyes to allow discrimination. Regardless of the repeat sequence, the repeat number can be determined by subtracting the flanking regions from the size of amplicons (total number of bases).

Although microvariants alleles (x.1, x.2, and x.3) can be detected by CE system, it is not possible to know the structure of the variants, for example, whether they are formed due to a deletion/insertion in the repeat region, or whether this happened in the flanking region. In addition, discordance in allele calling, between commercial kits may be observed, as they do not necessarily use the same primer pairs. For example, at SE33, a sample showed two alleles (heterozygote) using genRES MPX-2 kit (Serac, Bad Homburg, Germany) (Figure 1.6 A) while it showed one allele (homozygote) genRES MPX-2sp (Serac, Bad Homburg) (Lederer *et al.* 2008) (Figure 1.6 B). Both alleles had 23.2

repeats by sequencing; however, one of them had 60 bp deletion in the 5' flanking region, which is included by the primer pair of the genRES MPX-2 (Figure 1.7).

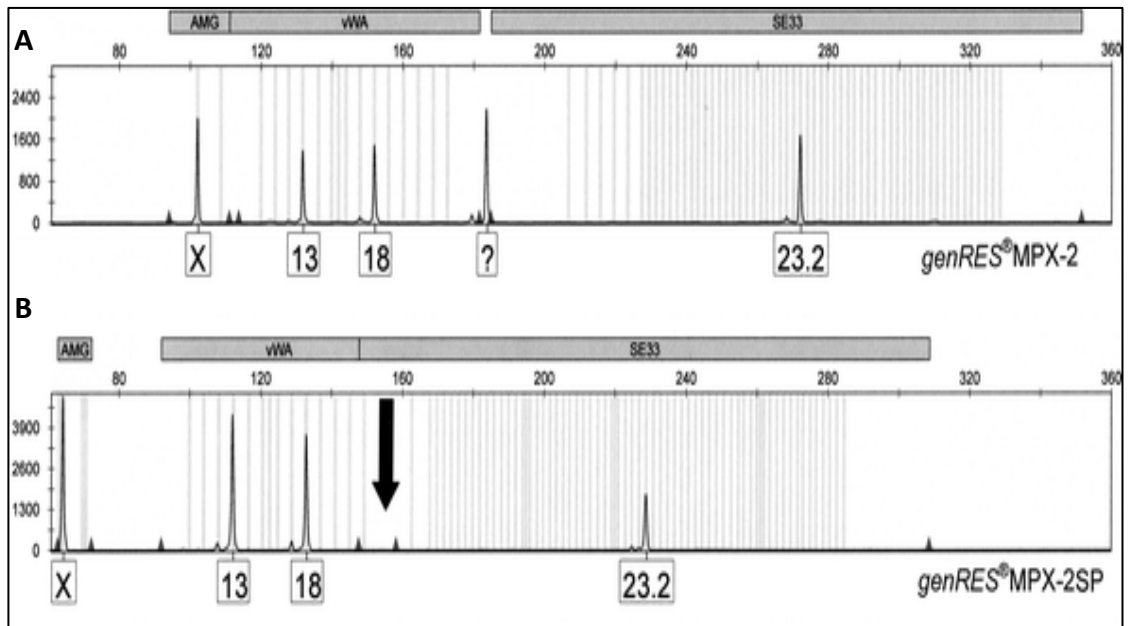


Figure 1.6. An example of discordance in allele calling between two kits. The figure shows the genotype of the same sample at the SE33 locus using genRES MPX-2 (A) and genRES MPX-2sp (B) (Lederer *et al.* 2008).

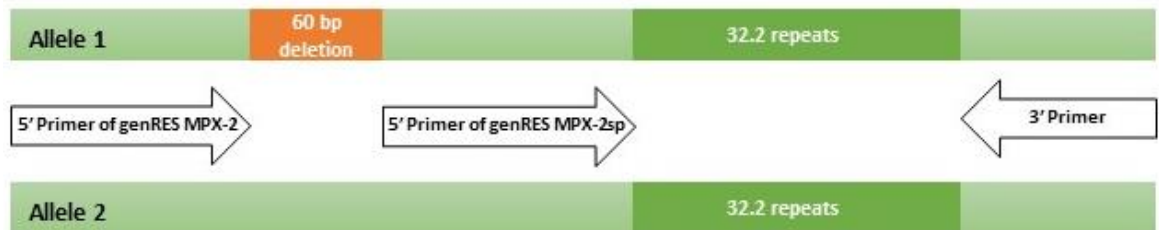


Figure 1.7. An explanation of one of the possible reasons of discordance between different kits. This figure explains the cause of the discordance of the same sample when using genRES MPX-2 and genRES MPX-2sp (Serac, Bad Homburg). The annealing region of the 5' primer of the genRES MPX-2 includes the 60 bp deletion present in one allele. The 5' primer of the genRES MPX-2sp is closer to the repeat region and the 60 bp deletion will not be reflected in the allele size (An original figure).

1.4 Internal validation of new multiplex kits

The ENFSI and the Scientific Working Group on DNA Analysis Methods (SWGAM) have developed guidelines for validation and verification of new kits for forensic applications (SWGAM 2016, ENFSI 2010). In both guidelines, the criteria include

repeatability, reproducibility, sensitivity and stochastic effect, heterozygote peak balances, stutter/corresponding allele ratios, concordance with other kits for the same STRs, performance when PCR inhibitors are present.

The repeatability and reproducibility assess the performance of a new multiplex kit when used by the same operator/instrument and by different operator/instrument, which maximises the likelihood of an identical result (DNA profile) at any time, by any operator, and using any instrument. The ENFSI guidelines have determined 5 replicates, as the minimum number, of the same sample that will be used for the repeatability and for reproducibility tests (ENFSI 2010).

The sensitivity test measures the limits of the detection of a multiplex kit using concentrations below the range defined by the manufacturer. A series of five dilutions (e.g. 500, 250, 125, 62, and 31 pg) each concentration replicated three times, was defined by the ENFSI to assess the sensitivity of a kit (ENFSI 2010). The sensitivity study can also assess the stochastic effect (allele imbalances or allele drop out) that result from low quantity/quality DNA (SWGDM 2016). In addition, it can also measure the ideal concentration of DNA/reaction that achieves higher peak balances of heterozygous genotypes.

The peak balances are expressed within three categories: intra-locus balances, intra-dye balances, and inter-dye balances. Low intra-locus balances (peak balance ratios within a locus) would increase the possibility of not detecting the second allele of heterozygote genotypes that may lead to mis-characterisation mixture samples. The intra-dye balances (peak balances ratios within one dye) is important to assess the quality of the samples and in the interpretation of mixture samples. As the performance of some loci differ from others and the level of fluorescence of dyes are not the same,

selection of a dye attached to markers combination is important that is expressed by the inter-dye balances. ENFSI has defined 60% as the minimum ratio of the peak balance between heterozygote genotypes (intra-locus) (ENFSI 2010).

PCRs are *in vitro* reactions that may lead to the creation of mis-copies observed in DNA profiles called stutter. While the most common type of stutter is a peak with one repeat smaller than the true allele, a stutter with one repeat larger can also be observed (Krenke *et al.* 2005). Studying the peak ratios of stutter height to true allele height is important in the interpretation of DNA profiles especially when multiple contributions are suspected. It was found that the number of repeat (allele size), structure complexity, and A - T content of the repeats are positively correlated to the height of the stutter peak (Brookes *et al.* 2012). The stutter ratios measured by the validation study have to be lower than ratios estimated by the manufacturer and will be ignored during the interpretation of DNA profiles (ENFSI 2010).

As mentioned above, STR multiplexes can use different primer pairs for the same loci and it is possible to observe discordance between kits at the same locus (Figure 1.6 and Figure 1.7). Therefore, previously genotyped samples can be tested to study loci concordance between different kits.

The performance of the kit when any of the common PCR inhibitors are present (stability), can also be tested. The inhibitors, which generally are either derived from the cell components or from the environment, interfere with amplification and may decrease the efficiency of amplifying the targeted DNA markers (Wilson 1997). Different concentrations of common inhibitors, like humic acid, tannic acid and collagen, are used to study the performance of kits (Lin *et al.* 2017).

The SWGDAM guidelines demand studying the precision and accuracy of the kit. The precision of a kit can be measured by identifying the size variations through all observations of each individual allele, which can be computed as a standard deviation (s.d). As the mobility of DNA fragments is affected by the attached dye, the accuracy study ensures that alleles will fit within the designated space (± 0.5 bp of the actual size) that can be carried out by measuring the differences between the size of genotyped alleles and their actual sizes (SWGDAM 2016). In general, the precision and accuracy tests show the reliability of a kit in identifying heterozygote genotypes that have a single base difference between the two alleles and show to what extent an allele can be sized outside the designated window.

For kits that include Y-specific loci (e.g. the DYS391 STR and the Yq11.221 indel in the GlobalFiler™ PCR Amplification Kit), the detection of male/female contents in artificially mixed samples with known male and female samples also needs to be evaluated if the kit is to be used for analysis of material recovered from a crime scene.

1.5 Applications of STR-based systems

1.5.1 Criminal investigation

Samples recovered from a crime scene are typed and are compared to the DNA database or to the suspect's profile. A full match links a suspect to the crime scene or links different crime scenes. However, would someone else, from the same population, have the same DNA profile? or would two unrelated individuals, from the same population, have the same DNA profile?. Although this could be answered if the whole population was tested, the option would involve some ethical and privacy implications and would also be expensive (Williamson and Duncan 2002, Jeffreys 2005). Alternatively, the match event can be statistically evaluated or quantified by estimating

the match probability (MP), that shows to what extent another unrelated individual could have the same DNA profile. The MP is calculated for independently inherited loci (see Section 1.7) by multiplying the genotype frequency generated from the allele frequencies (p^2 for homozygous and $2pq$ for heterozygous genotypes where p and q are the allele frequencies) and is based on population being in Hardy Weinberg equilibrium (HWE) (see Section 1.6)

While the MP decreases with the increased number of tested STRs and with the number of heterozygous genotypes, it increases in case of partial DNA profile (e.g. due to DNA degradation) and when the tested individual is related to the perpetrator or from the same sub-population (Jobling and Gill 2004).

1.5.2 Kinship testing

The number of relationship (kinship) tests reported by American Association of Blood Banks (AABB) has risen from 77,000 in 1988 to 371,719 in 2013. In 2013, almost 900,000 samples were tested by 19 AABB accredited laboratories to identify the father of a disputed child (AABB 2013, AABB 2010a). Unlike the criminal investigation of crime scene samples, kinship testing looks at inherited alleles within tested individuals where the inheritance pattern of the alleles, within relatives, varies based on the type of relationship (Figure 1.8).

The level of uncertainty is higher in kinship testing compared to the matching of crime scene samples due to the variations in the inheritance pattern between relationships and due to the possibility of the presence of the shared allele in an unrelated individual (Butler 2015). When kinship testing suggests an individual cannot be excluded from a claimed relationship due to allele sharing, the event can then be quantified to assess the strength of the evidence by calculating the relationship index likelihood ratio (RI-LR).

This can be done by using the equation: $RI-LR = X/Y$, where X is the probability of the genotypes when tested individuals are related as claimed and Y is the probability of the genotypes when a random individual from the same population has the shared allele (Allen 2013).

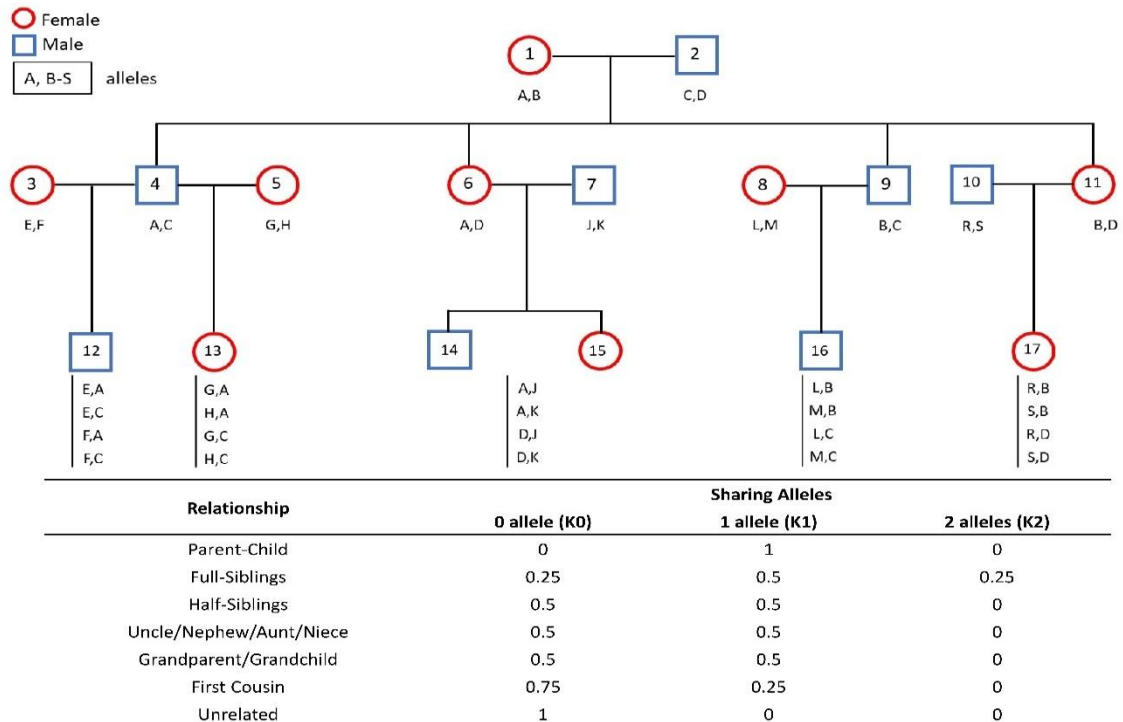


Figure 1.8. Inheritance pattern of alleles through generations. The pedigree shows three generations of a family and the portion of DNA shared between the family members. In the parent-child relationships, each of the offspring No. 4,6,9 and 11 has four expected genotypes (A,C),(A,D),(B,C),(B,D) each of which has 100% chance to have one shared allele with the father (No. 2) and 100% chance to have another shared allele with the mother (No. 1). In full sibling relationships (e.g. 4,6,9 and 11) there is 25% chance of having zero shared allele, 50% chance of having one shared allele and 25% chance of having two shared alleles. In half-sibling relationships (No. 12 and 13), there is 50% chance of having one shared allele, and 50% chance to have zero shared allele between them. There is 50% chance of having one shared allele and 50% of having zero shared allele between uncles (No. 4,6,9 and 11) and nephews (No. 12-17). There is 50% chance of having one shared allele and 50% of having zero shared allele between grandchild (NO. 12-17) with any of their grand grandparent (No. 1 and 2). Finally, there is a 25% (4/16) chance of having one shared allele and 75% (12/16) of having zero shared allele between the first cousins 13, (14 or 15),16 and 17. The shared alleles are termed as identical by descent (IBD) as they have originated from common ancestors. An original figure, and the table was adopted from (Butler 2015).

The simplest type of kinship testing is parentage testing that typically aims to identify the true father (paternal testing) by looking for shared alleles between the alleged father and the disputed child (Figure 1.9).

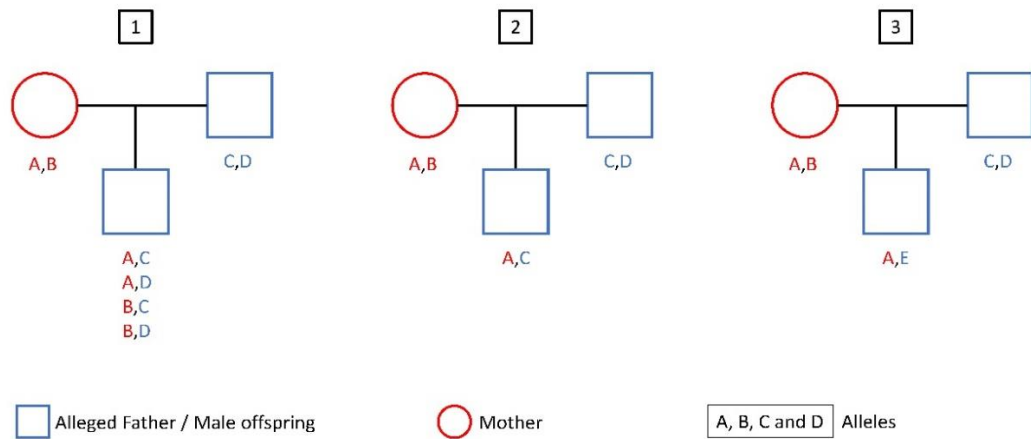


Figure 1.9. Inheritance pattern of the maternal and the paternal DNA component to offspring. 1) shows the maternal alleles (A, B), paternal (C, D), and the possible alleles combinations of the offspring. Each of the maternal and the paternal allele has 50% chance to be passed to the offspring. 2) shows a typical case of that the alleged father cannot be excluded from being the true father of the male offspring. 3) shows a typical exclusion case where the alleged father did not share any of his alleles with the disputed offspring with the assumption of no mutation event is suspected (an original figure).

In parentage testing, the calculation of paternity index (PI)/maternity index (MI) depends on the genotypes of the tested individuals (homozygote or heterozygote) by which specific equations are applied to calculate each of the probabilities (i.e. X and Y) (Stephenson 2010) (Appendix 1, Section 10.1.1, Table 10.1). When the child is missing (e.g. disaster victim identification (DVI)) or is needed to be identified (e.g. immigration cases), the case may involve reverse parentage analysis using the genotypes of the parent. In such cases, the PI-LR or MI-LR can be calculated using specific equations based on the genotypes of tested individuals too (homozygote or heterozygote) that are shown in (Appendix 1, Section 10.1.2, Table 10.2).

Kinship testing may also involve identifying distant relatives like sibling, half-sibling, grandparent/grandchildren, uncle/nephew, aunt /niece or first cousin. Table 10.3

(Appendix 1, Section 10.1.3) shows the scenarios specific equations that are used to calculate the sibling index-LR (SI-LR) and the half-sibling index (HSI-LR) based on the genotypes of tested individuals. Due to the complication of calculating the LR for grandparent/grandchildren, uncle/nephew, aunt/niece or first cousin relationships, the AABB recommends using validated kinship software to do these complex calculations (AABB 2010b).

STRs have shown relatively higher mutation rates (an average of $1.001E-3$ /locus/generation) for tetra-nucleotides STRs and the paternal origin of the mutations were estimated to be 3.3 times more those originated from the maternal side (Sun *et al.* 2012). This is a problematic in kinship testing as it is believed that it is possible to observe two inconsistencies with the true father and to observe same number of inconsistencies with a random man (not the true father) when using ~ 12 STRs in parentage testing (Brenner 2018). Therefore, the AABB emphasizes incorporating the mutational event in calculation of the LR (AABB 2008). Several ways are used to integrate the mutation event to the calculation that are described in detail in (Appendix 1, Section 10.1.4).

Once the RI-LR (i. e. PI, MI, SI or HSI) is calculated for each individual locus, the combined relationship index (CRI) can then be calculated by multiplying the RI-LRs for all independent loci. Although the RI-LR takes in account the probability of allele sharing showed in Figure 1.8, allele frequency and possible mutation events, it does not include non-genetic evidence (i.e., the prior probability). The prior probability (Pr) assesses the strength of non-genetic evidence before incorporating the DNA test data. In general, the prior probability of 0.5 is used for most kinship cases unless the court has assigned a different probability (Allen 2013). In case of missing person, the AABB (AABB 2010b)

emphasises the use of $1/N$ prior probability, where N is the number of missing people taking in account the number of males and female (e.g. mass graves).

Sections 10.1.5 and 10.1.6 (Appendix 1) show how the prior probability (Pr) and the genetic evidence are included in the calculation of the posterior probability (Po) (i. e. relationship probability) and how The RMNE (random man not excluded) is calculated, respectively.

In general, the accreditation standards of the AABB for kinship testing (9th edition) defined 100 as a threshold of combined paternity index (CPI) by which the evidence achieves acceptable level of certainty (Allen 2013). The 100 CPI means that there is 99 to 1 chance that the alleged father is the true father and it generates paternity probabilities of 91.7% at ($Pr = 0.10$) and of 99.89% at ($Pr = 0.90$). In Germany, new guidelines have defined 15 STRs as the minimum number of tested STRs and 99.999% as a threshold for exclusion probability (i. e. $CPI \geq 100,000$) to be accepted in the court (Poetsch *et al.* 2013). In addition, the guidelines necessitate testing additional STRs (16-20 STRs) in case of deficient pedigrees to allow the exclusion probability threshold (Poetsch *et al.* 2013).

In complex kinship cases, however, testing around 20 STRs may lead to inconclusive results especially when identifying distant relatives (Carboni *et al.* 2014). It has been demonstrated that additional STRs can increase the certainty of genetic testing in determining the true relation among parent-child, sibling, half sibling (O'Connor *et al.* 2010), and distant relatives (Carboni *et al.* 2014).

1.5.3 Ancestry testing

The adopted STRs for human identification (idSTRs) are not expected to be suitable for ancestry inference (Phillips *et al.* 2013), as they have been selected that are similarly

diverse in different populations. Therefore, attempts have been carried out to select ancestry informative STRs (aiSTRs) and to use them to infer the ancestry of a DNA profile (Rosenberg *et al.* 2002, Londin *et al.* 2010, Phillips *et al.* 2013, Rosenberg *et al.* 2003).

Rosenberg *et al.* (2002) have used 377 STRs and were successful in differentiating more than 1000 individuals from 52 populations to six major groups (Africans, Eurasians, East-Asians, Oceanians and Americans). However, the large number of STRs here is not suitable for forensic applications. Another attempt by using a set of 36 aiSTRs (33 di-nucleotides, 2 tetra-nucleotides and 1 tri-nucleotides) was successfully able to distinguish Africans, East-Asians, Oceanians, Americans and Caucasians (Londin *et al.* 2010). Although di-nucleotide STRs are more informative than tetra-nucleotides STRs for ancestry inference (Rosenberg *et al.* 2003), the high stutter products make them less suitable for forensic applications, especially when analysing mixed samples (Phillips *et al.* 2013). As recent as 2013, a set of 12 aiSTRs were selected to be a complementary set for the idSTRs (Phillips *et al.* 2013), by which this set alone, the success rate (correct assignment) ranged from 51.86% for Africans to 96.82 for Europeans. When combining this set with 20 idSTRs (32 STRs in total), the success rate has improved that ranged from 81.73% for Africans to 100% for Oceanians. Despite that STRs are multi-allelic markers and are more informative than binary markers (single nucleotides polymorphisms (SNPs)), the 32 STRs performance was still less efficient in comparison to 34 aiSNPs developed by Phillips *et al.* (2007). When analysing the same samples using the 34 aiSNPs, the success rates was higher and ranged from 92% (Oceanians) to 100% (Europeans).

Unlike SNPs, STR markers appear to be of limited use for ancestry prediction due to the rarity of finding population-specific alleles (Phillips *et al.* 2014), which might be a

consequence of having higher mutation rate comparing to SNPs ($\sim 2.5E-8$ (Nachman and Crowell 2000)).

1.6 Hardy Weinberg equilibrium (HWE)

The HWE law states that when a population is within the expectation of HW, the allele and genotypes frequencies are constant through generations. Thus, HW equation can be used to calculate the genotypes frequencies based on allele frequencies (Brooker 2012) (Figure 1.10).

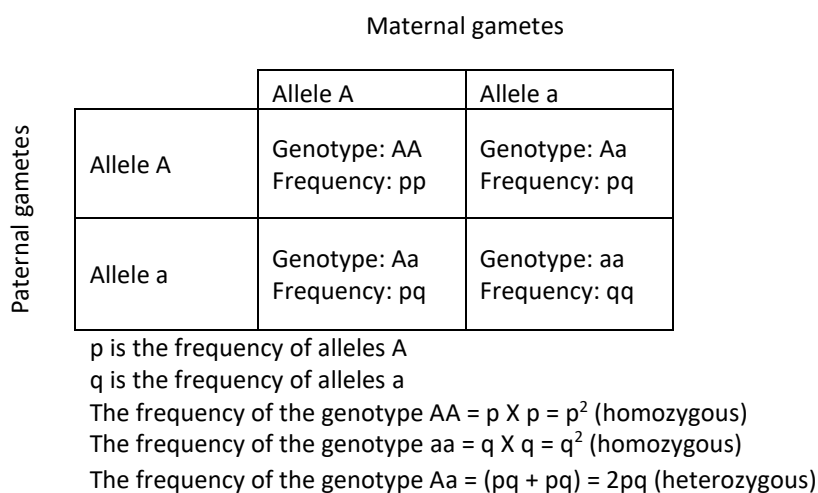


Figure 1.10. HW frequencies resulted from two alleles A and a and their frequencies p and q respectively, where $p + q = 1$. When a population is within the expectations of HWE, the equations of homozygous and heterozygous genotypes can be used to estimate the genotypes frequencies. Figure from (Brooker 2012).

There are five possible factors that may disturb HWE in a population:

- 1- Mutations that may introduce new alleles leading to a change in the allele frequencies.
- 2- Non-random mating that means that population members are mating based on specific genotypes or phenotypes.
- 3- Neutral selection that prevents members with specific genotypes from reproducing.

- 4- Migration that introduces new alleles from different population.
- 5- Genetic drift: the population is small that changes allele frequencies due sampling effect.

The HWE law predicts the genotypes frequencies based on the allele frequencies if none of the five factors disrupted HWE (Figure 1.11).

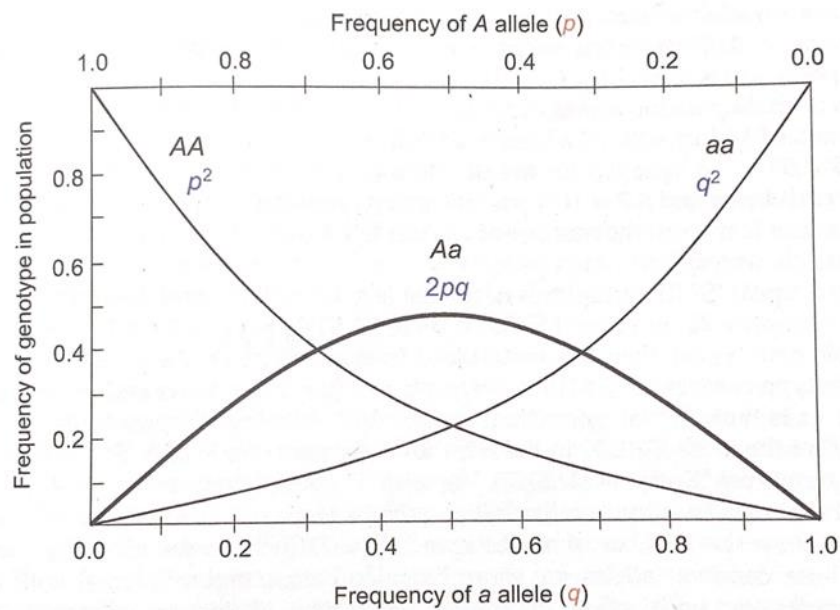


Figure 1.11. The expected genotype frequencies based on the alleles frequencies in a population that met the HWE expectations. It can be seen that the highest percentage of heterozygosity can be obtained when the two alleles (assuming only two alleles can be seen at a marker) have a frequency of 0.5. Figure from (Butler 2015).

Deviation from HWE in a data set can be tested by comparing the observed heterozygosity (H_o , the number of heterozygous genotypes divided by the total number of genotypes) to the expected heterozygosity (H_e , the expected number of heterozygous genotypes based on the allele frequency that can be predicted as in Figure 1.11).

1.7 Linkage disequilibrium (LD)

LD represents a non-random association of alleles at two or more markers within a population leading to that certain genotypes are more likely to be observed than others (genotypes show higher frequency more than expected based on their alleles frequencies) (Edge *et al.* 2017), which can even be observed between different types of markers (i. e. between STRs and SNPs) (Payseur *et al.* 2008).

LD can be caused by the genetic linkage between two closely located markers or as a result of other population genetic effects like genetic drift, neutral selection or population subdivision.

The genetic linkage between two closely located markers (syntenic loci) may influence the inheritance pattern of their alleles, by which having one allele from a marker influence the second allele from the other marker. Here, this association can be disturbed by the presence of recombinational hot spot between any two linked but not associated markers or by several mutational events through generations (Carothers and Wright 1992). Syntenic markers are regarded as independent (unlinked) if they are 50 centimorgans (cM) or more apart (at which point the probability of recombination between them is 0.5) (Lobo and Shaw 2008).

It is believed that unlinked markers (markers in different chromosomes or those are 50 cM or more apart), which showed LD, return to equilibrium faster (in fewer generations) than in those that are linked (Bright *et al.* 2014).

1.8 Common forensic statistical parameters

1.8.1 Match probability (MP) and power of discrimination (PoD)

MP is a measure that indicates the weight of a match (e.g. between a DNA profile from a crime scene and a DNA profile for a suspect). It shows the probability of that an unrelated person from the same population who could have the same genotype at a locus (National Research Council 1996). On the other hand, PoD presents the probability that two unrelated individuals from the same population would have different DNA genotype at a locus, which can be calculated by $1 - MP$.

1.8.2 Power of exclusion (PoE)

PE of a locus is defined by the probability of that an individual has a different genotype from a randomly selected individual in a paternity case, which can be calculated by $PE = h^2(1-2hH^2)$, where h is the heterozygosity and H is the homozygosity of the locus (Huston 1998).

1.8.3 Polymorphic information content (PIC)

PIC is described by Serrote *et al.* (2020) as an indicator of the ability of a marker to detect polymorphisms among individuals in a population, that was found to be impacted by the number of observed alleles and their distribution (frequency). Loci with > 0.5 PIC are recommended to genetic studies while those below 0.25 are not recommended (Serrote *et al.* 2020).

1.8.4 Analysis of molecular variance (AMOVA)

The AMOVA analysis is a common method to estimate F -statistics that includes the inbreeding coefficient (F_{IS}) and the total genetic variance (F_{ST}). The F_{IS} is the probability that a person has two identical alleles received from one ancestor; a high F_{IS} implies a higher level of inbreeding. The F_{ST} is the proportion of the total genetic variance

contained in a population. F_{ST} ranges from 0 to 1, where 1 demonstrates high level of differentiation between populations. The typical F_{ST} value among human populations was estimated to be <0.1 (Mathieson and McVean 2012).

Multidimensional scaling (MDS) is a common way to visualise the distances or the similarity between populations. The F_{ST} values generated by AMOVA analysis, are used to map the location of each population among others, and that are more similar appear closer together on the graph than populations that are less similar.

1.9 Limitations of STR-CE systems

The majority of STRs adopted in forensic applications, are tetranucleotide markers and the target regions are relatively longer (e.g. the sizes in the Identifiler kit are from 100 bp to 450 bp (Collins *et al.* 2004)). This increases the probability of degradation and can often lead to partial DNA profiles when processing material recovered from scenes of crimes. Some STRs were characterised as mini-STRs, where the primer can be designed to anneal close to the repeat region allowing shorter amplicons (50-150 bp) (Hill *et al.* 2008). The sensitivity was improved, when using this feature in the MiniFiler kit (amplicon sizes are 71-250 bp) (Mulero *et al.* 2008), by six-fold compared to the Identifiler kit; but a minimum quantity of 0.3 ng (300 pg) is still recommended to obtain a full profile (Luce *et al.* 2009). However, due to the limited number of labelling dyes (5-6 dyes), this feature is limited when using CE systems as labelled primers must be designed in a way that prevents overlap between adjacent loci labelled with the same dye and thus limits the number of STRs that can be genotyped with shorter amplicons.

1.10 Single Nucleotide polymorphisms (SNPs)

Building on earlier projects, such as the SNP Consortium, the 1000 Genome Project Consortium (GPC), has studied, in the final phase, 2500 individuals from 26 populations and has reported more than 88 million variants, almost half of which were newly reported by the project (Auton *et al.* 2015). This huge number of SNPs allows more selectivity on defining which SNPs suitable for forensic applications.

SNPs are a powerful tool for individual identification (identity informative SNPs (iiSNPs)) and paternity testing for two main reasons: Firstly, SNPs have a lower mutation rate of $\sim 2.5 \times 10^{-8}$ (Nachman and Crowell 2000). Secondly, SNPs are a single-base substitutions, which enable PCR-primers to anneal close to the target polymorphic nucleotides in many cases, and therefore can be designed to generate very short amplicons (e.g. 65-115 bp) (Børsting *et al.* 2012).

SNPs can also infer the bio-geographical ancestry by testing SNPs that show specific allele frequency variations (ancestry informative SNPs (aiSNPs)) for a population and by testing SNP variants related to specific appearance traits (hair, skin and eyes colours) (phenotypic informative SNPs (piSNPs)).

Y-STRs and mitochondrial-DNA (mtDNA) haplotypes exhibit higher differentiation between ancestries because they are not disrupted by recombination and are preserved through generations (if no mutation has occurred). However, the limited database coverage of some populations has reduced the benefit of using these markers for ancestry inference and may lead to misinterpretation of DNA evidences (Phillips 2015). For example, the presence of an African-specific Y-haplotype has been reported in a native Northern-England branches (King *et al.* 2007), showing the risk of misinterpretation that may occur. In addition, uniparental markers need relatively larger

sample sets to establish adequate estimations of allele frequencies for a population in comparison to autosomal markers (Phillips 2015). Therefore, inferring the ancestry was focusing on autosomal SNPs and thus many panels have been developed as summarized in Table 1.2.

Table 1.2. A list of the developed aiSNPs panels. This table summarizes a list of recently developed aiSNPs that have been adopted in forensic laboratories (an original table based on information from (Fondevila *et al.* 2013, Gettings *et al.* 2014, Rogalla *et al.* 2014).

aiSNPs Number	SBE Reaction	Targeted Population	Accuracy
34 aiSNPs (Fondevila <i>et al.</i> 2013)*	1 Reaction	Europeans, Africans, Americans, East Asians, and Oceanians	Europeans: 99.37%, Africans: 100%, Americans: 100%, East Asians: 94.71%, and Oceanians: 100%
50 aiSNPs (Gettings <i>et al.</i> 2014)	3 Reactions	US populations: (Africans Americans, East Asians, European Americans, and Hispanic Americans/ Native Americans.	98% Accuracy 61% Eye-colour
14 aiSNPs (Rogalla <i>et al.</i> 2014)	2 Reactions	Europeans, Africans, and East Asians.	This panel could differentiate the three major populations with 100% accuracy, but it has categorised Middle Eastern as Europeans.

*This panel is a revised panel of SNP panel published in (Phillips *et al.* 2007), the rs727811 was replaced with more informative SNP rs3827760

In 2016, a new global assay tests 31 of the most informative aiSNPs in a single SNaPshot reaction was developed to differentiate between five populations (Europeans, Africans, Native Americans, East Asians, and Oceanians), which was examined by known-ancestry control samples and could infer the population group correctly (De La Puente *et al.* 2016).

In general, the procedure of selecting SNP markers was influenced by the number of the characterised SNP markers in the human genome, and the availability of population data. To select global iiSNP markers five conditions were applied: 1) high heterozygosity in each population tested (≥ 0.4 heterozygosity), 2) almost the same allele frequencies between populations ($F_{st} \leq 0.06$), 3) feasibility of multiplexing, 4) located in non-coding region, and 5) unlinked SNPs. Having non-coding SNPs would help to avoid political and

legal considerations, and non-linked SNPs would allow higher discrimination even when relatives are suspected (Kidd *et al.* 2006, Sanchez *et al.* 2006, Pakstis *et al.* 2007, Pakstis *et al.* 2008, Pakstis *et al.* 2010, Wang *et al.* 2016, Lou *et al.* 2011) (Table 1.3).

Table 1.3. iiSNPs panels developed for human identification. The table shows the iiSNPs panels developed for human identification and the progress in match probability (MP) from 2006 – 2016. This table summarizes effort by research groups to select informative iiSNPs that can be applied globally leading to the 54 SNPs with 1.3E-22 match probability (MP).

Publication	Selected iiSNPs	MP	Conditions
(Kidd <i>et al.</i> 2006)	19 SNPs	1E-6 to E-7	<ul style="list-style-type: none"> – ≥ 0.4 heterozygosity. – low differences between all studied populations ($F_{st} \leq 0.06$). – feasibility of multiplexing. – non-coding region. – unlinked SNPs.
(Sanchez <i>et al.</i> 2006)	52 SNPs	5E-19	
(Pakstis <i>et al.</i> 2007)	40 SNPs	1E-14 to E-17	
(Lou <i>et al.</i> 2011)	44 SNPs	1E-19	
(Wang <i>et al.</i> 2016)	54 SNPs	1.3E-22	

On the other hand, nominating aiSNPs have only two conditions: 1) aiSNPs should express the differences between the target populations (higher allele frequency variations), 2) should be non-linked (De La Puente *et al.* 2016).

However, most SNPs are binary markers and less informative than STRs and thus increases the number of SNPs that would meet the power of STRs. It was found that 44 SNPs are required to meet the power of 15-16 STRs (~ 3 SNPs = 1 STR) (Amorim and Pereira 2005).

Phillips *et al.* (2015) reported 41 newly characterised tetra-allelic SNPs which have four possible alleles, 24 SNPs were found to be useful as iiSNPs. Having four possible alleles rather than two alleles increases the informativeness for each SNP, and therefore lowers the number of SNPs required to meet the power of STRs. In addition, those tetra-SNPs have been studied in three populations (Europeans, Africans, and East Asians), and the possibility of finding at least one heterozygote locus in a tetra-SNP-profile is 99.93%

in Europeans, 99.9% in Africans, and almost 93% in East Asians. This elevated heterozygosity allows better recognition of mixture samples that could not be achieved by bi-allelic SNPs (Phillips *et al.* 2015) (Table 1.4).

Table 1.4: the maximum possible heterozygosity calculation based on maximum allele frequencies of SNP types.

SNPs type	Maximum allele frequencies				Maximum possible heterozygosity
Bi-allelic SNPs	0.5		0.5		0.5
Tri-allelic SNPs	0.33	0.33		0.34	0.67
Tetra-allelic SNPs	0.25	0.25	0.25	0.25	0.75

A set of 19 multi-allelic SNPs (11 tetra-allelic SNPs and 8 tri-allelic SNPs) has provided 6.07×10^{-11} MP for the Chinese Han population (Gao *et al.* 2018) while the same number of bi-allelic-SNPs provided an average of 1×10^{-7} MP (Kidd *et al.* 2006).

Two genotyping methods have been commonly used in the forensic field for characterising SNPs: TaqMan[®] Real-Time PCR assay (Pakstis *et al.* 2010), and SNaPshot assay (Wang *et al.* 2016). Both assays have the advantage of using infrastructure that already established in most forensic laboratories. The TaqMan[®] assay is a reliable method, utilizing the minor groove binder (MGB) at the 3' terminal which increases the specificity of relatively shorter probes and increases the sensitivity of the probe even with 1 bp mismatch (Kutyavin *et al.* 2000). MGB reduces the effect of background noises comparing with probes with no-MGB (Kutyavin *et al.* 2000), allowing easier interpretation. However, only 1 di-allelic SNP (two possible alleles) can be genotyped in each reaction (Applied Biosystems 2014) (cannot be multiplexed) and two reactions for multi-allelic SNPs (a maximum of two alleles per reaction), that does not suit forensic applications especially with low quantities of DNA (Gao *et al.* 2018).

SNaPshot is a high throughput assay that can genotype, for example, 52 SNPs in only two reactions (29 SNPs and 23 SNPs) (Sanchez *et al.* 2006). Furthermore, SNaPshot assay has been applied in some laboratories which have been accredited the ISO 17025 (Børsting *et al.* 2009). In addition, the 52 SNPs multiplex has been validated against low-template DNA (Lt-DNA) and shows high sensitivity with only 50-100 pg of DNA (Børsting *et al.* 2013). However, some considerations have been raised regarding the allele calling as each labelled ddNTP for the same polymorphism is separated in a different channel (Phillips 2012). To illustrate, a person who has G as allele 1, and C as allele 2, and the ddGTP labelled with 6FAM dye, the ddCTP labelled with VIC dye, each allele will be in different channel and then it is difficult to be interpreted with another 23 SNPs. Although, guidelines have developed for electropherogram interpretation based on the ratio of peaks high for heterozygote alleles and the ratio of peak high to background noise for homozygote alleles, these guidelines fail when there are mixtures as most SNPs are binary polymorphism (Børsting *et al.* 2009).

1.11 Massively Parallel Sequencing (MPS)

MPS (also called Next Generation Sequencing (NGS)) systems have larger capabilities than conventional sequencing (Sanger sequencing). MPS systems can be used for sequencing the whole DNA “shotgun sequencing” where the DNA is fragmented to be as short as 50-500 bp prior the sequencing, or for sequencing specific regions on the DNA (targeted sequencing). Although the whole genome sequencing allows a huge amount of data to be gathered from a sample that may be useful, it does not suit forensic applications as micrograms of DNA are needed (Børsting and Morling 2015). In addition, the huge amount of data needs an extensive analysis and may eventually produce non-concordant DNA profiles for the same sample when using different MPS

platforms (Ratan *et al.* 2013). Therefore, all forensic kits that are commercially available, are based on targeted sequencing method, where the regions of interest are amplified, which reduces the amount of DNA template needed and avoids the problem of characterising coding regions.

The developed systems simultaneous test many markers. In addition, MPS facilitates the combination of different types of markers with detailed sequences. Three STR-MPS kits are available: ForenSeq™ DNA Signature Prep (Verogen), PowerSeq™ Auto/Mito/Y system (Promega Corporation) and Precision ID GlobalFiler™ NGS STR (AB) (Table 1.5). Illumina provides the MiSeq FGx, a bench-top instrument, that includes a data analysis software (ForenSeq™ Universal Analysis Software (UAS)) for the ForenSeq™ and the PowerSeq™ kits.

The rest of this section provides details of the ForenSeq™ DNA Signature Prep kit with the MiSeq FGx platform for forensic application, as they were used in this project. Verogen provides the ForenSeq™ DNA Signature Prep kit with two primer mixes A and B, and the purpose of the application defines which one will be used. The Primer Mix A targets 27 autosomal-STRs (aSTRs), 24 Y-STRs, 7 X-STRs, and 94 iSNPs, while the Primer Mix B includes additional 78 SNPs (56 aiSNPs and 22 piSNPs) (Table 1.5).

Table 1.5. DNA markers included in three STR-MPS kits commercially available (Faith and Scheible 2016, Applied Biosystems 2017, Verogen 2018a).

Locus	Amplicon length range (bp)		
	ForenSeq™ DNA signature (Primer Mixs A&B)	PowerSeq™ Auto/Mito/Y	GlobalFiler™ NGS STR
D1S1677	-	-	151-191
D1S1656	141-189	161-208	167-215
D2S441	144-180	158-204	163-195
TPOX	85-145	196-244	167-199
D2S1776	-	-	163-195
D2S1338	114-182	197-269	133-197
D3S4529	-	-	167-195
D3S1358	138-186	192-240	129-177
D4S2408	93-117	-	167-191
FGA	150-306	176-268	137-299
D5S2800	-	-	171-211
D5S818	102-150	191-239	141-173
CSF1PO	85-129	185-229	143-183
D6S1043	163-227	-	163-227
D6S474	-	-	158-186
D7S820	135-179	211-255	130-166
D8S1179	86-138	203-255	151-199
D9S1122	108-140	-	-
D10S1248	128-172	135-179	155-199
TH01	100-148	220-264	129-173
D12S391	237-281	202-254	149-193
vWA	132-192	202-262	147-207
D12ATA63	-	-	126-146
D13S317	138-186	209-257	149-181
D14S1434	-	-	163-195
PentaE	362-467	179-284	168-273
D16S539	132-180	198-253	139-179
D17S1301	114-142	-	-
D18S51	140-227	190-277	156-232
D19S433	154-212	193-253	155-195
D20S482	125-165	-	-
D21S11	158-276	203-273	179-245
PentaD	209-293	192-266	139-204
D22S1045	193-229	129-176	178-211
DYS393	-	294-256	-
DYS505	154-194	-	-
DYS456	-	141-165	-
DYS570	162-214	157-217	-
DYS576	183-235	155-203	-
DYS522	294-334	-	-
DYS458	-	171-199	-
DYS481	102-129	139-184	-
DYS19	261-345	168-294	-
DYS391	123-167	147-178	-
DYS635	214-306	155-179	-
DYS437	178-210	181-197	-
DYS439	199-239	204-224	-
DYS389I	231-275	258-294	-
DYS389II	255-299	-	-
DYS438	144-169	202-242	-
DYS390	242-286	204-248	-
DYS643	115-215	150-210	-
DYS533	198-258	242-284	-
GATA-H4	151-203	231-251	-
DYS612	215-248	-	-
DYS385 a	-	202-303	-
DYS385 b	316-354	-	-
DYS460	356-380	-	-
DYS549	214-262	189-230	-
DYS392	346-358	143-164	-
DYS448	288-324	213-255	-
DYF387S1a	-	-	-
DYF387S1b	123-255	-	-
DXS10074	211-309	-	-
DXS10103	161-185	-	-
DXS10135	228-334	-	-
DXS7132	176-208	-	-
DXS7423	147-215	-	-
DXS8378	430-462	-	-
HPRTB	193-237	-	-
94 iiSNPs	63-170	-	-
56 aiSNPs*	73-227	-	-
22 piSNPs*	67-200	-	-
Mito	-	10 amplicons (cover the mitochondrial control region)	-

* Primer mix B includes the 56 aiSNPs and the 22 piSNPs in addition to those markers in primer mix A.

Processing samples using the ForenSeq™ DNA Signature Prep kit with the MiSeq FGx platform includes four main steps: library preparation, cluster generation, sequencing and data analysis. Here, the word “library” means a combination of DNA fragments for the target regions that were prepared for sequencing for an MPS system.

1.11.1 Library preparation

This stage contains two PCR steps, PCR1 is to amplify the target regions using a tagged primer pair for each locus; each strand is tagged with a different sequence (Figure 1.12). The tags have specific sequences of nucleotides that are used as complementary to adapters that are added in the PCR2 and to a sequencing primer (in sequencing stage).

In PCR2, a combination of one i5 adapter (8 indexed adapters) and one i7 adapter (12 indexed adapters) are added to the tagged fragments. The adapters comprised two parts: indices that are used as unique identifiers for samples and complementary sequences to those primers attached to the flow cell (Figure 1.12). As each sample will have a unique identifier (an index), this enables pooling up to 96 samples for sequencing in a single run.

Then, libraries are purified by removing excess tagged primers, dNTPs, adapters and unamplified DNA using sample purification beads (SPB). The purified libraries are then normalised using library normalisation beads (LNB), to approximate all the sample’s concentration, that allows equal amounts of DNA be used in the downstream steps. Finally, the libraries are pooled into one tube for denaturation.

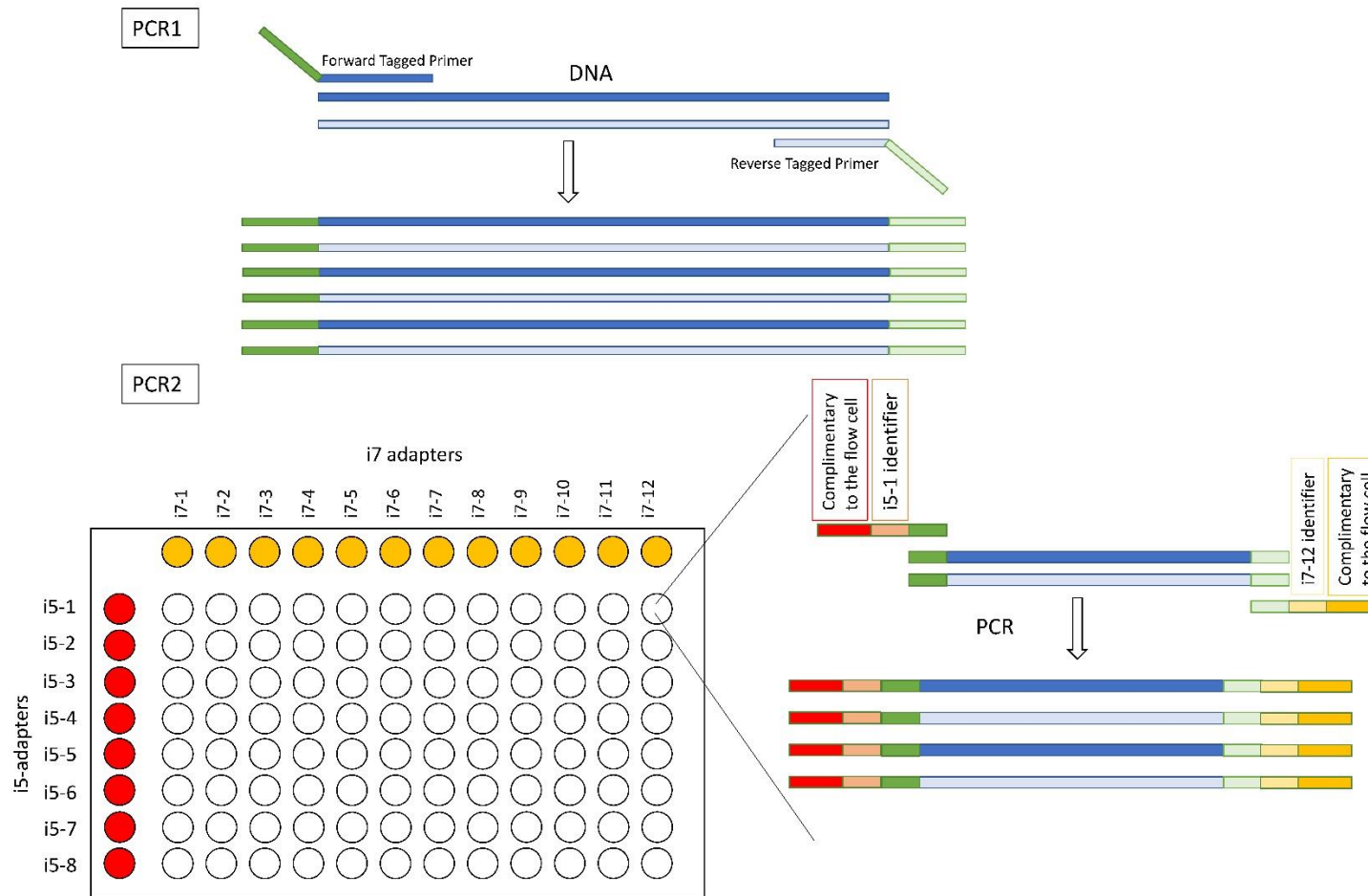


Figure 1.12. The steps of PCR1 and PCR2 during the library preparation using the ForenSeq™ DNA Signature Prep kit. PCR1 is for amplifying the target regions with tagged primers (green colours). In PCR2, the tags on the amplicons are used to attach adapters and to attach sequencing primer in the sequencing stage. The adapters contain a part used as a unique index for each library (light red and light yellow) and a part complimentary to primers attached to the flow cell (dark red and dark yellow) (An original figure based on information from (Verogen 2018a)).

1.11.2 Cluster generation

In this stage, the pooled libraries are denatured and applied into the flow cell that has two types of attached primers, where only one of them is enabled for hybridisation. When the fragments are hybridised, the DNA polymerase create a complimentary strand for each fragment and the original fragment is then removed by washing. The other primer (attached to the cell) is enabled allowing the other end of complimentary strands to anneal, which promotes bridge amplification. This process is repeated (cycles) to create millions of copies attached to the flow cell. At the end of this stage, all reverse strands are removed and only forward strands that are sequenced (Bentley *et al.* 2008) (Figure 1.13).

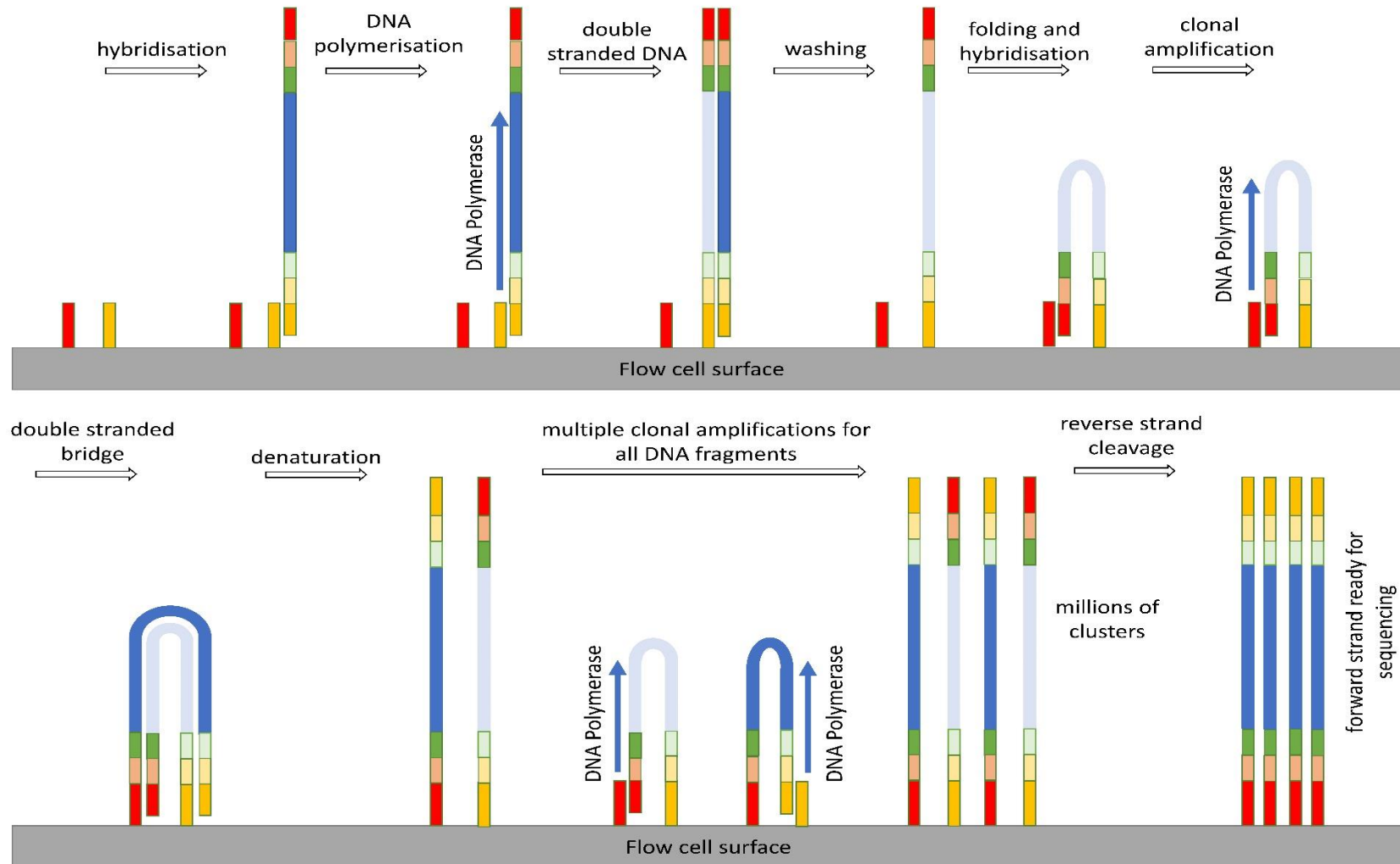


Figure 1.13. Cluster generation of the ForenSeq™ DNA Signature Prep kit. The figure is showing a detailed process for cluster generation. By the end of this stage, millions of the forward strands are ready for sequencing (An original figure based on information from (Verogen 2018a)).

1.11.3 Sequencing

Illumina systems utilise sequencing by synthesis (SBS) that uses reversible termination strategy (Bentley *et al.* 2008). The polymerase stops adding nucleotides if the 3' is terminated (i. e. di-deoxynucleotides tri-phosphate (ddNTPs) that are being used in conventional sequencing and SBE) (Figure 1.14).

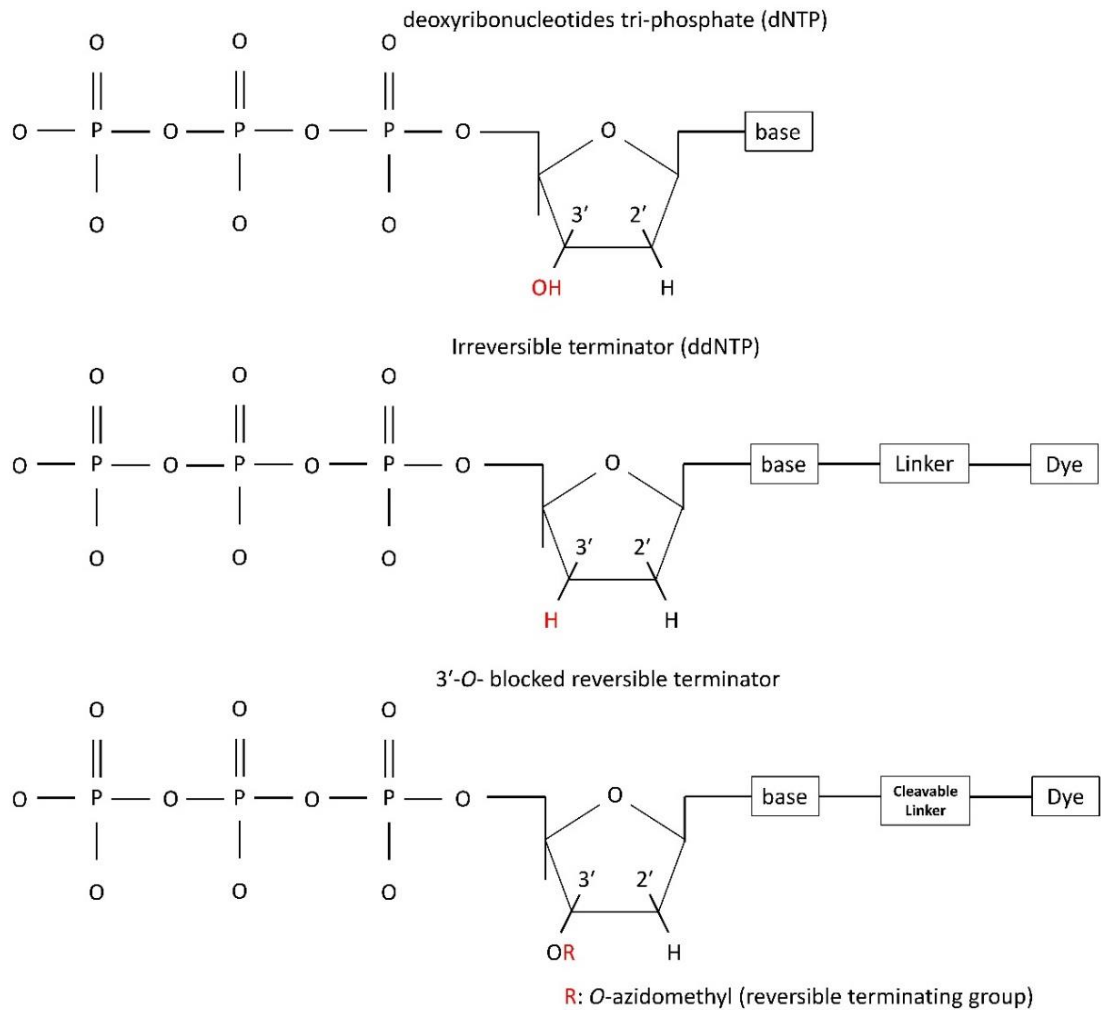


Figure 1.14. Reversible termination strategy used by Illumina systems. The figure shows two types of termination strategies (Irreversible and reversible). The Irreversible strategy blocks the 3' by hydrogen atom. The reversible strategy caps the hydroxy group at the 3' position by a removable cap (O-azidomethyl) (an original figure adopted from (Chen *et al.* 2013).

The illumina systems employ the *O*-azidomethyl at the 3' position as a reversible terminating group that can be cleaved to allow annealing the next base (Figure 1.14) (Bentley *et al.* 2008). In the flow cell, a universal sequencing primer is annealed to the

forward strands (attached to the flow cell) and the four labelled nucleotides (A, G, C, T with reversible terminating groups) are added simultaneously and are competing to be incorporated to the target base. Once the first base is incorporated, tris-2-carboxyethyl phosphine (TCEP) is added to simultaneously remove the dye, the reversible terminating groups, and to generate hydroxy group at the 3' position (Bentley *et al.* 2008). The dye fluorescence is then imaged and the 3' position allows the next base to be attached in the second cycle. The illumina systems perform multiple cycles of sequencing and real time imaging for each cluster and generate the data automatically (Chen *et al.* 2013) (Figure 1.15).

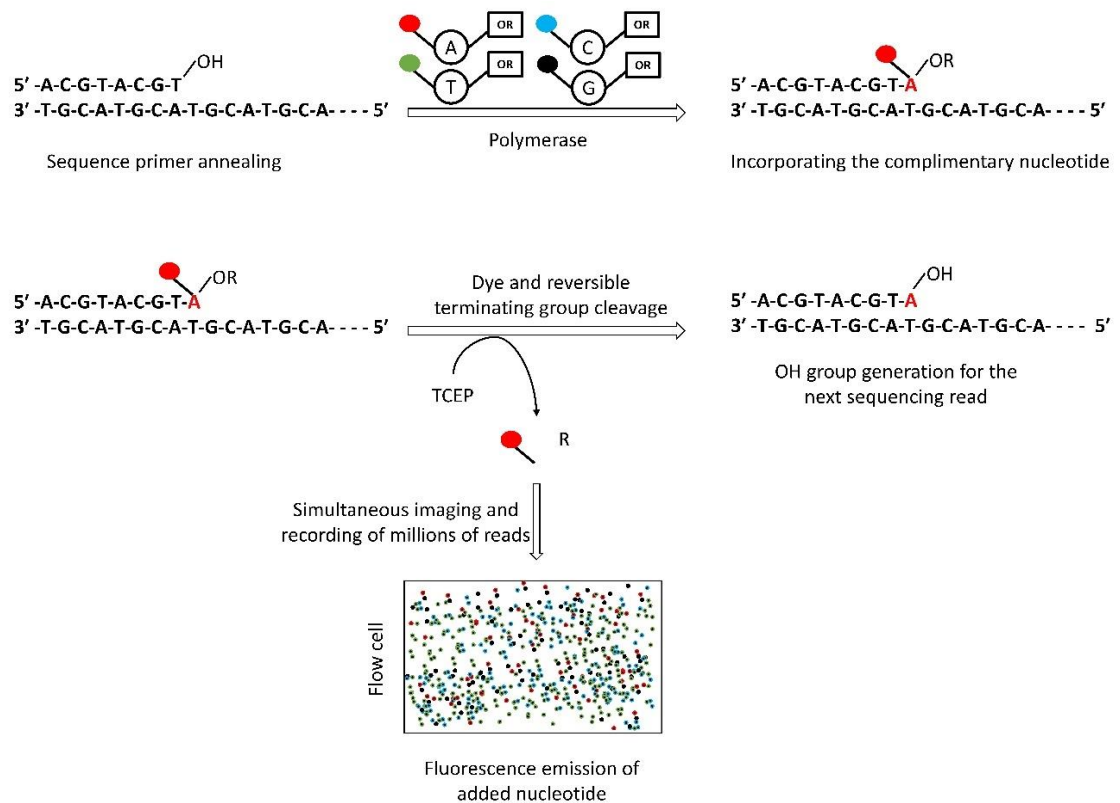


Figure 1.15. Sequencing by synthesis used by the illumina systems. Once the sequencing primer is annealed to the forward strand, four labelled nucleotides A, G, C, T with reversible terminating groups are added simultaneously and are competing to be incorporated to the target base. The complimentary nucleotide is annealed and the TCEP is added to remove the dye, the reversible terminating groups, and to generate hydroxy group at the 3' position simultaneously. This allows second base annealing in the second read. The fluorescent of the cleaved dye is imaged and recorded (an original figure adopted from (Chen *et al.* 2013)).

All reports are generated based on the setting of the analytical and interpretation thresholds (AT and IT respectively) and the stutter filters.

1.11.5 Advantages and disadvantages of MPS systems

MPS systems reveal more information about variations in the repeat region of STRs and the flanking regions than STR-CE systems do. For example, allele 20 at D2S1338 locus could have 6 different sequences that will be designated as allele 20 using the STR-CE systems (Gettings *et al.* 2016) (Table 1.7). This has led to the fact that some individuals may show homozygous genotypes using the size-based systems while they have heterozygote genotypes when using sequence-based systems (Gelardi *et al.* 2014).

Table 1.7. Different sequences of allele 20 at D2S1338 locus using MPS technologies. STR-CE systems distinguish alleles by their sizes (Gettings *et al.* 2016).

The Size-based allele call	Sequence
20	[TGCC]6[TTCC]14
	[TGCC]7[TCCC][TTCC]12
	[TGCC]7[TTCC]10[GTCC][TTCC]2
	[TGCC]7[TTCC]13
	[TGCC]7[TTCC]2[TTTC][TTCC]10
	[TGCC]8[TTCC]12

Moreover, STR loci can be even more variable when SNPs in the flanking regions are associated in the statistics (Table 1.8).

Table 1.8. Comparison of the power of discrimination of four loci by using STR-CE systems, MPS systems for variants in the repeat region and for variants in both repeat and flanking regions. This study was conducted to examine STR loci variations for the Koreans population (Kim *et al.* 2017).

Locus	Size-based systems	Power of Discrimination (PoD)	
		Include variations in repeat region	Include variation in repeat and flanking regions
D2S441	0.900	0.930	0.931
D7S820	0.904	0.904	0.947
D13S317	0.928	0.930	0.956
D21S11	0.927	0.983	0.983

Although that by looking to the repeat region variants, some loci would gain more alleles, some do not gain any additional alleles. A study of sequence variants (within repeat region) of 22 autosomal STRs using MPS in 183 volunteers from the three most common populations in the USA, found that, for example, D2S1338 locus gained 233.3% more alleles by sequencing (Table 1.9).

Table 1.9 : Comparison between the number of alleles obtained by size-based systems and by MPS systems. This table showing data that compares the number of alleles obtained by size-based systems and by MPS systems for 23 of the most common used autosomal STRs (Gettings *et al.* 2016), SE33 data (Gettings *et al.* 2015).

Locus	Alleles obtained by length	Alleles obtained by sequence	Difference
D2S1338	12	40	233.3%
D12S391	17	53	211.8%
SE33	50	152	204.0%
D21S11	19	46	142.1%
D3S1358	8	19	137.5%
vWA	8	19	137.5%
D8S1179	10	22	120.0%
D1S1656	14	23	64.3%
D2S441	9	14	55.6%
CSF1PO	8	10	25.0%
D5S818	9	11	22.2%
PentaE	16	19	18.8%
FGA	16	19	18.8%
D18S51	18	21	16.7%
D19S433	14	16	14.3%
D10S1248	9	10	11.1%
PentaD	14	14	0.0%
D22S1045	11	11	0.0%
D13S317	8	8	0.0%
D7S820	7	7	0.0%
D16S539	7	7	0.0%
TPOX	7	7	0.0%
TH01	6	6	0.0%

In addition, the MPS systems allow the amplification of STRs with shorter amplicons as they are not separated based on their sizes. A study for the ForenSeq™ DNA Signature Prep kit has shown that 0.1 ng of DNA template was enough to profile more than 98% of DNA markers included in Primer Mix A (Xavier and Parson 2017). However, some STRs, where the flanking regions are highly repetitive, and the amplicon sizes cannot be constricted (e.g. SE33).

The deep information about the repeat structure and about variants in the flanking region encouraged the International Society of Forensic Genetics (ISFG) to increase the STR nomenclature minimum requirements for MPS systems. The ISFG recommended including a description of the reference of the genome assembly sequence, locus name and allele name for the CE, version of the human genome assembly, STR region, description of the repeat region, and the location of flanking region variants (Parson *et al.* 2016) (Figure 1.16).

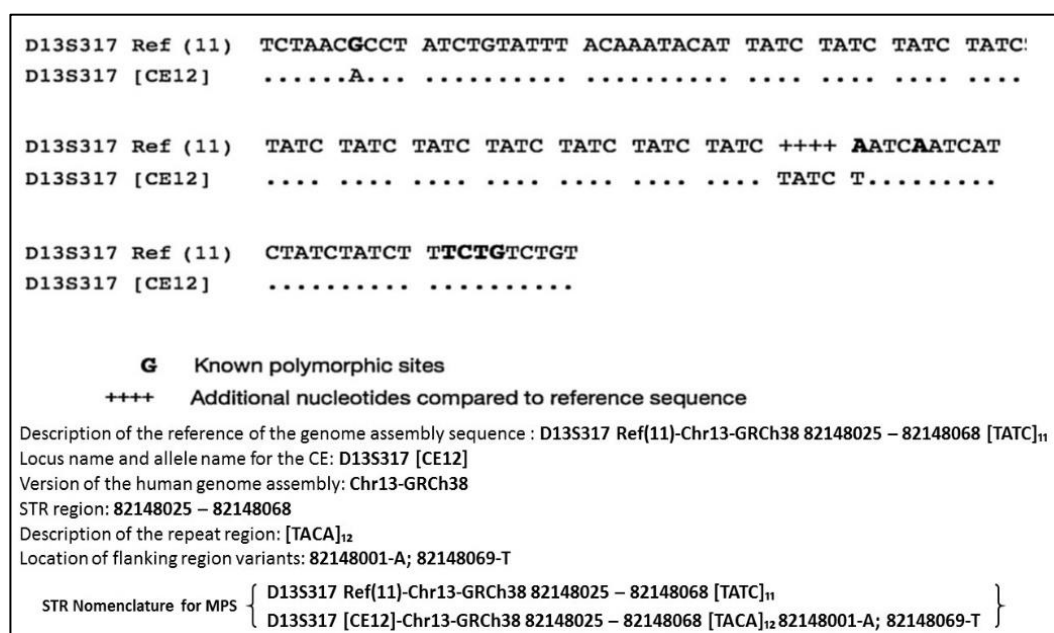


Figure 1.16. The minimum requirements for STR nomenclature system. This figure showing an example of the minimum requirements for STR nomenclature when using MPS systems that were recommended by the ISFG (Parson *et al.* 2016).

However, MPS systems are not adopted in all forensic genetics laboratories. The cost of the instruments and the kits may delay adopting MPS in the routine work of forensic genetics laboratories. In addition, the huge amount of data generated by MPS systems increases the demands to establish a sophisticated software that capable to analyse this data (Børsting and Morling 2015).

However, the benefits can be significant, for example the ForenSeq DNA Signature Prep Kit was used to test 62 samples from Native American tribe (Yavapai) and result in

a combined MP (STRs and iiSNPs) more than $3E-61$, where 1 ng was enough to generate full DNA profiles (Wendt *et al.* 2016). A recent study found that the kit is a powerful tool in kinship testing especially in paternity and full sibling with zero error rate (Li, R. *et al.* 2019). Interestingly, the true positive (TP) and the true negative (TN) of testing second generation relationships, including half-siblings, and uncle/aunt-nephew/niece, and grandparent-grandchild; were 93.6% and 92.4% respectively (Li, R. *et al.* 2019).

1.12 Project Background

Saudi Arabia, in the Southwest region of Asia, occupies the majority of the Arabian Peninsula. It shares borders with eight Arab countries: Bahrain, Qatar, and the United Arab Emirates (UAE) to the East; Oman and Yemen to the South; Jordan and Iraq to the North and Kuwait to the Northeast. Saudi Arabia is divided into 13 administrative provinces: Makkah, Al-Madinah, Riyadh, Eastern Province, Al-Qassim, Asir, Hail, Tabuk, Northern Borders, Jizan, Al-Baha, Al-Jouf, and Najran (Figure 1.17) (Alsafiah *et al.* 2017).



Figure 1.17. Saudi Arabia administrative divisions. This map is showing the 13 administrative provinces: Makkah, Al-Madinah, Riyadh, Eastern Province, Al-Qassim, Asir, Hail, Tabuk, Northern Borders, Jizan, Al-Baha, Al-Jouf, and Najran. It also shows the eight Arab countries (image from <https://www.123rf.com/>).

As of the 2016 census, the Saudi population was 31,742,308 (20,064,970 were Saudis and 11,677,338 were non-Saudis). Half of the population resides in two administrative provinces of Riyadh and Makkah (General Authority for Statistics in Saudi Arabia 2016). Saudi Arabia is an Arab country where African and Asian surrounding populations have influenced the genetic structure of its population (Abu-Amero *et al.* 2007, 2009). The majority of Saudi Arabian Y-chromosome composition was estimated to be of Levantine origins (69%); with significant contributions from the east via Iran (17%), and Africa (14%) (Abu-Amero *et al.* 2009). The intermediate location between Africa and Asia, and the coastal borders of the Red Sea and the Persian Gulf, have facilitated migrations between Africa and Asia, and trading between neighbouring areas. In addition, the

Arabian Peninsula is connected to the Levant by a long landlocked area that has contained important routes for trading caravans and migration. The movement of people has increased over the last centuries through the presence of the holy Cities of Mecca and Al-Madinah, which have received millions of Muslims performing the Haj for more than 1,400 years, some of whom have remained for many generations (Alsafiah *et al.* 2017).

The first forensic genetics laboratory in the Criminal Evidences Administration was established in 1991, in the capital city of Riyadh. Since then, another 12 forensic genetics laboratories have been established, ten of the 13 laboratories, are accredited to ISO17025:2005/2017 (Alsafiah *et al.* 2017). The main contribution of Saudi laboratories is toward fighting terrorism, solving crimes, and the identifications of human remains resulting from terrorist attacks or mass disasters (e.g. explosions in petro-chemical factories, or accidents during the Haj). Although paternity testing regulations are very strict due the tribal nature of the Saudi population, paternity cases are being addressed by DNA analysis, when directed by the courts.

Laboratory Information Management Systems (LIMS) administrates the workflow in the Saudi forensic genetics laboratories. DNA IQ™ System (Promega Corporation) and biomek 2000 laboratory automation workstations (Beckman Coulter, USA) are used in the extraction laboratories. Extracted DNA from forensic samples is quantified by using Quantifiler Human DNA Quantification Kit and 7500 Real-Time PCR System (AB).

AmpFISTR Identifiler Plus kit is the standard STR kit in Saudi Arabia that provides a typical match probability of $2.2278E-18$, D19S433 is the most informative locus, and TPOX is the least informative locus (Alsafiah *et al.* 2017). In some cases, AmpFLSTR™ Yfiler™ PCR Amplification Kit (AB) for Y-STR markers and Investigator Argus X-12 QS Kit

(Qiagen) for X-STR markers, are used. Capillary electrophoresis is performed using 3500 Genetic Analysers (AB).

The main laboratory, in the capital City of Riyadh, holds the Saudi DNA Data Bank (SDDB) and the other 12 laboratories deal with the SDDB as clients either for adding or for searching.

At the time of the study, four studies have described the genetic diversity of forensic STRs in the Saudi population. The first was a study of 207 samples with eight STR loci (Sinha *et al.* 1999); another two studies investigated 13 STR loci in Saudi individuals residing in Dubai (94 samples) (Alshamali *et al.* 2005) and 15 STR loci in individuals residing in Kuwait (250 samples) (Al-Enizi *et al.* 2013). The most recent study was carried out in 2015, testing 190 individuals from the Riyadh province using the AmpFISTR Identifiler PCR amplification kit (Osman *et al.* 2015). However, Saudi individuals residing in Dubai, Kuwait or even in the Riyadh province are not necessarily representative of the entire population of Saudi Arabia (Alsafiah *et al.* 2017).

In addition, consanguineous marriage is a major factor in shaping the genetic structure of the Saudi population. Previous studies conducted by questionnaires on 3212 families (El-Hazmi *et al.* 1995) and on 4498 pregnant women (Wong and Anokute 1990) found that the percentages of consanguinity were 56.8% and 54.3% respectively. First cousin marriage prevalence was 25.8% and 31.4% while the prevalence of second cousin marriage was 31% and 22.9% of the Saudi population (El-Hazmi *et al.* 1995, Wong and Anokute 1990) respectively. El-Hazmi *et al.* (1995) studied five provinces and the highest rate (67.7%) was observed in the North Western province (Al-Madinah and Tabuk provinces based on the new division system) and the lowest (52.1%) was in the Northern Borders province with an overall inbreeding coefficient of 0.024.

Recently, three studies were published about the Saudi population (Khubrani *et al.* 2018, Khubrani *et al.* 2019a, Khubrani *et al.* 2019b). Khubrani *et al.* (2018) studied 597 male samples from five different regions, which are Central (Riyadh and Al-Qassim provinces), Northern (Northern borders, Tabuk, Al-Jauf and Hail provinces), Southern (Asir, Jazan, Bahah and Najran provinces), Eastern (Eastern province) and Western (Makkah and Al-Madinah provinces). The study used the Yfiler®Plus PCR Amplification Kit (AB) to generate Y-chromosome haplotypes for 27 STRs. By comparing the predicted haplogroups, the Central and Northern regions showed low diversity, while high diversity was observed in the Eastern and the Western regions. In addition, high similarity was observed between samples from the Central and Northern regions and between samples from Eastern and Western. However, the Southern region was distinguished from all other regions (Khubrani *et al.* 2018).

This confirms the heterogeneity of the Saudi population. It is more likely due to the geographical isolation of the Central and Northern regions and the coastal borders of the Eastern and Western areas that allow historical immigrations between Africa and Asia. On the other hand, the Southern region is an agricultural region, where the lands are valuable, and inhabited by tribes who preserve the land within the families by consanguineous marriages.

Khubrani *et al.* (2019a) studied 523 male samples from the population of Saudi Arabia using the GlobalFiler kit. The study highlighted excess of homozygosity in 20/21 aSTRs and the data set showed 0.0476 inbreeding coefficient (F_{IS}), suggesting history of consanguineous marriages.

The excess of homozygosity and the elevated F_{IS} was also observed in the sequence-based data when 89 male samples from the Saudi population were examined using

ForenSeq™ DNA Signature Prep Kit (Khubrani *et al.* (2019b)). The study reported excess of homozygosity in 23/27 aSTRs and 63/91 tested with an overall F_{IS} of 0.04131.

The heterogeneity nature of the Saudi population and the elevated level of consanguinity increase the importance of studying the genetic diversity to evaluate to what extent new STR markers can be utilized for crime scene investigations and for expanded kinship testing (testing beyond just parent-child relationships).

1.13 Project Aims

To evaluate the GlobalFiler PCR amplification kit for use in Saudi Arabia for human identification and kinship testing. In addition, to evaluate SureID®23comp Human Identification kit as a supplementary STR kit for complex kinship testing. Finally, to evaluate ForenSeq™ DNA Signature Prep kit for human identification and kinship testing in Saudi Arabia.

1.14 Objectives

- 1- Gain an ethical approval for the PhD project.
- 2- Collecting around 500 samples from the population of Saudi Arabia.
- 3- Using the GlobalFiler PCR Amplification Kit to genotype 21 aSTRs included in the kit.
- 4- Characterising microvariant alleles observed when using the GlobalFiler kit, which have not been characterised before.
- 5- Examine 17 non-CODIS STR loci in a new kit specifically designed to complement the existing kits: the kit is SureID 23comp (Health Gene Technologies, China). This will generate data 38 STR loci which will be evaluated for human identification and kinship testing in Saudi Arabia. This will include concordance study for five

loci common with the GlobalFiler kit and an evaluation of the kit following the minimum criteria for validation recommended by the ENFSI and the SWGDAM.

- 6- Using ForenSeq™ DNA Signature Prep kit to examine micro variation in the STR loci studied to date and to generate information on a selection of SNP markers. Concordance study for loci that are common with the GlobalFiler kit and the SureID 23comp will also be carried out. This includes generating information about SE33 sequence-based data for the Saudi population.
- 7- Using the data generated from Objectives 3, 5 and 6 (42 aSTRs and 94 iiSNPs) to assess their performance in kinship testing in Saudi Arabia.

2 Chapter Two: Materials and Methods

2.1 Background

This chapter describes the materials and methods used throughout the experimental work. Table 2.1 shows all reagents used in the experimental work of the project.

Table 2.1. The reagents and suppliers used in the experimental work.

Item	supplier
Whatman® FTA® card	Whatman, UK
Unistik® 3 Normal (single use safety lancets)	Owen Mumford, USA
QIAamp DNA Mini Kit	Qiagen, Germany
DNA Ladder Plus	NBS-biologicals, UK
SafeView	
GelRed®	Biotium, USA
Qubit® assay tubes	Invitrogen, USA
Qubit® dsDNA HS Kit	
Qubit® Fluorometer 3.0	
Control DNA (G147A)	Promega, USA
GlobalFiler PCR kit	Applied Biosystems (AB), USA
POP-6™ polymer	
50 cm capillary array	
Hi-Di™ Formamide	
600 LIZ™ v2	
2X ReddyMix PCR Master Mix	
SE33-1 and SE33-2 primers	
PureLink™ Quick Gel Extraction Kit	
BigDye™ Terminator v3.1 Cycle Sequencing Kit	
Shrimp Alkaline Phosphatase (SAP)	
PrepFiler™ BTA Forensic DNA Extraction Kit	
Quantifiler™ Trio DNA Quantification Kit	
MicroAmp™ optical 96-well reaction plates	
MicroAmp™ Optical adhesive films	
Agarose gel	
SureID®23comp kit	Health Gene Technologies, China
HGT 5-Dye Matrix Standard	
Size-500-Plus	
2800M control DNA	Promega, USA
G147A control DNA	
Humic acid	Sigma-Aldrich, USA
Tannic acid	
The ForenSeq DNA Signature Prep Kit	Verogen, USA
MiSeq FGx ForenSeq Reagent Kit	

2.2 Samples collection and preparation

2.2.1 Ethical approval

Before collecting samples, a communication started with the Security Forces Hospitals Programme (SFHP, Saudi Arabia) to allow sample collection in the facilities of six branches in Saudi Arabia. In addition, the proposed application with the title 'Forensically Relevant Polymorphisms (STRs/SNPs) in the population of Saudi Arabia', Participant Information Sheet (Appendix 2, Section 10.2.1) and consent form (Appendix 2, Section 10.2.2) were sent to the ethics committee to be studied and approved.

2.2.2 Samples collection

Blood samples were collected by utilizing the facilities of the SFHP in Saudi Arabia. There are six branches of Security Forces Hospitals in different cities (different administrative provinces) Makkah, Al-Madinah, Riyadh, Dammam, Tabuk and Abha. In the Riyadh and Dammam branches, the project had been presented in their lecture halls that allowed participants to gain a better understanding of the project and the consent process.

Although dealing with volunteers who already have medical or scientific backgrounds (staff and trainees in the hospitals) was easier, collecting 500 samples from six cities, in different provinces, within three weeks (15 working days) was the main challenge. Figure 2.1 illustrates the location and order of sampling sites.



Figure 2.1. This map is showing distribution of collection sites. Collection started from Riyadh to Dammam, Riyadh, Abha, Makkah (Mecca), Al-Madinah, Tabuk, and then back to Riyadh.

In each branch, the collection event was announced by sending an e-mail to the staff and the trainees. The e-mail included the Participant Information Sheet and instructed people who are willing to participate to respond to the e-mail. Responders were given a specific date and an approximate time for collection. The collection was from 8 am to 6 pm over the five working days. To ensure that samples are from unrelated representative, every volunteer was asked if she/he has a related person in the same or in any of the other branches. Volunteers had signed consent forms before sample were collected.

Around 200 μ l of liquid blood was spotted onto FTA card (Whatman, UK). Cards were left in a clean ventilated fume hood to dry the blood spots before placing the card in an envelope. Each card and its envelope were identically numbered by a unique number and has the sex of the volunteer.

2.2.3 DNA extraction

DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) following the manufacturer's guidelines (Qiagen 2016) with two modifications. First: samples were incubated in the ATL buffer for at least 6 h/overnight at 56 °C (additional step) before proceeding with the manufacturer's procedure. Second, a volume of 100 µl of the AE buffer was used for the elution stage rather than 150 µl (amended step). Before applying these modifications, the impact of each modification was evaluated by running the extracted DNA of five samples on a 1% agarose gel (Section 2.8.1).

After optimising the extraction procedure, collected samples were extracted in batches, each batch contained 20 samples. For quality control purpose, one blank tube (no DNA) was processed with every extraction batch. Five (2 mm diameter) punches per sample were placed in a 1.5 ml tube. Then, 180 µl of ATL were added and samples were incubated at 56 °C for 6 h/overnight. The manufacturer's procedure started from this point where the temperature of the incubation was raised to 85 °C for 10 min. This was followed by adding 20 µl proteinase K (PK) and an incubation at 56 °C for 1 h. Then, 200 µl of AL buffer were added and samples were incubated at 70 °C for 10 min. Subsequently, 200 µl of absolute ethanol were added. It is important to note that every tube was briefly centrifuged to remove drops from the lid before any addition and was vortexed thoroughly after any addition.

The 600 µl mixture was moved to a QIAamp Mini spin column inserted in a 2 ml collection tube. Then, the tube was centrifuged at 8000 rpm for 1 min. Subsequently, the collection tube was emptied, 500 µl of AW1 buffer was added and the samples were centrifuged at 8000 rpm for 1 min. The final washing step included adding 500 µl of AW2 buffer, and centrifuge at 14000 rpm for 3 min. To collect the extracted DNA, the spin

column was placed in a labelled cross-linked 1.5 ml tube, 100 µl of the AE buffer was added, and samples were incubated at the room temperature for 1 min. Finally, the 1.5 ml tube (with the spin column) was centrifuged at 8000 rpm for 1 min and tubes with the eluted DNA were placed in a -20 °C freezer.

To monitor the performance of the extraction stage, five random samples from each batch (5/20 samples) were run on a 1% agarose gel (Section 2.8.1). If any of the five samples did not show an obvious band, all samples from the same batch were run on a gel to identify samples that need re-extraction.

2.2.4 Quantification of the extracted DNA

The concentrations of the extracted DNA were estimated using Qubit® dsDNA HS Kit and Qubit® Fluorometer 3.0 (Invitrogen, USA). The samples were processed in ten batches, each batch contained 50 DNA samples.

Qubit® assay tubes (0.5 ml) were used for the quantification reaction. Based on the manufacturer's protocol, 10 µl of extracted DNA or of the Qubit® standards (Qubit® dsDNA HS Standard #1 and #2) was added to 190 µl Qubit® working solution. However, the workflow of the assay was reversed by adding the extracted DNA or the standards first, followed by adding Qubit® working solution to all the tubes. The Qubit® working solution was prepared by 1 µl of Qubit® dsDNA HS Reagent and 189 µl of Qubit® dsDNA HS Buffer.

The workflow reversal was to reduce the effect of variations in the incubation time between the tested samples. In addition to the standard tubes, an addition tube that contained a known concentration DNA (G147A control DNA, Promega) was analysed with each batch. The G147A control was diluted to 1.92 ng/µl (original concentration

192 ng/ μ l) to place it within the detection range of the kit (0.001 ng/ μ l to 100 ng/ μ l). In total, 53 tubes were proceeded in every batch.

After adding the Qubit[®] working solution, the tubes were vortexed for 2-3 s and then incubated for 2-3 min at room temperature. Each of tested sample was read twice and the average of the two reads was defined as the sample's concentration.

2.3 GlobalFiler™ PCR kit.

2.3.1 DNA amplification

The amplification used the half volume reactions that contained 3.75 μ l Master Mix, 1.25 μ l Primer Set, 0.5 ng extracted DNA in 7.5 μ l. The half volume reaction was optimised and validated by comparing the profiles of the positive control using the manufacturer's guidelines (full volume) and the half volume. This was done in three replicates.

Once the half volume reaction was validated, the 500 samples were amplified in batches (90 samples/batch). The amplification reactions were monitored using positive and negative controls (Table 2.2). A Veriti thermal cycler (AB) was utilized to carry out the PCR as following: [95 °C (60 s)] / [94 °C (10 s) 59 °C (90 s)] 29 cycles / [60 °C (10 min)].

Table 2.2. The components of amplification tubes of the GlobalFiler™ PCR kit. The table shows the components of amplification tubes using the half volume reactions of the GlobalFiler™ PCR kit used in the project.

Tube	Master Mix (μ l)	Primer Set (μ l)	DNA (ng)	DNase free Water (μ l)	Total Volume (μ l)
Positive control	3.75	1.25	0.5	2.5	12.5
Negative control	3.75	1.25	0	7.5	12.5
Samples	3.75	1.25	0.5	Up to 7.5	12.5

2.3.2 DNA separation, detection and analysis

An ABI 3500 DNA Genetic Analyser with an 8-capillary array (AB, USA) was employed for separation and detection following the manufacturer's guidelines, except that POP-6™ polymer and a 50 cm capillary array (AB) were used. In addition, the run time, in the Data Collection Software v3 (AB), was increased to 3800 s due to the use of the 50 cm capillary array (Alsafiah *et al.* 2017).

Samples were prepared for injection by adding 0.5 µl of PCR amplicons to 9.5 µl of Liz-Formamide mixture. The Liz-Formamide mixture contained 9.25 µl Hi-Di™ Formamide and 0.25 µl Size Standard 600 LIZ™ v2 (AB) per sample. As recommended by the manufacturer, one well of an allelic ladder, which contained 9.5 µl Liz-Formamide mixture and 0.5 µl allelic ladder, was included every three injections (Alsafiah *et al.* 2017).

Following the published nomenclatures and the guidelines of the International Society for Forensic Genetics (ISFG) (Schneider 2007), alleles from the 21 STRs were identified using the allelic ladder and GeneMapper™ ID-X Software v1.2 (AB) (Alsafiah *et al.* 2017).

2.4 Characterisation of six unusual alleles at SE33 and D1S1656 STR loci.

2.4.1 Sequencing the SE33 alleles

Samples that exhibited alleles of interest were amplified in 25 µl volume reactions. It contained 12.5 µl of 2X ReddyMix PCR Master Mix (AB), 1.25 µl of a 10 µM concentration of each primer, and a total of 2 ng DNA (up to 15 µl). A primer pair (SE33-1 and SE33-2), as published in (Gill *et al.* 1994), was used for the amplification reactions: SE33-1 [5'-AAT CTG GGC GAC AAG AGT GA-3'] and SE33-2 [5'-ACA TCT CCC CTA CCG CTA TA-3'].

Samples were amplified using a Veriti™ Thermal Cycler (AB): [95 °C / 2 min] [(95 °C / 25 s) (60 °C / 30 s) (72 °C / 40 s)] 30 cycles [72 °C / 5 min] (Alsafiah *et al.* 2018).

A 20-cm-long 3% agarose gel was employed to separate target fragments (section 2.8.2). The bands of interest were then sliced and placed into a 15 ml falcon tube for recovery and purification. PureLink™ Quick Gel Extraction Kit (AB) was used for the DNA recovery and purification following the manufacturer's procedure. Based on the weight of each slice and the percentage of the used gel, the volume of gel solubilization buffer (L3) was determined (e.g. 400 mg weight of 3% agarose gel needed 2.4 ml of L3 buffer). Therefore, the volumes of L3 buffer added to the gel slices ranged from 1.3 – 3.1 ml. Samples were then incubated at 50 °C for 15 min or until the gel was completely dissolved. This was followed by adding an equal volume to the L3 buffer of isopropanol to each tube. For the purification stage, the quick gel extraction column inserted in 2 ml collection tube was used. The mixture from the above steps was placed into the column and was centrifuged at 14,000 rpm for a 1 min. Then, 500 µl of W1 wash buffer was added to the column and was centrifuged at 14,000 rpm for a 2 min. The purified DNA was eventually eluted by 50 µl E5 elution buffer that incubated at the room temperature for 1 min. Finally, the DNA was collected in 1.5 ml tubes by centrifuging the column at 14,000 rpm for a 1 min. Collected DNA was placed at -20 °C for storage.

DNA concentrations of the purified fragments were estimated using Qubit™ dsDNA HS Assay Kit and Qubit® Fluorometer 3.0 (AB) following the above procedure (Section 2.2.4)

DNA fragments were sequenced directly using BigDye™ Terminator v3.1 Cycle Sequencing Kit (AB) following an internally validated 10 µl reaction volume. For each DNA strand, the 10 µl sequencing reaction contained 0.75 µl of BigDye® Terminator v3.1

Ready Reaction Mix, 1.7 µl 5X Sequencing Buffer, 0.32 µl of 10 µM primer (forward or reverse), and 3-6 ng of DNA (extracted from the gel). A Veriti™ Thermal Cycler was used for sequencing reaction: [95 °C / 1 min] [(96 °C / 10 s) (50 °C / 5 s) (60 °C / 4 min)] 25 cycles (Alsafiah *et al.* 2018).

Post-sequencing purification was carried out by adding 2 µl Shrimp Alkaline Phosphatase (SAP) (AB) to 5 µl of sequencing products that was followed by an incubation at 37 °C for 60 min then at 65°C for 15 min as recommended by the manufacturer (Alsafiah *et al.* 2018).

Purified products were prepared for separation by adding 5 µl Hi-Di™ Formamide (AB). An ABI 3500 DNA Genetic Analyser, POP-6™ polymer and 50 cm capillary array were employed for separation using the run modules StdSeq50_POP6 and the base calling protocol BDTv3.1_PA_Protocol-POP6. Sequencing raw data was then analysed by sequencing analysis software v5.4 (AB) (Alsafiah *et al.* 2018).

2.4.2 Sequencing the D1S1656 alleles

This part is described in Section 2.6.1.

2.5 An evaluation of the SureID®23comp Human Identification kit.

The SureID®23comp kit was evaluated for forensic applications as a supplementary STR kit. The minimum criteria for validation recommended by the European Network of Forensic Science Institutes (ENFSI) (ENFSI 2010) and by the scientific working group on DNA analysis Methods (SWGAM) (SWGAM 2016) were followed.

2.5.1 Preparation ABI 3500 DNA Genetic Analyser.

The preparation included spectral calibration and run optimisation due to the use of 50 cm capillaries and POP-6™ polymer (AB). The spectral calibration mix was prepared

by adding 8 μl HGT 5-Dye Matrix Standard (Health Gene Technologies) to 200 μl of Hi-Di™ Formamide (AB); 10 μl were dispensed to each well. In the data collection software (AB), the dye set of SureID®23comp was created based on the G5 template as recommended by manufacturer. Based on the manufacturer guidelines, the run time in the run module of HID36_POP4 should be 1,210 – 1,500 s when using a 36 cm capillary. In this study, the run time was increased to 3900 s due to the use of the 50 cm capillaries (Alsafiah *et al.* 2019a).

2.5.2 DNA Samples

Initial tests of the SureID®23comp kit, were carried out using the 2800M control DNA (Promega Corporation). The control DNA was also used for sensitivity and stochastic tests by preparing five serial dilutions of (500, 250, 125, 62, and 31) pg. In addition, 0.5 ng of the control was amplified with the presence of different concentrations (50, 75, 100, 120 and 150) ng/ μl of common PCR inhibitors humic and tannic acids (Sigma-Aldrich, USA), for stability tests (Alsafiah *et al.* 2019a).

The study of precision, accuracy, peak balances, concordance and stutter peak ratios were carried out using the 500 samples from the population of Saudi Arabia. The sensitivity and the stability of the kit were further assessed using nine bone samples collected from a mass grave in Iraq. The bone samples were previously extracted using PrepFiler™ BTA Forensic DNA Extraction Kit (AB) and were quantified using Quantifiler™ Trio DNA Quantification Kit (AB). The concentrations of the small fragments of the Quantifiler™ Trio ranged from 0.0173 ng/ μl to 0.3271 ng/ μl and the Degradation Indexes (DI) were from 1.6758 to 57.666. These samples were previously profiled using one or more of the commonly used STR kits (Table 2.3) (Alsafiah *et al.* 2019a).

Table 2.3. Bone samples used in the evaluation tests of the SureID® 23 comp kit. Nine bone samples, collected from a mass grave in Iraq, were extracted using PrepFiler™ BTA Forensic DNA Extraction Kit (AB), and were quantified using Quantifiler™ Trio DNA Quantification Kit (AB). This table shows Quantifiler™ Trio small fragment concentrations (ng/μl) and degradation indexes (DI) of the samples (Alsafiah *et al.* 2019a).

Sample #	Quantifiler™ Trio		% of detected alleles using different kits		
	Small fragment concentration (ng/μl)	Degradation Index (DI)	PowerPlex®21 (20 STRs)	GlobalFiler™ (21 aSTRs)	PowerPlex®Fusion 6C (23 aSTRs)
76 c	0.0173	57.666	60%	66.60%	60.80%
78 a	0.0194	16.166	90%	95.20%	82.60%
93 b	0.3271	2.7464	100%	N/A	N/A
76 e	0.093	2.2962	100%	N/A	N/A
81 a	0.0571	1.929	100%	76.20%	N/A
97 b	0.0548	1.6758	100%	N/A	N/A
94 a	0.0685	2.4204	100%	N/A	N/A
25 a	0.0463	4.9784	95%	N/A	N/A
46 b	0.0412	3.1937	100%	N/A	N/A

N/A: sample was not profiled using the kit.

2.5.3 DNA amplification

The Initial tests of the SureID®23comp used two reaction volumes that were optimised by the manufacturer. A 25 μl volume that contained 12.5 μl master mix, 6.25 μl primer mix and up to 6.25 μl of DNA template; and a 10 μl volume that contained 5 μl master mix, 2.5 μl primer mix and up to 2.5 μl of DNA template. The range of recommended DNA quantity was 0.5 - 4 ng. To validate both volumes, two operators carried out the initial tests independently with 0.5 ng of control DNA in five replicates (20 tests in total) (Alsafiah *et al.* 2019a).

Three DNA concentrations (0.5, 0.35, and 0.25) ng were used for the first 90 samples from Saudi population, to test the ideal concentration that can achieve the highest peak balances with the 10 μl volume. Then, the rest of samples were genotyped using the 10 μl volume and 0.5 ng as the total DNA input per reaction (Alsafiah *et al.* 2019a).

MicroAmp™ optical 96-well reaction plates and MicroAmp™ Optical adhesive films (AB), were used for amplification. The amplification contents were prepared by adding the 2.5 µl of the DNA and DNase/RNase-free water, followed by aliquoting 7.5 µl the SureID®23comp mix (5 µl master mix and 2.5 µl primer mix). A Veriti thermal cycler (AB) was employed to carry out the amplification reactions as following [95 °C (5 min)] / [94 °C (10 s) 61 °C (60 s) 70 °C (30 s)] 28-30 cycles / [60 °C (15 min)]. The 28-cycles protocol was used for the initial tests, stability tests and for the 500 samples. For sensitivity and stochastic study, the serial dilution samples were amplified in five replicates using both reaction volumes, each volume was tested with 28 and 30 PCR cycles. For the bone samples, the 25 µl volume and 30 PCR cycles were used (Alsafiah *et al.* 2019a).

2.5.4 DNA separation, detection and analysis

Samples were prepared for separation and detection by adding 1 µl of PCR products or of an allelic ladder (Health Gene Technologies) to 9 µl of Formamide/Size-500-Plus mix. This mix was prepared by 9 µl of Hi-Di™ Formamide (AB) and 0.25 µl Size-500-Plus (Health Gene Technologies), for each sample. An ABI 3500 DNA Genetic Analyser with 50 cm capillaries and POP-6™ polymer (AB) was used for the separation and detection by applying 3900 s as the run time as validated in Section 2.5.1 (Alsafiah *et al.* 2019a).

Alleles from the 23 markers were called using GeneMapper™ *ID-X* Software v1.2 (AB) with an allelic ladder mix supported by panels and bins provided by the manufacturer. For the sensitivity and stability tests, the minimum relative fluorescent units (RFU) was 50 RFU for heterozygous genotypes and was 150 RFU for homozygous genotypes (Alsafiah *et al.* 2019a).

2.6 ForenSeq™ DNA Signature Prep kit.

The kit was used in two parts of the project, sequencing the two allele variants of the D1S1656 locus (Chapter 4) and for generating ForenSeq™ data of the Saudi population (Chapter 6). For both parts the Verogen ForenSeq™ DNA Signature Prep kit (Verogen 2018a) was used for the library preparation.

2.6.1 Library preparation and sequencing for the D1S1656 variants.

The library preparation of the two samples showed alleles 7 and 8 at the D1S1656 was carried out using the Primer Mix B for the initial PCR (PCR1). The primer mix B targets additional 78 SNPs (56 aiSNPs and 22 piSNPs) to those markers included in Primer Mix A (27 autosomal STRs, 7 X, 24 Y haplotype markers and 94 iiSNPs) (Table 1.5). All other library preparation steps were carried out following the manufacturer's guidelines (Verogen 2018a) (Alsafiah *et al.* 2018).

The prepared libraries were then sequenced using a MiSeq FGx™ Instrument and a standard flow cell following the manufacturer's guidelines (Verogen 2018c), in the Applications Laboratory (Illumina, Cambridge, United Kingdom) (Alsafiah *et al.* 2018).

2.6.2 Library preparation and sequencing for the Saudi population data.

For the population genetic study, 94 samples from the population of Saudi Arabia were sequenced using the Primer Mix A in the PCR1. All other library preparation steps were carried out following the manufacturer's guidelines (Verogen 2018a), except that the volume of the pooled normalised libraries (PNL) that was added to the human sequencing control (HSC) mixture was increased from 7 µl to 12 µl as validated by Devesse *et al.* (2018).

The denatured normalised libraries (DNL) were then transferred to the reagent cartridge which was then loaded to a MiSeq FGx™ Instrument alongside with a standard flow cell, SBS solution (PR2) Bottle, and the waste Bottle following the manufacturer's guidelines (Verogen 2018c). This part was carried out in Alec Jeffreys Forensic Genomics Unit, Department of Genetics and Genome Biology (University of Leicester, United Kingdom).

2.6.3 Universal analysis software

The run was created using the ForenSeq™ Universal Analysis (UAS) by entering the samples' information including the indices combinations for each sample as described the manufacturer's guide (Verogen 2018b). The UAS was also used to perform sequences analysis, allele call and to generate the samples' report and the run Flanking Region Report. The default setting of the UAS uses, for the analytical and interpretation thresholds (AI and IT), 1.5% and 4.5% of the total number of reads of the most frequent sequences on a locus and applies minimum limits of 10 and 30 reads for the thresholds respectively. This study applied the default setting for the AT, IT and the stutter filter.

2.6.4 Concordance study

The data of 23 autosomal STRs gathered from the CE kits (Sections 2.3 and 2.5) common with ForenSeq™ data was compared using an Excel workbook and any differences were considered as a non-concordance. However, in few cases where a known heterozygote genotype (CE data) at the D22S1045, and the ForenSeq™ data showed only one allele with low coverage, the cases were considered as drop out not a discrepancy event due to the lower allele count ratios (ACR) feature of this locus. In addition, due to the lack of CE data for D4S2408, D6S1043, PentaE and PentaD STRs, they were not included in the concordance study.

2.6.5 Further analysis using the STRait Razor (SR)

The FASTQ files of the samples that were generated from the MiSeq FGx™ instrument were loaded to STRait Razor (SR) v3.0 (Woerner *et al.* 2017). To recover as many as possible of sequences, the parameters of 0.20 for the heterozygote threshold and 2 for coverage threshold were used. Discordance events appeared in Section 2.6.2, were further investigated for possible allele drop-out, drop-in, or alleles imbalances. Additionally, allele calling generated from the SR was compared to those generated by the UAS to investigate any bioinformatical discordance.

2.6.6 SE33 sequence-based data

The SR (Woerner *et al.* 2017), was also used to recover the SE33 sequences from the FASTQ files generated by the MiSeq FGx after modifying the configuration file by adding the 5' and 3' anchors and motif sequence provided in (Borsuk *et al.* 2018). For the SE33 sequences, all sequences with ≥ 10 reads (depth of coverage, DoC) were included and heterozygous sequences that showed $\geq 20\%$ of ACR were considered as true sequences. Sequences that showed less than 20% ACR were then recovered manually (Alsafiah *et al.* 2019b).

2.6.7 Novelty assessment

The novelty assessment of an allele was started with the STRait Razor v3.0 by which alleles showed “Novel Sequence” were further assessed. The alleles were then searched for in the literature that included samples from the Middle East (Phillips *et al.* 2018a), from the Qatari population (Almohammed and Hadi 2019) and from Saudi Arabia (Khubrani *et al.* 2019b). Finally, unreported alleles were searched for in the GenBank database.

For the SE33, the novelty of a motif pattern or of an allele sequence was assessed based on those motifs and sequences reported in (Borsuk *et al.* 2018) and in the GenBank database.

2.7 Evaluation of DNA markers

2.7.1 Forensic parameters

The parameters included power of discrimination (PoD), power of exclusion (PoE), matching probability (MP), polymorphism information content (PIC), observed homozygosity (Ho) and typical paternity index (PI).

For 38 autosomal STRs which were generated from Sections 2.3 and 2.5, the statistical parameters were calculated using PowerStat v 1.2 (Promega Promega Corporation). For DNA markers included in the ForenSeq™ DNA Signature Prep kit, GenAlEx 6.5 Excel software (Peakall and Smouse 2012), was used.

2.7.2 Hardy-Weinberg equilibrium

Convert software (Glaubitz 2004), was employed to convert an Excel sheet, which contains the data, to the input file (ARP files) for Arlequin Software. The expected heterozygosity (He) and the exact test for detecting deviation from the Hardy-Weinberg equilibrium (HWE) was carried out by Arlequin v3.5.2 Software (Excoffier *et al.* 2007), using values of 1,000,000 steps for the Markov chain and 100,000 for the dememorization steps.

2.7.3 Linkage disequilibrium test

Arlequin v3.5.2 Software (Excoffier *et al.* 2007), was also used to test linkage disequilibrium (LD) between syntenic loci (STR-STR, STR-SNP, and SNP-SNP). The data of the 500 samples were used to test LD between 12 syntenic pairs at the same arm resulted from combining GlobalFiler™ (Section 2.3) and SureID®23comp (Section 2.5)

kits. The data of the 87 samples sequenced by ForenSeq™ DNA Signature Prep kit (Section 2.6) were used to test LD between 166 syntenic loci at the same arm resulted from using ForenSeq™ alone. The LD test was carried out by applying the values of 1000 in the permutations and of 2 in the Expectation-Maximisation (EM) algorithm.

2.7.4 Population differentiation test, F_{ST} calculation, and inbreeding coefficient (F_{IS})

Arlequin v3.5.2 Software was also used to perform a population differentiation exact test and to calculate the F_{ST} values.

For GlobalFiler™ data (Section 2.3), fourteen populations were compared that included previous studies in the Saudi population reported by Sinha *et al.* (1999) (207 samples), Osman *et al.* (2015) (190 samples) and by Khubrani *et al.* (2019a) (523 samples). In addition, the comparison included Saudi individuals residing in Kuwait (Al-Enizi *et al.* 2013) (250 samples) and in Dubai (Alshamali *et al.* 2005) (94 samples). Gulf Cooperation Council (GCC) populations were also included: Kuwait (Al-Enizi *et al.* 2013) (502 samples), United Arab Emirates (Jones *et al.* 2017) (477 samples), Qatar (Perez-Miranda *et al.* 2006) (133 samples), Yemeni (101 samples) and Omani (79 samples) residing in Dubai (Alshamali *et al.* 2005). Egyptian (421 samples), Iraqi (146 samples), Iranian (287 samples), and Indian (415 samples) residing in Kuwait (Al-Enizi *et al.* 2013) were also included in the comparison. This part used the allele frequency data.

For the SureID®23comp data (Section 2.5), four populations were included in the comparison that are European (321 samples), South Asian (315 samples), African (284 samples) (Iyavoo *et al.* 2019), and Ningbo population (284 samples) (China, data provided by the Health Gene Technologies). Here, the genotype data were used in the comparison.

The AMOVA test was also carried out using Arlequin v3.5.2 Software to estimate the average F_{IS} for 21 loci included in GlobalFiler™ kit (Section 2.3), 17 loci included in SureID®23comp data (Section 2.5), and 122 loci included in ForenSeq™ data (Section 2.6).

2.7.5 RStudio platform and packages used in the project.

RStudio platform (RStudio Team 2016) and DNA tools package (James and Curran 2017), were used to identify the maximum number of matched loci and partial matches within the tested samples. The ggplot2 package (Wickham 2016) was also used for plotting figures. Finally, the cmdscale function (Ingwer and Patrick 2005) was used in R software to generate a multi-dimensional scale (MDS).

2.8 Gel electrophoresis

2.8.1 Assessment of extraction procedure and DNA yield

Gel electrophoresis was employed to study the effect of each of the modifications during the optimisation of the extraction procedure and to give an initial assessment of the extracted DNA in all batches (Section 2.2.3). For both tests, a 1% agarose gel was used that was prepared by 0.5 g of agarose gel (AB) dissolved in 50 ml of Tris-Acetate with Ethylenediaminetetraacetic acid (EDTA) buffer (TAE) and 5 µl of a nucleic acid stain SafeView (NBS Biologicals, UK) or GelRed® (Biotium, USA). A total of 5 µl of the extracted DNA and 2 µl of 100 bp DNA Ladder Plus (NBS-biologicals) were loaded into the gel and the electrophoresis was at 100 v for 20 min.

2.8.2 Preparation of the 20-cm-long 3% agarose

Gel electrophoresis was also employed for separation DNA fragments of the SE33 alleles (Section 2.4.1). The 20-cm-long 3% agarose gel was prepared by adding 6 g of agarose gel (AB) to 220 mL of TAE buffer and 16 µl of GelRed® (Biotium). A total of 25 µl

of the PCR products of SE33 locus and 10 µl of 100 bp DNA Ladder Plus (NBS-biologicals) were loaded into the gel. Electrophoresis was at 120 v for 6 h (Alsafiah *et al.* 2018).

2.9 An evaluation of 136 DNA markers for kinship testing

The data of 42 aSTRs and 94 iiSNPs (136 markers) generated in Sections 2.3, 2.5 and 2.6 were used in the simulation studies to evaluate the extent that they can improve the resolution of kinship testing in Saudi Arabia. An in-house Excel sheet was used to create a hypothetical pedigree based on the allele frequencies of the DNA markers obtained (Figure 2.2).

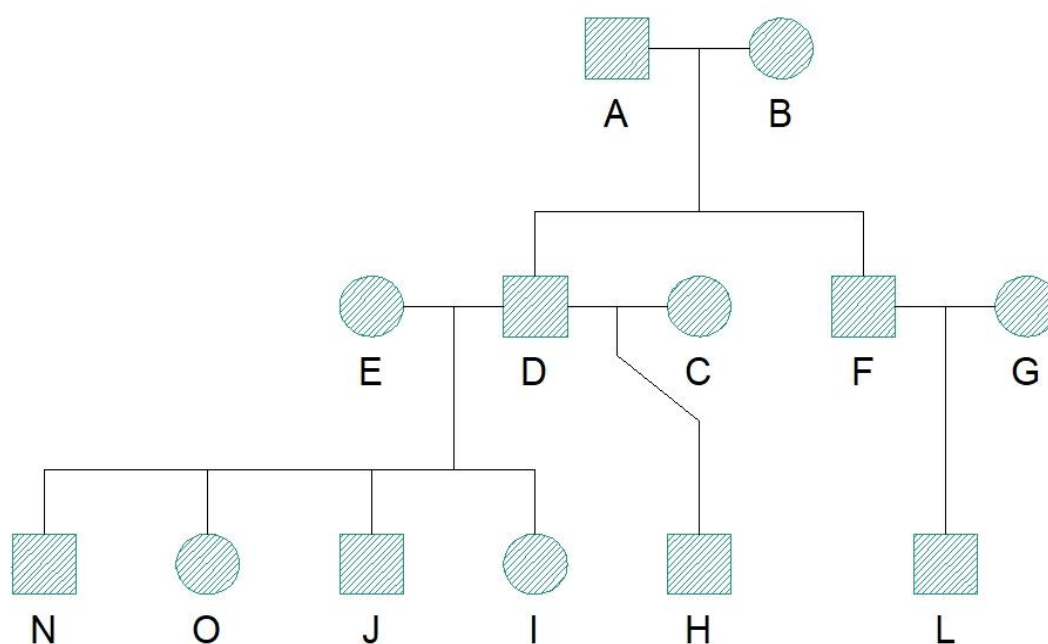


Figure 2.2. A hypothetical pedigree created by an in-house Excel sheet. The hypothetical pedigree comprised of three generations and 13 members. Circles represent female members and squares represent male members. The profiles of the members were generated by the in-house Excel sheet based on the allele frequencies of the 136 loci.

Familias3 software v3.2.7 (Kling *et al.* 2014) was used to test the parent-Child relationships of the hypothetical pedigree's member to validate the members' profiles, before starting the simulation study, using the blind search option. The software was

then used to do the simulation study using the allele frequency data of the 136 markers and the DNA profiles of the pedigree's members after setting the mutation rates.

2.9.1 Setting up the mutation rates in the Familias3 software

For aSTRs, the sex-specific (Maternal/Paternal) mutation rates of 38/42 STRs were adopted from (Butler 2015, Lan *et al.* 2018, Jin *et al.* 2016) (no data were available for D3S1744, D4S2366, D19S253 and D21S2055) (Table 2.4).

Table 2.4. Mutation rates for aSTRs that were reviewed from literatures. The mutation rates of 38 aSTRs were reviewed from (Butler 2015, Lan *et al.* 2018, Jin *et al.* 2016) and were used in the simulation study. No mutation rates were available for D3S1744, D4S2366, D19S253 and D21S2055.

STRs	Mutation rates		STRs	Mutation rates	
	Maternal	Paternal		Maternal	Paternal
D1S1656 (Butler 2015)	0	0.0025	TH01 (Butler 2015)	0.0001	0.0001
D2S441 (Butler 2015)	0	0.0025	D11S2368 (Jin <i>et al.</i> 2016)	0.00047	0.00189
D2S1338 (Butler 2015)	0.0002	0.001	D12S391 (Butler 2015)	0.0003	0.003
TPOX (Butler 2015)	0	0.0001	vWA (Butler 2015)	0.0003	0.0017
D3S1358 (Butler 2015)	0.0002	0.0013	D13S317 (Butler 2015)	0.0004	0.0014
D3S1744	-	-	D13S325 (Jin <i>et al.</i> 2016)	0	0.00095
FGA (Butler 2015)	0.0005	0.0032	D14S1434 (Lan <i>et al.</i> 2018)	0	0.00103
D4S2366	-	-	D15S659 (Jin <i>et al.</i> 2016)	0	0.00081
D4S2408 (Lan <i>et al.</i> 2018)	0	0.00103	PentaE (Butler 2015)	0.0007	0.0013
D5S818 (Butler 2015)	0.0003	0.0012	D16S539 (Butler 2015)	0.0003	0.0011
D5S2800 (Lan <i>et al.</i> 2018)	0	0	D17S1301 (Lan <i>et al.</i> 2018)	0.00103	0.00206
CSF1PO (Butler 2015)	0.0003	0.0015	D18S51 (Butler 2015)	0.0006	0.0022
SE33 (Butler 2015)	0.003	0.0064	D18S1364 (Jin <i>et al.</i> 2016)	0	0.00141
D6S474 (Lan <i>et al.</i> 2018)	0	0	D19S433 (Butler 2015)	0.0005	0.0008
D6S1043 (Butler 2015)	0.0003	0.0006	D19S253	-	-
D7S820 (Butler 2015)	0.0001	0.0012	D20S482 (Lan <i>et al.</i> 2018)	0.00103	0
D7S3048 (Jin <i>et al.</i> 2016)	0.00095	0	D21S11 (Butler 2015)	0.0011	0.0015
D8S1179 (Butler 2015)	0.0002	0.0016	D21S2055	-	-
D8S1132 (Jin <i>et al.</i> 2016)	0.00095	0.00143	PentaD (Butler 2015)	0.0006	0.0009
D9S1122 (Lan <i>et al.</i> 2018)	0.00103	0	D22S1045 (Butler 2015)	0	0.0025
D10S1248 (Butler 2015)	0	0.0025	D22GATA198B05 (Jin <i>et al.</i> 2016)	0	0.00144

The mutation rates were applied for each locus in the rate box in the software (Figure 2.3). In addition, due the lack of allele-specific mutation rates for every allele at a locus, the extended stepwise model was used by applying a fixed probability (one tenth less) for each \pm unit difference (0.1 in the range box) (i. e. the mutation rates in Table 2.4 will be decreased by one tenth (0.1) for allele $x \pm 1$ repeat, 0.01 for allele $x \pm 2$...etc). The mutation rate of 0.000001 was used for microvariants (e.g. allele $x.1 >$ allele $x.2$) (rate

2). For iiSNPs, equal probability type was used by applying the mutation rate of $2.5E-8$ as reported in (Nachman and Crowell 2000) (Figure 2.3). All mutation settings were as recommended by Daniel Kling (kinship testing workshop in ISFG2019, Prague).

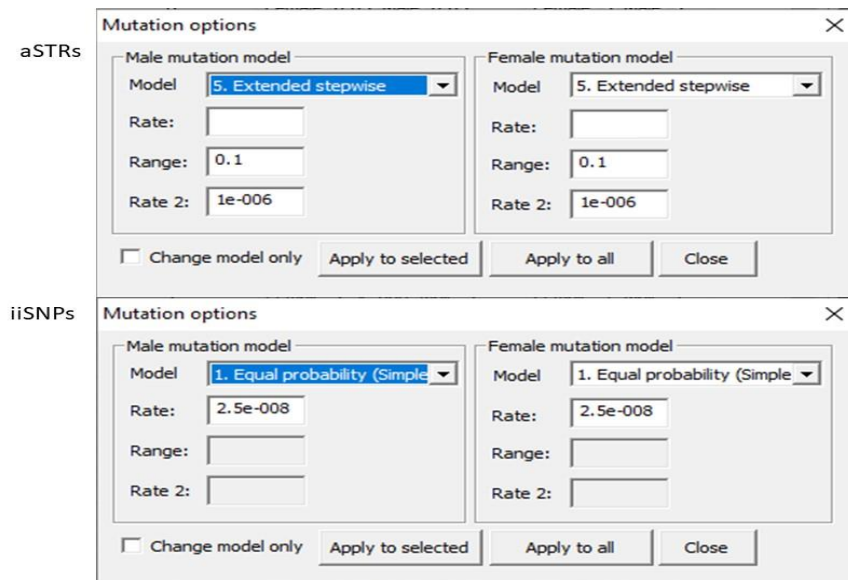


Figure 2.3. Mutation rate settings in Familias3 software. For aSTRs, extended stepwise model was used by applying the mutation rate in Table 2.4 in the rate box, 0.1 in the range box, 0.000001 in the rate 2 box for microvariants alleles. For iiSNPs, equal probability was used by applying $2.5E-8$ in the rate box.

2.9.2 Simulation study

The simulation was conducted using Familias3 software v3.2.7 (Kling *et al.* 2014) to test different combinations of DNA markers that make up the commercially available kits of Identifiler Plus (15 aSTRs), GlobalFiler (21 aSTRs), GlobalFiler and SureID (38 aSTRs), Fusion 6C and SureID (40 aSTRs), and ForenSeq DNA Signature Prep kit (27 aSTRs and 94 iiSNPs). In addition, all loci (42 aSTRs and 94 iiSNPs) and the 94 iiSNPs alone were also tested. Although none of the samples were typed by Identifiler Plus or Fusion 6C, the data of all loci included in the kits have been obtained from Sections 2.3 and 2.6.

Eight different scenarios were assumed to test potential five types of relationships, each of which was based on two hypotheses as shown in (Table 2.5).

In the simulation, members included in the simulation (genotyped) were simulated 1000 times using the random seeds. When the simulation finished, different LR thresholds (1, 10, 100, 1,000, 10,000 and 100,000) were tested to find out the limits (the true positive (TP) and the false positive (FP)) of each LR threshold.

Table 2.5. The hypotheses 1 and 2 that were used in the simulation study. The simulation study was conducted using Familias3 software v3.2.7 (Kling *et al.* 2014). A total of 8 scenarios for five different relationships were tested. The table also shows members who were simulated (genotyped) in each run (orange colour).

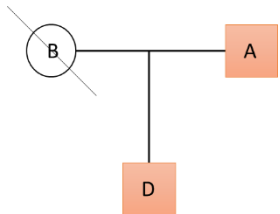
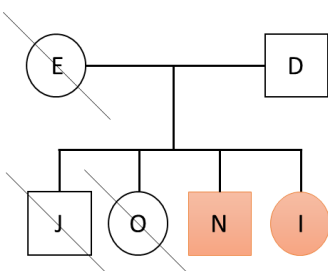
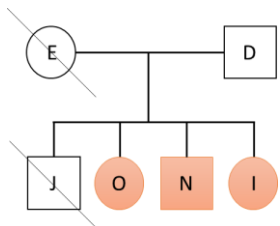
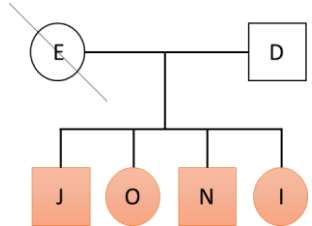
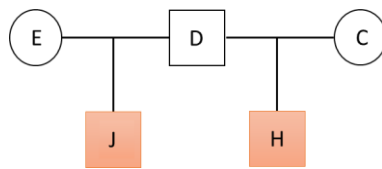
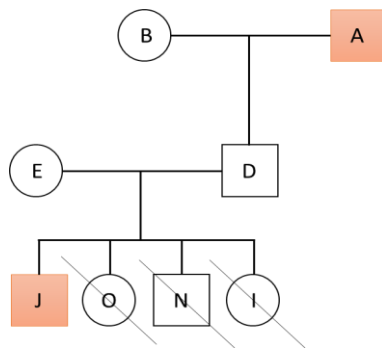
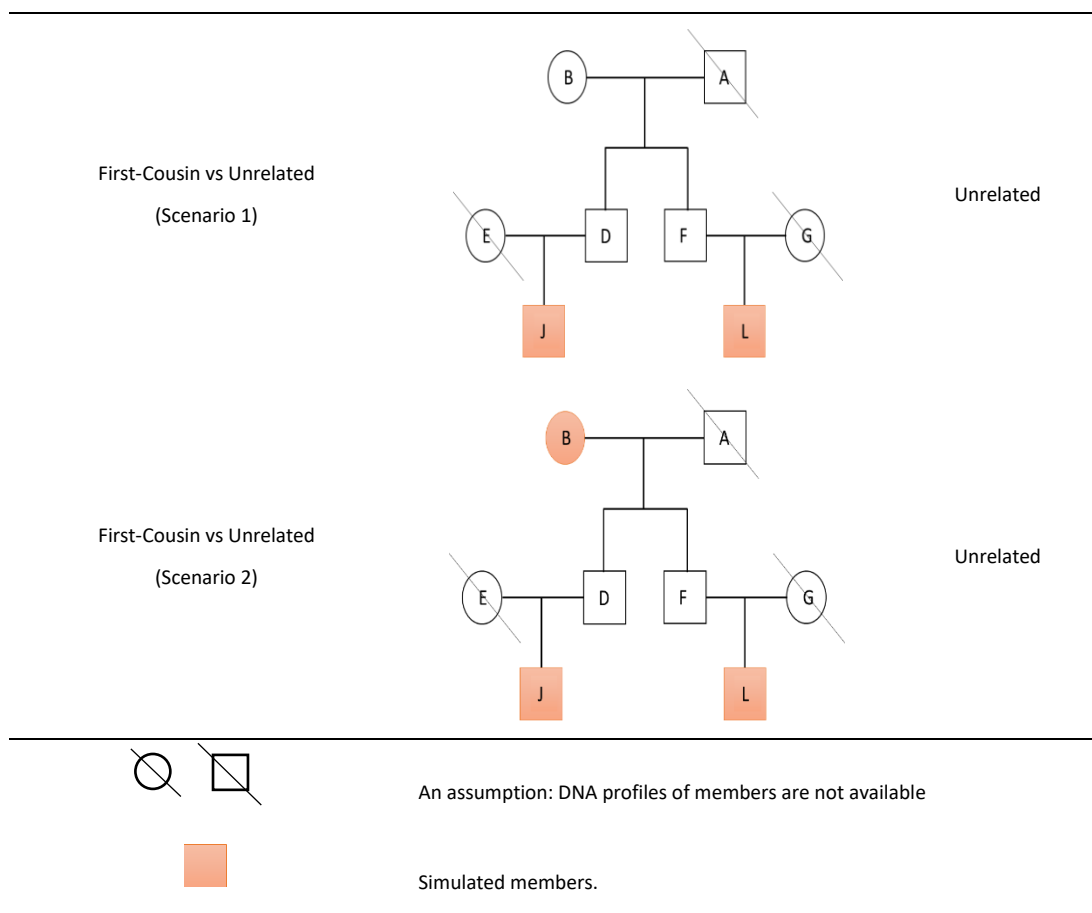
Tested relationship	Hypothesis (1)	Hypothesis (2)
Parent-Child vs Unrelated (Mother not genotyped)		Unrelated
Full-Siblings vs Unrelated (Scenario 1)		Unrelated
Full-Siblings vs Unrelated (Scenario 2)		Unrelated
Full-Siblings vs Unrelated (Scenario 3)		Unrelated
Half-Siblings vs Unrelated		Unrelated
Grand-Parent or Child/ Unrelated		Unrelated

Table 2.5. continued.



In addition, the Familias3 software generates a data file (Simulation_LRs) that could be visualised by plotting in RStudio platform and produced two plots: LR distributions and exceedance probability (a figure that shows the improvement in probabilities at different LR thresholds).

2.9.3 Estimating the genetic distance between syntenic pairs

The genetic distances in cM of syntenic pairs resulted from combining the 136 markers were estimated as described by Phillips *et al.* (2012). This needed to estimate the cumulative genetic map distance (cM) for each marker first.

The cumulative genetic map distance for 41/42 aSTRs were already published and were reviewed from (Phillips 2017) (D16S539 was not available).

For D16S539 and the 94 iiSNPs, the HapMap recombination map was retrieved from (ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/) that was used to approximate the cumulative genetic map distance for each marker. The rs925658351 SNP (Chr. 16:86386300, GRCh37) located at the 5' flanking region 8 bp before the repeat region of D16S539 STR was used to approximate the cumulative genetic map distance of the D16S539 STR (repeat region starts from 86386308 to 86386351, GRCh37) as recommended by Phillips, C. (personal communication).

Then, the locations (bp) of the 95 SNPs (94 iiSNPs and the rs925658351 SNP for the D16S539 STR) were retrieved from the 1000 Genome Browser (<https://www.internationalgenome.org/1000-genomes-browsers/>) using the GRCh37 coordinates. The SNP location (bp) was then used to find the closest map position to the SNP using the HapMap recombination map, by which the approximate cumulative genetic map distance (cM) could be estimated. Figure 2.4 shows how the genetic distance in cM was estimated for the rs560681 and rs10495407 SNPs (Chr. 1), as an example.

2.9.4 Calculation of recombination fraction (RF) using Kosambi mapping function.

Once the cumulative genetic map distance in cM was estimated for each marker, the distance between any two markers in cM was then calculated (subtraction the two values). The RF rate was then calculated by Kosambi mapping function using the Excel tool provided by Phillips *et al.* (2012) (Figure 2.4).

1- Defined the SNP location using the 1000 Genome Browser (GRCh37).

rs560681 SNP

Most severe consequence	intron variant See all predicted consequences
Alleles	A/G Ancestral: A MAF: 0.34 (G) Highest population MAF: 0.47
Location	Chromosome 1:160786670 (forward strand) VCF: 1 160786670 rs560681 A G
Evidence status	
HGVS names	This variant has 11 HGVS names - Show
Synonyms	This variant has 2 synonyms - Show
Genotyping chips	This variant has assays on 7 chips - Show
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) View in dbSNP
About this variant	This variant overlaps 7 transcripts, has 3685 sample genotypes and is mentioned in 7 citations.

rs10495407 SNP

Most severe consequence	intergenic variant
Alleles	G/A Ancestral: G MAF: 0.24 (A) Highest population MAF: 0.43
Location	Chromosome 1:238439308 (forward strand) VCF: 1 238439308 rs10495407 G
Evidence status	
HGVS name	NC_000001.10:g.238439308G>A
Synonyms	This variant has 2 synonyms - Show
Genotyping chips	This variant has assays on 8 chips - Show
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) View in dbSNP
About this variant	This variant has 3816 sample genotypes and is mentioned in 10 citations.

2- Defined the closest map position to the SNP using the HapMap recombination map.

rs560681

	A	B	C	D	E
1	Chromosc	Position(bp)	Rate(cM/Mb)	Map(cM)	
2	chr1	160777713	0.078871	173.516685	-8957
3	chr1	160779495	0.083403	173.516825	-7175
4	chr1	160780555	0.48498	173.516914	-6115
5	chr1	160782561	0.150676	173.517886	-4109
6	chr1	160787725	0.273362	173.518665	1055
7	chr1	160788131	0.345124	173.518776	1461
8	chr1	160788217	0.582114	173.518805	1547
9	chr1	160790411	0.507136	173.520082	3741
10	chr1	160791411	0.602325	173.52059	4741
11	chr1	160791880	0.589007	173.520872	5210
12	chr1	160791892	0.57576	173.520879	5222

rs10495407

	A	B	C	D	E
1	Chromosome	Position(bp)	Rate(cM/Mb)	Map(cM)	
2	chr1	238438554	0.261527	264.509405	-754
3	chr1	238439308	0.335417	264.509602	0
4	chr1	238440236	0.244815	264.509913	928
5	chr1	238440844	0.236624	264.510062	1536
6	chr1	238441600	0.235338	264.510241	2292
7	chr1	238441781	0.234729	264.510283	2473
8	chr1	238441850	0.234225	264.5103	2542
9	chr1	238442713	0.234475	264.510502	3405
10	chr1	238443711	0.234702	264.510736	4403

HapMap recombination map for Chromosome 1

3- The cumulative genetic map distance (cM):

Rs560681 : 173.51866

Rs10495407 :264.5096

4- The genetic map distance (cM) between rs560681 and rs10495407:

= 264.5096 - 173.51866

= 90.9909 cM

RF (Kosambi function) = 0.47441

	A	B	C	D	E	F	G	H
1	Supplementary File S2. Kosambi mapping function calculator							
2								
3	(formula components)				enter cM value here		Kosambi MF Rc values returned here	
4								
5	0.948820274	0.026261902	0.909909	90.9909		0.474410137		

Figure 2.4. An example of how the genetic distance (cM) between syntenic pairs was calculated. This figure shows how the genetic distance (cM) between syntenic pairs was calculated as described by Phillips *et al.* (2012).

3 Chapter Three: An evaluation of 21 autosomal STRs for the population of Saudi Arabia using the Globalfiler™ PCR Amplification Kit.

3.1 Overview of experiment

This chapter presents the sample collection and preparation for downstream applications. In addition, the samples were genotyped using Globalfiler™ PCR amplification kit (AB) for 21 aSTRs included (D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, SE33, D10S1248, D1S1656, D12S391 and D2S1338). The 21 aSTRs were then evaluated for forensic applications in the population of Saudi Arabia including assessment of HWE, population differentiation, and calculations of forensic statistical parameters.

3.2 Aims of the study

The initial aims of this chapter were to obtain ethical approval for the research and collect sufficient blood samples from unrelated representatives from the population of Saudi Arabia. This was followed by DNA extraction and quantification to prepare the samples for downstream applications.

The second aim was to use the Globalfiler™ PCR amplification kit (AB) to genotype 21 aSTRs to evaluate their performance in human identification applications in Saudi Arabia and compare it with the currently used kit (Identifiler® Plus). This include generating allele frequency data for the Saudi population that facilitates the estimation of match probabilities of DNA profiles in Saudi Arabia.

This chapter also aimed to compare the Saudi population with neighbouring populations.

3.3 Objectives

- 1- To obtain ethical approval for the project from a recognised foundation in Saudi Arabia and from the Ethics Committee in UCLan before the sample's collection.
- 2- Collection around 500 blood samples from unrelated volunteers from the population of Saudi Arabia.
- 3- Extract DNA from all samples using QIAamp DNA Mini Kit (Qiagen), after evaluating of modifications applied to the extraction protocol.
- 4- Estimation of the concentration of the extracted DNA using Qubit® dsDNA HS Kit (Invitrogen).
- 5- Validate the use of half volume reactions and of using 50 cm capillary with POP6 before processing the 500 samples.
- 6- Amplifying DNA extracts of the 500 samples using half volume reactions.
- 7- Using the ABI 3500 DNA Genetic Analyser (AB) for separation and detection of PCR products.
- 8- Analysing the raw data using GeneMapper™ ID-X Software v1.2 (AB) and transport the results using the export option.
- 9- Evaluating the data for HWE, LD and other forensic statistical parameters.
- 10- Carry out AMOVA analysis to estimate the inbreeding coefficient (F_{IS}) and compare the results with previous studies in the population of Saudi Arabia.
- 11- Carry out the population comparison tests to compare the data from this study to other published data for the Saudi population and neighbouring countries.

3.4 Materials and Methods

All materials and methods used in this chapter are detailed in Sections 2.2, 2.3 and 2.7.

3.5 Results and discussion

3.5.1 Ethical approval

Initially, the sample collection was approved by the Security Forces Hospitals Programme (Saudi Arabia) (Appendix 3, Section 10.3.1). Following on from this, the UCLan Ethics Committee has granted an approval for the proposed application 'Forensically Relevant Polymorphisms (STRs/SNPs) in the population of Saudi Arabia', and was given the reference number of STEMH 557. The approval was granted in 28th October 2016 for five years or to the end of the project (Appendix 3, Section 10.3.2).

3.5.2 Sample collection

A total of 500 blood samples was collected from unrelated individuals (116 Females and 384 Males) from the population of Saudi Arabia (Figure 3.1). Every donor confirmed that to the best of her/his knowledge and belief there were no relatives working or involved in a training programme at any of the six branches of the Security Forces Hospitals. The Security Forces Hospitals Programme are military hospitals established to serve military bases in those cities and staff or trainees are all Saudi citizens and are offspring of Saudi parents. This allowed more confidence regarding the origin of the collected samples.

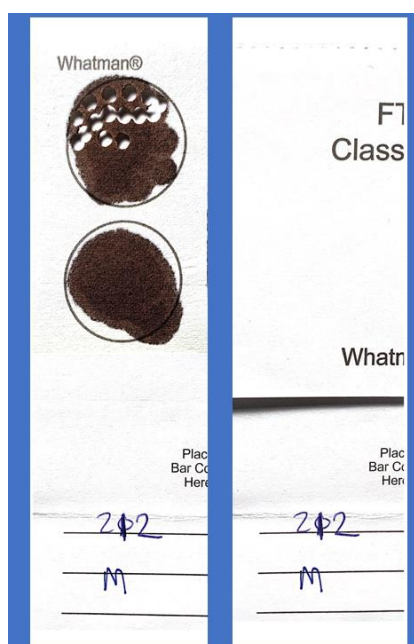


Figure 3.1. An example of an FTA card used for sample collection. Each card contained the sample number and the donor sex (F/M). The same number was printed in the consent form. A series of 2 mm punches can be seen in the upper blood spot.

Riyadh and Dammam had the highest number of participants, which may have been due the opportunity to give a presentation about the project to the staff; Tabuk had the lowest number. Table 3.1 shows the total number of samples that collected from each city.

Table 3.1. The number of participants per each city. This table is showing samples numbers collected from each city where Riyadh and Dammam cities had the highest number of samples.

City	Riyadh	Dammam	Abha	Makkah	Al-Madinah	Tabuk
Samples No.	158	120	82	102	31	7

Defining the “sufficient” number of samples can represent a population was addressed by Chakraborty (1992) who concluded that 100-150 may be adequate for statistical evaluation. However, the latest guidelines for the publication of genetic population data in the forensic science international (FSI): genetics defined 500 samples as the minimum required number for publication of autosomal markers detected through capillary electrophoresis (Gusmão *et al.* 2017), and thus this number was the

target number of collected samples. Although using blood samples on FTA cards is considered as an invasive procedure, it was practical as a method to collect material and has been demonstrated to preserve the DNA quality and quantity of samples for 16 years (Rahikainen *et al.* 2016).

3.5.3 DNA extraction

Two modifications were applied to the manufacturer's protocol (Qiagen 2016). The effect of these modification was assessed using 1% agarose gel with five samples before proceeding to extract the remaining 495 samples (Figure 3.2).

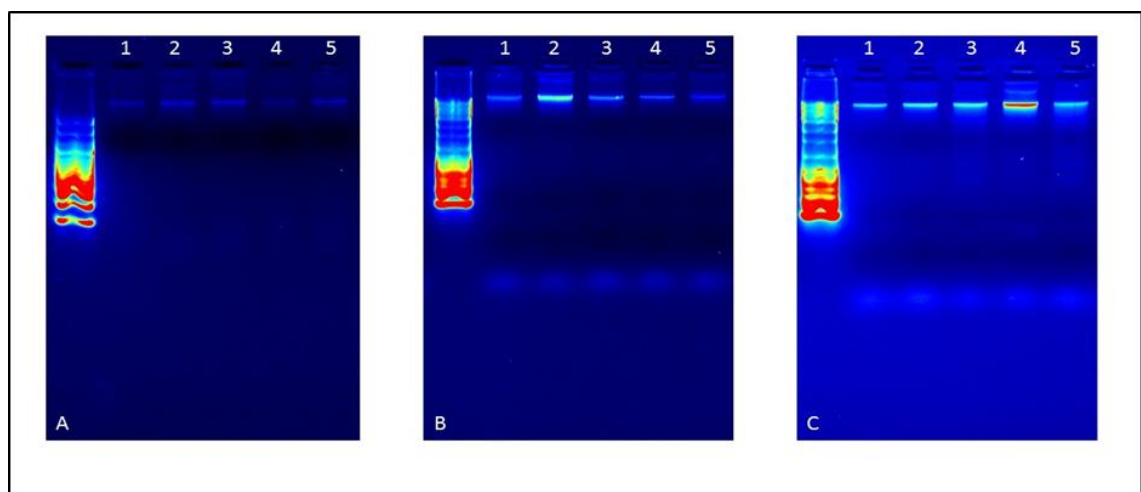


Figure 3.2. Extracted DNA run on agarose gels (1%) (A, B, and C). A) Shows results of the 5 samples following manufacturer's protocol. B) Shows results of the same samples after applying the 6 hours/overnight incubation in the ATL buffer before starting the manufacturer's protocol. C) Shows results of the 5 samples when using a volume of 100 μ l of the AE buffer for the elution stage rather than 150 μ l, in addition to the first modification.

Despite the time consumed in the extraction, DNA yield was increased by adopting the modifications. The first step of the original procedure was an incubation at 85 °C for 10 min in the ATL buffer; however, this failed to release all the blood components from the FTA punches, which still showed staining. Although the DNA contents are in the white blood cell, this still an indication that the overall blood contents were not released from the paper. The FTA punches became whiter when they were incubated in the ATL

buffer for at least 6 hours at 56 °C, and this correlated with a higher yield of DNA (Figure 3.2). Additionally, using 100 µl of AE buffer for elution was also correlated with a higher yield of DNA (Figure 3.2) which, theoretically, increases the concentration by one-third.

Due to the number of samples that can be tested by the Qubit® dsDNA HS Kit (500 samples/pack), the improvement in the DNA yield was not measured by the quantification kit.

3.5.4 Quantification of the extracted DNA

While pipetting and reaction time are critical when using the Qubit dsDNA HS Kit (Invitrogen 2019), the calibration sample was used to monitor the batches of assays. In addition, variations in the reaction time between the first tube and the last tube (53 tubes/batch) was reduced by reversing the workflow of the assay.

The calibration sample, which had a known concentration of 1.92 ng/µl, measured between 1.43 and 1.90 ng/µl in the ten extraction batches. This allowed more confidence in the concentrations estimates of the 500 samples. DNA extracts from the 500 samples showed an average concentration of 1.5 ng/µl that ranged from 0.07 - 13.5 ng/µl (Figure 3.3).

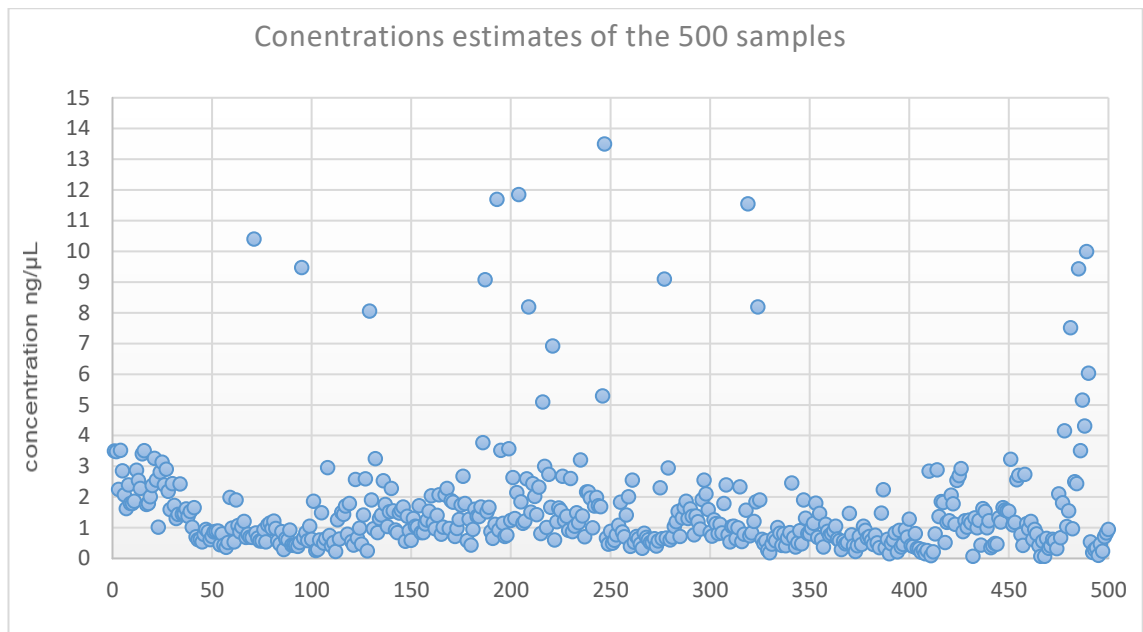


Figure 3.3. The average concentration of each DNA samples extracted. The average of the samples concentrations was 1.5 ng/μl that ranged from 0.07 - 13.5 ng/μl. Each dot represents the average of the two reading/sample.

3.5.5 Validation of half volume reaction and the 50 cm capillary with POP6.

Prior to genotyping the 500 samples, the use of half volume reaction and the 50 cm capillary were validated. Three replicates of the positive control were amplified using the manufacturer’s guidelines and using half volume (6 reactions in total). Both reaction volumes showed a full profile from 0.5 ng DNA; however, the half volume was less balanced at TH01 and D2S1338 in all replicates (Figure 3.4). In addition, based on the user guide of the manufacturer (Applied Biosystems 2016), the kit uses 36 cm capillary and POP4 while 50 cm capillary and POP6 were used in this study. Increasing the run time to 3800 s allowed detection amplicons up to 480 bp that included the designated area for largest locus (SE33) that allowed the local Southern method to be used (Figure 3.4).

Based on the validation, the 500 samples were then successfully genotyped using the half volume reaction and the 50 cm capillary with POP6.

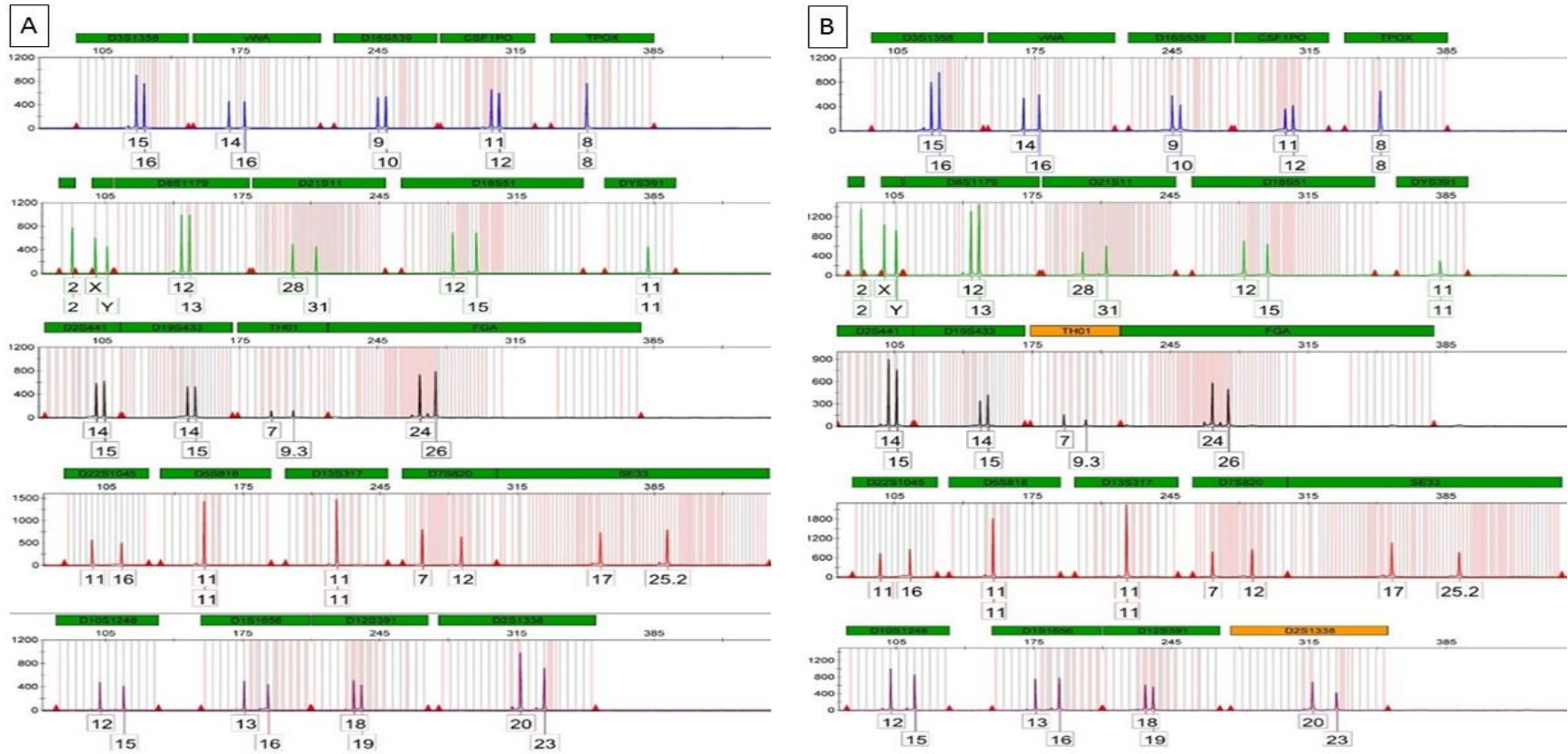


Figure 3.4. Internal validation of half volume reaction and 50 cm capillary/ POP6. The figure shows one of the replicates of two Globalfiler™ profiles for the positive control using the manufacturer's protocol (A) and the half volume reaction (B). Due to the use of 50 cm capillary, the run time was increased to 3800 s which was sufficient to cover the designated area of the largest locus SE33 that allowed the local Southern method to be used.

3.5.6 Allelic ladder variants

After analysing the 500 samples, eight allelic ladder variants were detected at SE33: allele 7.3 (10 samples), allele 13.3 (two samples), allele 17.2 (one sample), allele 22 (8 samples), allele 23 (3 samples), allele 28 (one sample), allele 33 (two samples) and allele 34 (5 samples). All these have already been observed and had been reported (size-based and sequence-based alleles were reported) in STRBase (Ruitberg *et al.* 2001). Two variants were also detected at the D1S1656: allele 7 (one sample), allele 8 (one sample) where the size-based alleles were reported (no sequence data available) in STRBase (Ruitberg *et al.* 2001) (Figure 3.5) (Alsafiah *et al.* 2017).

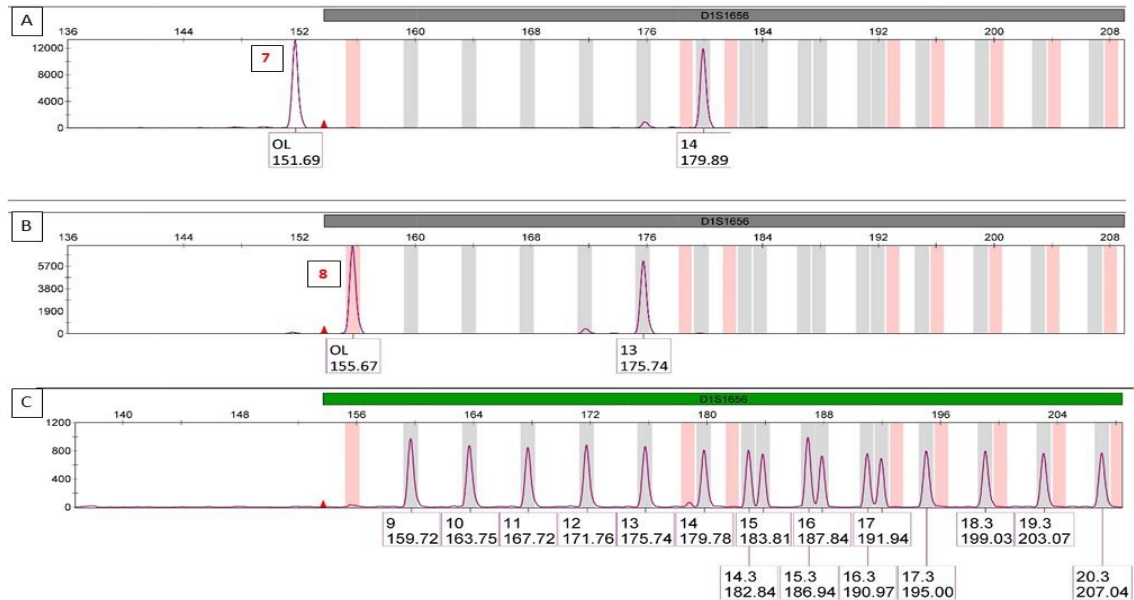


Figure 3.5. Allele variants of 7 and 8 at D1S1656. The figure shows allele 7 (A) and 8 (B) at the D1S1656 locus and the allelic ladder (C). Allele 7 is located outside the designated area of the locus. Both alleles are not represented in the allelic ladder of the Globalfiler™ PCR amplification kit. The alleles were reported in STRBase (Ruitberg *et al.* 2001) but no sequence data was available (Alsafiah *et al.* 2017).

Non-reported variants were also detected in SE33: allele 14.3 (one sample), allele 20.3 (one sample) and allele 38 (one sample) (Figure 3.6) (Alsafiah *et al.* 2017).

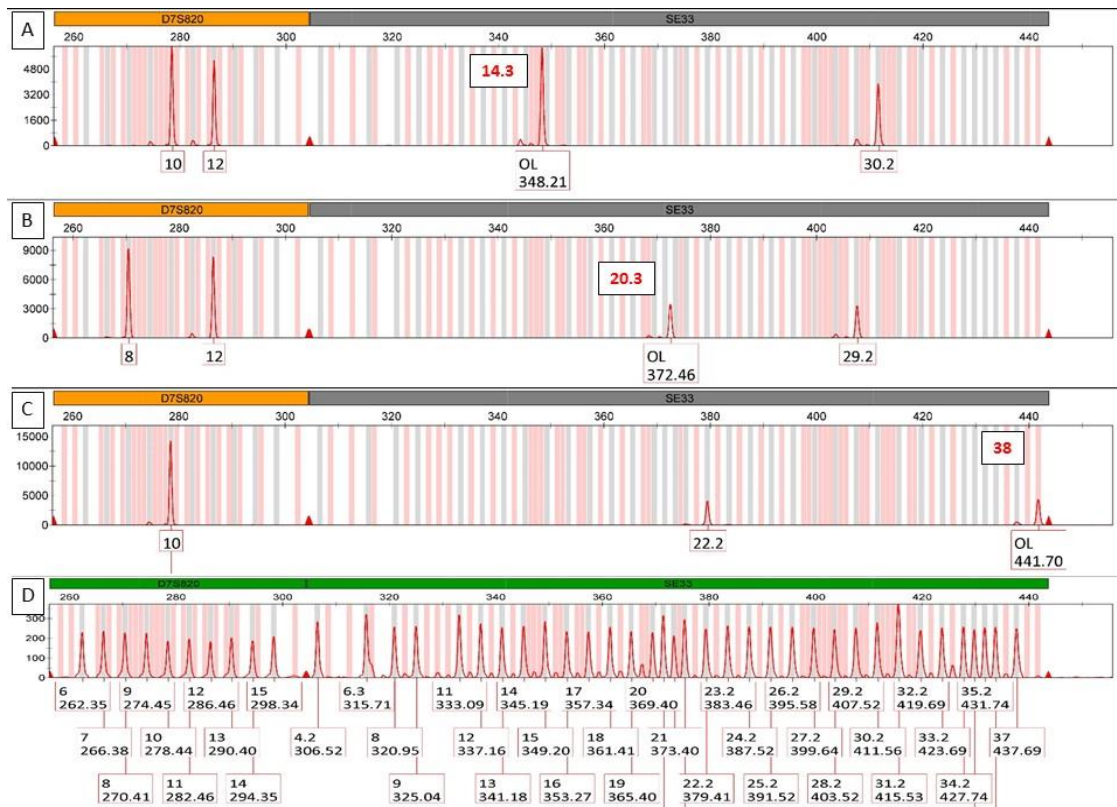


Figure 3.6. Non-reported Allele variants at SE33. The figure shows alleles 14.3 (A), 20.3 (B), and 38 (C) at SE33 and the allelic ladder (D). The alleles have not been reported before in the STRBase (Ruitberg *et al.* 2001) and are not represented in the allelic ladder of the Globalfiler™ PCR amplification kit (Alsafiah *et al.* 2017).

Interestingly, one sample showed three alleles (9, 12, OL) at D7S820, and showed homozygous allele (16) at SE33 (Figure 3.7 A). This suggests that the OL allele either belongs to D7S820 forming a triplet allele phenomenon, or is an unusual short allele belonging to SE33 forming a heterozygote genotype (OL, 16). To resolve this the D7S820 was genotyped using the PowerPlex®21 System (Promega Corporation) following the manufacturer’s guidelines, and gave only two alleles (9, 12) (Figure 3.7 B). This demonstrated that the OL allele belonged to the SE33 locus and because of the adjacent locations of D7S820 and SE33 in the GlobalFiler® PCR amplification kit, the OL allele appeared within the allelic window of D7S820 (Alsafiah *et al.* 2017).

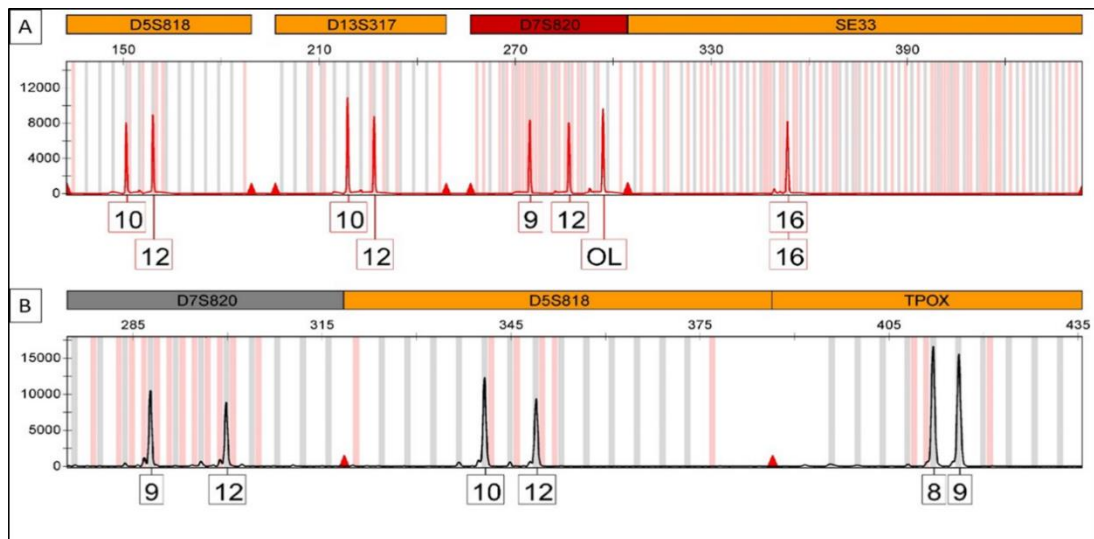


Figure 3.7. Two electropherograms (A & B) for the same sample using two different STR kits. (A) shows the genotype of D7S820 locus (9, 12, OL) using the GlobalFiler® PCR amplification kit. (B) shows the genotype of the same locus (9,12) using PowerPlex® 21 System. This confirmed that the OL allele belonged to SE33 and the OL allele appeared within the allelic window of D7S820 because of the adjacent locations of the D7S820 and SE33 in the GlobalFiler® PCR amplification kit (Alsafiah *et al.* 2017).

Based on the sizes of the OL allele (296.85 bp) and allele 4.2 (306.55 bp) in the allelic ladder (Figure 3.8), the OL allele was designated as allele 2, which had not been reported in STRBase (Ruitberg *et al.* 2001). Therefore, the genotype of this sample at the SE33 was designated (2, 16) rather than (16, 16). However, the stutter artefact of the OL allele (Figure 3.8) suggested that the allele has more than two repeats and a deletion in the flanking region may lead to the reduced size allele (Alsafiah *et al.* 2017).

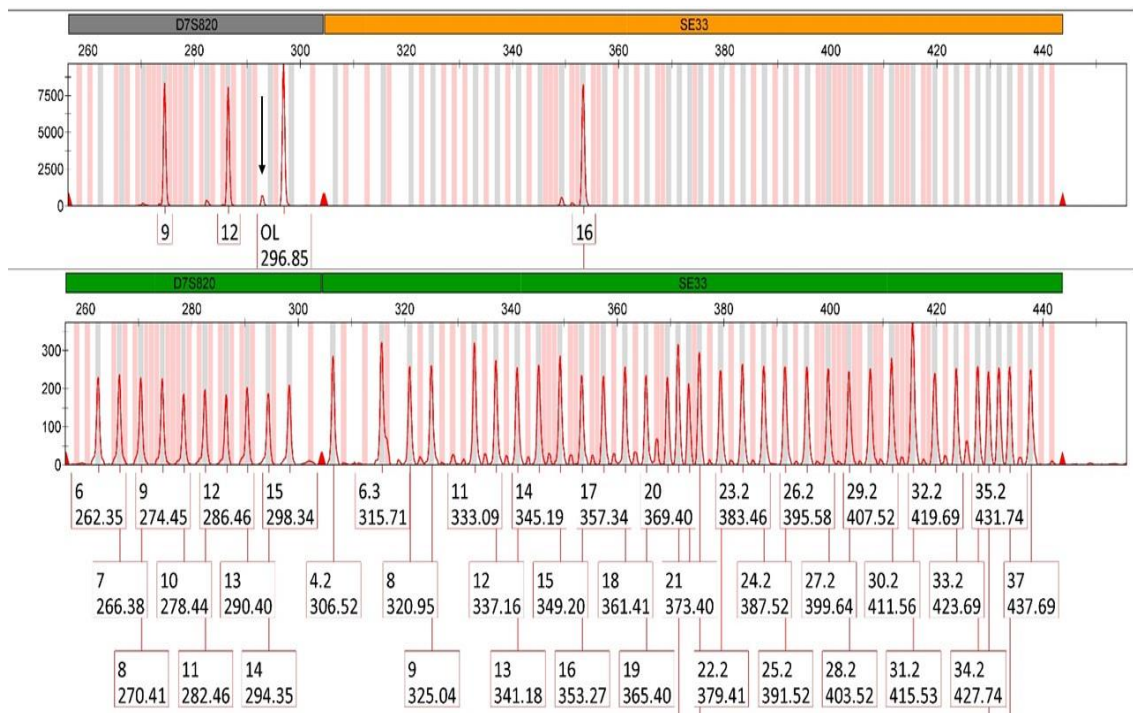


Figure 3.8. The OL allele at the SE33 and the allelic ladder of the GlobalFiler® PCR amplification kit. The figure shows the size of the OL allele comparing to the allelic ladder. As it had been confirmed that the OL allele belonged to the SE33 locus, it was possible to calculate the repeat numbers based on the sizes of the OL allele and the nearest allele in the allelic ladder (4.2); the OL was called as allele 2 (size-based call). The black arrow points to stutter artefact of the OL allele (Alsafiah *et al.* 2017).

Alleles outside the designated area of a locus or that are not represented by allelic ladder of a kit can be misinterpreted. More information about these alleles was gathered by sequencing that is addressed in Chapter 4.

3.5.7 Population genetics

Although the D18S51, D2S441, D22S1045, D7S820 and the SE33 have shown deviation from the Hardy–Weinberg equilibrium (HWE) (P value < 0.05), no significant deviation was detected after applying Bonferroni correction (P value < 0.002). The observed heterozygosity ranged from 0.660 in the TPOX to 0.914 in the SE33 (Table 3.2) (Alsafiah *et al.* 2017).

Table 3.2. Results of expected heterozygosity calculation and of Hardy-Weinberg equilibrium exact test, conducted by Arlequin v3.5.2.1 software for the 21 STR loci. The *P* values after Bonferroni correction is significant if $P < 0.002$. The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/21$ STRs = 0.002.

Locus	Alleles No	Observed Heterozygosity	Expected Heterozygosity	Exact test <i>P</i> value	Standard Deviation	Steps done
D3S1358	1000	0.73800	0.76693	0.071	0.00018	1001000
vWA	1000	0.74400	0.77924	0.286	0.00032	1001000
D16S539	1000	0.76400	0.76466	0.320	0.00029	1001000
CSF1PO	1000	0.70200	0.73067	0.239	0.00037	1001000
TPOX	1000	0.66000	0.65855	0.144	0.00031	1001000
D8S1179	1000	0.81200	0.83246	0.135	0.00029	1001000
D21S11	1000	0.79400	0.82356	0.057	0.00012	1001000
D18S51	1000	0.82400	0.87074	0.038	0.00016	1001000
D2S441	1000	0.70400	0.76151	0.042	0.00018	1001000
D19S433	1000	0.83600	0.87290	0.234	0.00036	1001000
TH01	1000	0.72200	0.76965	0.089	0.00025	1001000
FGA	1000	0.83400	0.86542	0.698	0.00018	1001000
D22S1045	1000	0.67000	0.69240	0.006	0.00006	1001000
D5S818	1000	0.73800	0.75465	0.989	0.00012	1001000
D13S317	1000	0.73800	0.75901	0.922	0.00027	1001000
D7S820	1000	0.76400	0.78855	0.011	0.00006	1001000
SE33	1000	0.91400	0.95122	0.045	0.00010	1001000
D10S1248	1000	0.71400	0.74877	0.747	0.00029	1001000
D12S391	1000	0.84600	0.87630	0.058	0.00018	1001000
D2S1338	1000	0.87400	0.88513	0.163	0.00027	1001000
D1S1656	1000	0.82200	0.85961	0.068	0.00020	1001000

The potential linkage of five syntenic pairs, three of which located in the same arm, was tested. The linkage disequilibrium (LD) test showed that all syntenic pairs has a *P* value > 0.05 (Table 3.3). Therefore, based on this population sample, the product rule can be used when calculating the frequencies of STR profiles with these 21 aSTRs.

Table 3.3. Results of linkage disequilibrium tests of syntenic loci included the GlobalFiler® PCR amplification kit. The results showed that the tested samples did not show linkage disequilibrium (*P* value > 0.05). (p-q) indicates syntenic loci located in different arms.

Chromosome	Syntenic Pair	LD test <i>P</i> value
Chr.2	TPOX	0.97764
	D2S441	
Chr.2 (p-q)	D2S1338	0.99164
	D2S441	
Chr.2 (p-q)	TPOX	0.79338
	D2S1338	
Chr.5	D5S818	0.69008
	CSF1PO	
Chr.12	vWA	0.89307
	D12S391	

The 21 STRs have a combined match probability (CMP) of 1.42091E-26, a combined power of discrimination (PoD) of 0.999999999999999999999999999986 and a combined power of exclusion of 0.999997405. Most of the 21 STRs (18/21) show ≥ 0.9 PoD: SE33 was the most informative locus with 0.993 PoD and TPOX was the least informative locus with 0.84 PoD. Allele ranges varied from 6 alleles in TH01 to 44 alleles in SE33. Some alleles show very high frequencies in the Saudi population; for example, allele 8 in the TPOX and allele 15 in the D22S1045 displayed the highest frequencies of 0.520 and 0.463 respectively (Table 3.4) (Alsafiah *et al.* 2017).

Table 3.4. Allele frequency data and forensic statistical parameters of 21 aSTRs included in GlobalFiler® PCR amplification kit for the population of Saudi Arabia. The parameters included: matching probability, power of discrimination, polymorphism information content, power of exclusion, observed homozygosity and observed heterozygosity that were generated using the PowerStat v 1.2 (Alsafiah *et al.* 2017).

Allele	D3S1358	VWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01
6					0.007						0.317
7				0.003	0.003						0.179
8			0.028	0.013	0.520						0.098
9			0.148	0.028	0.173	0.003			0.006		0.274
9.3											0.117
10			0.085	0.300	0.109	0.055		0.007	0.127		0.015
10.2								0.001			
11			0.382	0.303	0.172	0.121		0.020	0.332	0.015	
11.3									0.068		
12			0.202	0.290	0.016	0.156		0.145	0.091	0.097	
12.2								0.001		0.005	
13	0.002	0.003	0.139	0.056		0.224		0.226	0.017	0.187	
13.2										0.047	
13.3									0.001		
14	0.063	0.035	0.011	0.007		0.181		0.132	0.313	0.194	
14.2										0.065	
15	0.249	0.122	0.005			0.203		0.120	0.041	0.129	
15.2								0.002		0.118	
16	0.284	0.296				0.046		0.111	0.004	0.082	
16.2								0.002		0.042	
17	0.272	0.267				0.010		0.087		0.005	
17.1						0.001					
17.2										0.014	
18	0.114	0.207						0.002			
18.2	0.001							0.059			
19	0.015	0.060						0.042			
20		0.008						0.022			
21		0.002						0.007			
22								0.007			
23								0.004			
24								0.003			
27							0.013				
28							0.135				
29							0.253				
30							0.259				
30.2							0.010				
31							0.053				
31.2							0.077				
32							0.005				
32.1							0.001				
32.2							0.128				
33							0.002				
33.2							0.050				
34							0.001				
34.2							0.003				
35							0.003				
35.2							0.001				
36							0.002				
37							0.002				
38							0.002				
Number of alleles	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Matching Probability	0.091	0.082	0.087	0.122	0.160	0.051	0.055	0.031	0.091	0.030	0.085
Expressed as 1 in ...	10.988	12.169	11.471	8.190	6.237	19.785	18.019	32.027	10.983	33.069	11.754
Power of Discrimination	0.909	0.918	0.913	0.878	0.840	0.949	0.945	0.969	0.909	0.970	0.915
Polymorphic Information Content (PIC)	0.727	0.744	0.732	0.681	0.616	0.809	0.801	0.857	0.726	0.859	0.732
Power of Exclusion	0.489	0.500	0.534	0.431	0.369	0.621	0.588	0.644	0.434	0.667	0.463
Observed Homozygosity	0.262	0.256	0.236	0.298	0.340	0.188	0.206	0.176	0.296	0.164	0.278
Observed Heterozygosity	0.738	0.744	0.764	0.702	0.660	0.812	0.794	0.824	0.704	0.836	0.722

Table 3.4. continued.

Allele	FGA	D22S1045	D5S818	D13S317	D7S820	SE33	D10S1248	D151656	D12S391	D2S1338
2						0.001				
6			0.001		0.001					
6.3						0.007				
7				0.002	0.010			0.001		
7.3					0.001	0.010				
8			0.016	0.112	0.183	0.001		0.001		
9			0.073	0.046	0.117	0.006	0.009			
9.3					0.003					
10		0.006	0.090	0.059	0.322			0.004		
11		0.120	0.298	0.220	0.204		0.007	0.052		
12		0.018	0.335	0.400	0.139	0.006	0.045	0.142		
13		0.003	0.177	0.114	0.018	0.015	0.144	0.100		
13.3						0.002				
14	0.001	0.063	0.007	0.047		0.060	0.371	0.104	0.005	0.005
14.3						0.001		0.001		
15		0.463	0.003			0.038	0.285	0.161	0.021	
15.3								0.041		
16		0.268				0.059	0.090	0.209	0.009	0.072
16.1	0.001									
16.3								0.061		
17		0.058			0.001	0.077	0.046	0.069	0.135	0.226
17.1					0.001					
17.2						0.001				
17.3								0.029		
18	0.007	0.001				0.095	0.003	0.003	0.186	0.102
18.3								0.015	0.006	
19	0.081					0.068			0.117	0.119
19.1									0.001	
19.3								0.006	0.004	
20	0.091					0.028			0.102	0.217
20.2						0.007				
20.3						0.001		0.001		
21	0.149					0.015			0.079	0.056
21.2	0.002					0.034				
22	0.136					0.008			0.097	0.034
22.2	0.005					0.010				
23	0.174					0.003			0.129	0.070
23.2	0.002					0.034				
24	0.198								0.070	0.049
24.2	0.001					0.037				
25	0.093								0.032	0.041
25.2						0.023				
26	0.040								0.004	0.004
26.2						0.049				
27	0.005								0.003	0.004
27.2						0.057				
28	0.007					0.001				0.001
28.2						0.043				
29	0.004									
29.2						0.031				
30	0.002									
30.2						0.053				
31.2						0.057				
32.2						0.032				
33						0.002				
33.1										
33.2						0.012				
34						0.005				
34.2						0.005				
35						0.001				
35.2						0.002				
36						0.002				
38						0.001				
48.2	0.001									
Number of alleles	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Matching Probability	0.033	0.138	0.098	0.087	0.076	0.007	0.098	0.030	0.026	0.035
Expressed as 1 in ...	29.926	7.268	10.161	11.495	13.092	150.421	10.253	33.557	37.948	28.210
Power of Discrimination	0.967	0.862	0.902	0.913	0.924	0.993	0.902	0.970	0.974	0.965
Polymorphic Information Content (PIC)	0.850	0.648	0.715	0.728	0.757	0.948	0.711	0.863	0.873	0.844
Power of Exclusion	0.664	0.383	0.489	0.489	0.534	0.824	0.450	0.687	0.743	0.640
Observed Homozygosity	0.166	0.330	0.262	0.262	0.236	0.086	0.286	0.154	0.126	0.178
Observed Heterozygosity	0.834	0.670	0.738	0.738	0.764	0.914	0.714	0.846	0.874	0.822
Combined Match Probability (CMP)	1.42091E-26									
Combined Power of Exclusion (CPE)	0.999997405									
Combined Power of Discrimination (CPD)	0.99999999999999999999999999999986									

To assess the GlobalFiler™ kit for kinship testing, a typical paternity case (an alleged father, a child and a known mother) was assumed, and the combined typical paternity index (CPI) was used to calculate the paternity probabilities with different prior probabilities (Pr): 0.90, 0.50 and 0.10. The probabilities of paternity were 99.99999974% (Pr = 0.90), 99.99999765% (Pr = 0.50) and 99.99997886% (Pr = 0.10), which was expected, are much higher (~ 300 fold) than those probabilities calculated when using the currently used kit in Saudi Arabia (Identifiler® Plus) (Table 3.5).

Table 3.5. An assessment of the 21 loci included in the GlobalFiler™ kit for kinship testing. This table shows the paternity probabilities for a typical paternity case by using combined typical paternity index for different prior probabilities (Pr = 0.90, 0.50 and 0.10). The GlobalFiler™ kit showed much higher (~300-fold) probabilities comparing to those probabilities calculated when using the currently used kit in Saudi Arabia (Identifiler® Plus).

kit	CPI	Paternity probabilities (%)		
		Pr = 0.90	Pr = 0.50	Pr = 0.10
GlobalFiler™	42,569,026.49	99.99999974	99.99999765	99.99997886
Identifiler® Plus	126,843.32	99.9999124	99.99921163	99.99290514

The data of the 500 samples were analysed by DNA tools package and R studio platform to define the maximum number of matched loci between any two DNA Globalfiler™ profiles. The result showed that, within the 500 samples, two sample pairs matched in 9/21 loci (42.8%), which was the maximum number of matched loci (Table 3.6). On the other hand, one pair had partial match (i.e. one of the two alleles) in 19/21 loci (Table 3.6).

Table 3.6. The maximum matching loci within the 500 samples. In the 500 samples, only two pairs of samples showed full matching in 9 loci (i.e. both alleles); this was the maximum number of matched loci (shaded row). One pair showed partial matching (i.e. one of the two alleles) at 19 out of 22 loci (shaded column). This table was generated by the R studio using the package of DNA tools.

		No. of partial match per any sample pair																					
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
No. of matched loci per any sample pair	0	0	0	3	21	76	233	651	1433	2539	3793	4587	4467	3685	2483	1399	595	199	67	14	1	0	0
	1	0	0	7	58	174	644	1638	3300	5291	6768	7681	6841	5016	2947	1347	493	143	26	5	0	0	
	2	0	3	13	52	226	709	1724	3206	4978	5967	5798	4590	2831	1401	619	165	40	4	1	0	0	
	3	0	1	11	60	183	535	1144	2012	2710	2988	2487	1752	1023	426	132	28	3	0	0	0	0	
	4	0	0	5	36	84	259	546	875	1012	907	761	437	190	81	20	3	1	1				
	5	0	1	4	13	39	105	177	249	260	202	139	61	25	11	0	0	0					
	6	0	0	0	5	19	38	52	53	45	31	12	9	6	0	0	0						
	7	0	0	0	2	5	5	7	4	3	1	1	0	1	0	0							
	8	0	0	0	0	0	0	1	1	0	0	1	0	0	0								
	9	0	0	0	0	0	0	1	0	0	0	1	0	0									
	10	0	0	0	0	0	0	0	0	0	0	0	0										
	11	0	0	0	0	0	0	0	0	0	0	0											
	12	0	0	0	0	0	0	0	0	0	0												
	13	0	0	0	0	0	0	0	0	0													
	14	0	0	0	0	0	0	0	0														
	15	0	0	0	0	0	0	0															
	16	0	0	0	0	0	0																
	17	0	0	0	0	0																	
	18	0	0	0	0																		
	19	0	0	0																			
	20	0	0																				
	21	0																					

3.5.8 Consanguinity in the population of Saudi Arabia.

The level of consanguinity in the population of Saudi Arabia was found to be 56.8% - 54.3% (El-Hazmi *et al.* 1995, Wong and Anokute 1990), that is similar to those levels in neighbouring countries like UAE, Kuwait, Iraq, Jordan and Egypt, but is significantly higher than Europeans, Eastern Asians, Americans and Africans (El-Hazmi *et al.* 1995). This was supported by an inbreeding coefficient (F_{IS}) value of 0.024 overall the population of Saudi Arabia (El-Hazmi *et al.* 1995). In addition, the most recent study in the population of Saudi Arabia (Khubrani *et al.* 2019a), which studied the same 21 aSTRs investigated here, found that 20/21 aSTRs (D10S1248 was the exception) showed deficiency of heterozygotes with 0.0476 inbreeding coefficient (F_{IS}).

The current data set also showed deficiency of heterozygotes in 20/21 aSTRs (Table 3.2), but TPOX was the exception. The AMOVA analysis was carried out to estimate the inbreeding coefficient (F_{IS}) that had an F_{IS} value of 0.03560. Although higher F_{IS} value could be an evidence of the presence of null alleles, none of the aSTRs showed significant deviation from HWE (Table 3.2). In addition, the results of this study are in line with previous studies conducted either by questionnaires (El-Hazmi *et al.* 1995, Wong and Anokute 1990) or by aSTRs analysis (Khubrani *et al.* 2019a) showing an evidence of consanguinity in the population of Saudi Arabia.

3.5.9 Population comparison

The comparison included previous studies in the population of Saudi Arabia (Sinha *et al.* 1999, Osman *et al.* 2015, Khubrani *et al.* 2019a), Saudi individuals residing in Kuwait and in Dubai (Al-Enizi *et al.* 2013, Alshamali *et al.* 2005). Populations from Gulf Cooperation Council (GCC) countries: Kuwait (Al-Enizi *et al.* 2013), United Arab Emirates (Jones *et al.* 2017), Qatar (Perez-Miranda *et al.* 2006), Yemeni and Omani populations

residing in Dubai (Alshamali *et al.* 2005) were included. Other populations such as Egyptian, Iraqi, Iranian, and Indian residing in Kuwait (Al-Enizi *et al.* 2013) were also included in the comparison.

After the Bonferroni correction, the population differentiation exact test showed that the data of the Saudi populations previously reported in Al-Enizi *et al.* (2013), Sinha *et al.* (1999), and Khubrani *et al.* 2019a were consistent with the data reported in this study, i.e. no significant pairwise differences were observed. However, the data of the Saudi population in Dubai (Alshamali *et al.* 2005) showed significant difference in the TH01 locus (P value = 0), which was due in part to the notable differences in alleles frequencies at this locus. For example, allele 7 frequency was 0.179 in the current study while it had 0.08 frequency in (Alshamali *et al.* 2005), which is over 2-fold higher. This inconsistency may be due to the small number of Saudi participants (94 samples) in this study leading to an exaggerated sampling effect. There were also significant differences with the data from the Riyadh province (Osman *et al.* 2015) at three loci (vWA, CSF1PO and TH01). Despite the relatively small sample size (190 samples), alleles 5.3, 7.3, and 8.3 at TH01 locus were observed which have not been observed in the current study or in previous studies of the Saudi population. In addition, this study found that 9 out of 15 loci had significant deviation from HWE (P value < 0.05), which was attributed to the prevalence of consanguinity in Saudi Arabia (Alsafiah *et al.* 2017). The general percentage of consanguinity in the Riyadh province is 60%, which is higher than the average rate of Saudi Arabia (56%), and is even higher (74.3%) in rural areas (El-Mouzan *et al.* 2007).

As expected, the differentiation between the data obtained in this study and the data from the Yemeni, Omani (Alshamali *et al.* 2005), Kuwaiti, Egyptian, Iraqi, Iranian, Indian

(Al-Enizi *et al.* 2013), UAE (Jones *et al.* 2017) and Qatari populations (Perez-Miranda *et al.* 2006) varies, with a general trend of more significant differences being detected as the populations become more geographically separated. For example, there was no significant difference observed between the Saudi and the Kuwaiti population whereas there were significant differences between the Indian and the Saudi populations in 13 out of the 15 loci compared (Table 3.7) (Alsafiah *et al.* 2017).

Table 3.7. Population differentiation exact test results using the Arlequin v3.5.2.1 software. Shaded cells indicate significant differences (P value < 0.002). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/21$ STRs = 0.002. N/A values indicate data that was not collected during each of the previous studies (Alsafiah *et al.* 2017).

Loci	Saudi (Khubrani <i>et al.</i> 2019a)	Saudi (Sinha <i>et al.</i> 1999)	Saudi in Dubai (Alshamali <i>et al.</i> 2005)	Saudi in Kuwait (Al-Enizi <i>et al.</i> 2013)	Saudi in Riyadh (Osman <i>et al.</i> 2015)	Yemeni (Alshamali <i>et al.</i> 2005)	Omani (Alshamali <i>et al.</i> 2005)	Kuwaiti (Al-Enizi <i>et al.</i> 2013)
D3S1358	0.34860+0.0344	N/A	0.62163+0.0142	0.48355+0.0259	0.01642+0.0029	0.46105+0.0151	0.31956+0.0158	0.84557+0.0236
vWA	0.28936+0.0309	0.06000+0.0139	0.28266+0.0087	0.09801+0.0084	0.00000+0.0000	0.00000+0.0000	0.24164+0.0177	0.00726+0.0025
D16S539	0.00923+0.0031	N/A	0.11894+0.0071	0.85052+0.0092	0.04244+0.0082	0.67981+0.0203	0.71855+0.0121	0.41396+0.0463
CSF1PO	0.00250+0.0007	0.12223+0.0158	0.41111+0.0230	0.06154+0.0133	0.00000+0.0000	0.16060+0.0191	0.22793+0.0161	0.00665+0.0027
TPOX	0.40218+0.0362	0.35619+0.0230	0.85978+0.0078	0.86633+0.0152	0.13245+0.0108	0.00000+0.0000	0.00319+0.0008	0.11572+0.0128
D8S1179	0.09649+0.0125	N/A	0.71537+0.0125	0.39880+0.0247	0.04653+0.0080	0.00262+0.0009	0.86144+0.0146	0.07594+0.0155
D21S11	0.48562+0.0339	N/A	0.72354+0.0157	0.27445+0.0249	0.05186+0.0076	0.78571+0.0137	0.01399+0.0037	0.56529+0.0407
D18S51	0.58406+0.0417	N/A	0.15631+0.0200	0.59097+0.0318	0.00706+0.0020	0.87926+0.0063	0.14291+0.0134	0.44209+0.0443
D2S441	0.59406+0.0353	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D19S433	0.73678+0.0157	N/A	N/A	0.31263+0.0260	0.03886+0.0086	N/A	N/A	0.06922+0.0113
TH01	0.35251+0.0353	0.45546+0.0215	0.00000+0.0000	0.32640+0.0147	0.00000+0.0000	0.47472+0.0154	0.00000+0.0000	0.34986+0.0297
FGA	0.41584+0.0643	N/A	0.39644+0.0267	0.03796+0.0101	0.00978+0.0081	0.07879+0.0125	0.03816+0.0048	0.03756+0.0131
D22S1045	0.25156+0.0359	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D5S818	0.13395+0.0220	N/A	0.51783+0.0184	0.95086+0.0067	0.01434+0.0036	0.08981+0.0078	0.29418+0.0204	0.46785+0.0448
D13S317	0.04443+0.0154	N/A	0.07185+0.0094	0.0023+0.0002	0.07141+0.0114	0.25500+0.0134	0.05452+0.0124	0.01935+0.0068
D7S820	0.26756+0.0400	N/A	0.66590+0.0194	0.03002+0.0102	0.38965+0.0290	0.82962+0.0165	0.15146+0.0175	0.05470+0.0142
SE33	0.05082+0.0151	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D10S1248	0.30853+0.0232	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D12S391	0.27017+0.0212	N/A	N/A	N/A	N/A	N/A	N/A	N/A
D2S1338	0.13381+0.0159	N/A	N/A	0.62703+0.0337	0.17545+0.0157	N/A	N/A	0.23134+0.0254
D1S1656	0.49862+0.0347	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 3.7. continued.

Loci	Egyptian (Al-Enizi <i>et al.</i> 2013)	Iraqi (Al-Enizi <i>et al.</i> 2013)	Iranian (Al-Enizi <i>et al.</i> 2013)	India (Al-Enizi <i>et al.</i> 2013)	Qatari (Perez-Miranda <i>et al.</i> 2006)	UAE (Jones <i>et al.</i> 2017)
D3S1358	0.16526+-0.0403	0.28366+-0.0173	0.04000+-0.0101	0.00001+-0.0000	0.04233+-0.0056	0.64109+-0.0372
vWA	0.00000+-0.0000	0.00032+-0.0001	0.00149+-0.0005	0.00000+-0.0000	0.00172+-0.0006	0.05355+-0.0127
D16S539	0.14551+-0.0246	0.00401+-0.0027	0.10082+-0.0119	0.00000+-0.0000	0.00531+-0.0021	0.22141+-0.0340
CSF1PO	0.35504+-0.0323	0.09136+-0.0403	0.02183+-0.0049	0.00000+-0.0000	0.90715+-0.0106	0.64778+-0.0198
TPOX	0.01078+-0.0051	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.07145+-0.0096	0.00018+-0.0002
D8S1179	0.00080+-0.0003	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.13486+-0.0201	0.03907+-0.0150
D21S11	0.10283+-0.0249	0.00257+-0.0023	0.01676+-0.0071	0.00000+-0.0000	0.54356+-0.0284	0.02580+-0.0092
D18S51	0.00000+-0.0000	0.00081+-0.0004	0.00015+-0.0001	0.00000+-0.0000	0.60425+-0.0411	0.12998+-0.0171
D2S441	N/A	N/A	N/A	N/A	N/A	0.00009+-0.0001
D19S433	0.00000+-0.0000	0.00000+-0.0000	0.00728+-0.0034	0.00000+-0.0000	0.00000+-0.0000	0.01495+-0.0072
TH01	0.00000+-0.0000	0.82111+-0.0244	0.02256+-0.0061	0.00742+-0.0036	0.10273+-0.0103	0.18154+-0.0524
FGA	0.00001+-0.0000	0.04025+-0.0145	0.03340+-0.0086	0.08558+-0.0260	0.12379+-0.0225	0.21628+-0.0254
D22S1045	N/A	N/A	N/A	N/A	N/A	0.01316+-0.0047
D5S818	0.01659+-0.0046	0.25700+-0.0248	0.20914+-0.0159	0.00005+-0.0001	0.00000+-0.0000	0.01271+-0.0041
D13S317	0.00000+-0.0000	0.00000+-0.0000	0.00258+-0.0016	0.00000+-0.0000	0.00012+-0.0001	0.07272+-0.0193
D7S820	0.00114+-0.0012	0.11864+-0.0510	0.00015+-0.0002	0.00000+-0.0000	0.93375+-0.0081	0.01476+-0.0079
SE33	N/A	N/A	N/A	N/A	N/A	0.00000+-0.0000
D10S1248	N/A	N/A	N/A	N/A	N/A	0.05404+-0.0122
D12S391	N/A	N/A	N/A	N/A	N/A	0.04029+-0.0100
D2S1338	0.00000+-0.0000	N/A	0.00000+-0.0000	0.00000+-0.0000	0.11449+-0.0135	0.00000+-0.0000
D1S1656	N/A	N/A	N/A	N/A	N/A	0.00108+-0.0005

The F_{ST} values of the 13 STRs, which are common with the previous studies (Perez-Miranda *et al.* 2006, Jones *et al.* 2017, Al-Enizi *et al.* 2013, Alshamali *et al.* 2005, Osman *et al.* 2015, Khubrani *et al.* 2019a), was also calculated. The `cmdscale` function was used in R studio software to generate a multi-dimensional scale (MDS) for the average of F_{ST} values (Figure 3.9).

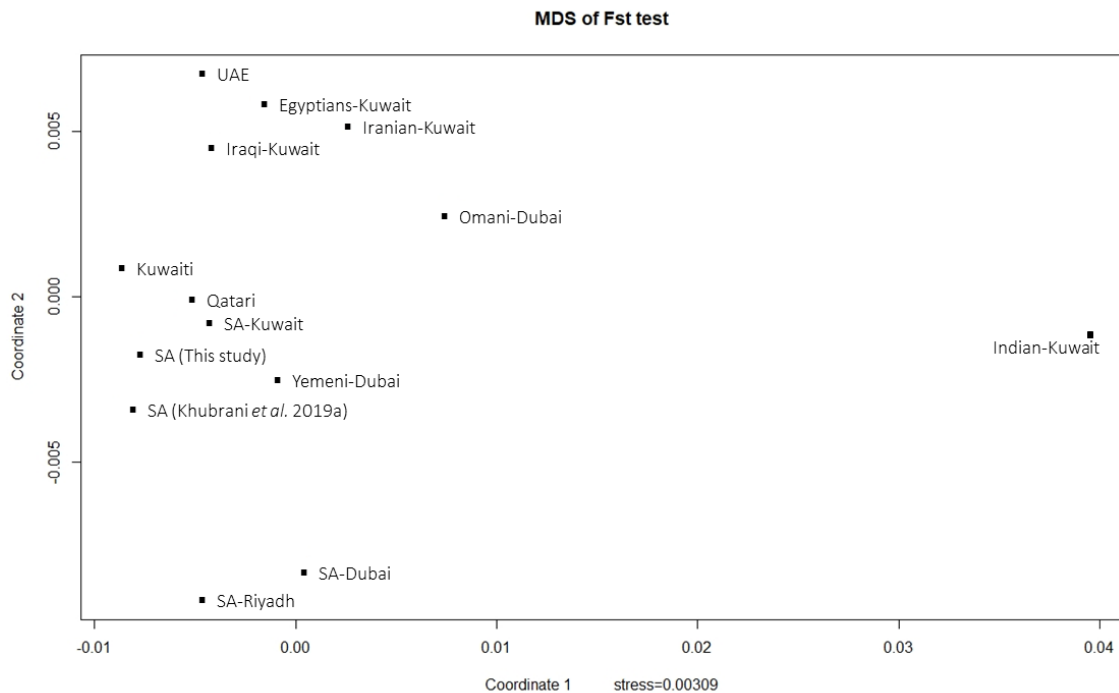


Figure 3.9. A multi-dimensional scale (MDS) for the average F_{ST} values of 13 common loci. Fourteen populations were included in the comparison: Saudi Arabian (this study), Saudi Arabian (Khubrani *et al.* 2019a), Saudi Arabian population in Riyadh (Osman *et al.* 2015), Saudi Arabian in Dubai (Alshamali *et al.* 2005), Saudi Arabian in Kuwait (Al-Enizi *et al.* 2013), Qatari (Perez-Miranda *et al.* 2006), UAE population (Jones *et al.* 2017), Kuwaiti (Al-Enizi *et al.* 2013), Omani-Dubai (Alshamali *et al.* 2005), Yemeni-Dubai (Alshamali *et al.* 2005), Iraqi-Kuwait (Al-Enizi *et al.* 2013), Egyptians-Kuwait (Al-Enizi *et al.* 2013), Iranian-Kuwait (Al-Enizi *et al.* 2013), and Indian-Kuwait (Al-Enizi *et al.* 2013). Note: the data of Saudi population in (Sinha *et al.* 1999) was not included in the F_{ST} test due to the limited number of common loci included in the study (four loci). SA: Saudi Arabian and UAE: United Arab Emirates. The `cmdscale` function was used in R software to generate a multi-dimensional scale (MDS).

As presented in the MDS plot, the results showed low differentiation between this study and the data of Saudi Arabia (Khubrani *et al.* 2019a), Qatari (Perez-Miranda *et al.* 2006), Saudi in Kuwait (Al-Enizi *et al.* 2013) and Kuwaiti (Al-Enizi *et al.* 2013). The greatest differentiation was observed in the Indian population in Kuwait (Al-Enizi *et al.* 2013),

which was included in the comparison as a control. In addition, the MDS plot confirms the exact test results for the Saudi population in Dubai and from the Riyadh city showing less similarity to the data generated from this study or from other studies in Saudi population (Khubrani *et al.* 2019a, Al-Enizi *et al.* 2013).

3.6 Conclusion

An ethical approval was granted for the project and 500 samples from unrelated volunteers (as far as could be ascertained) in Saudi Arabia were collected. High quality DNA was obtained from the samples after evaluating two modifications applied to the manufacturer's DNA extraction protocol. The quantities of the extracted DNA were as expected from blood samples and adequate for downstream applications.

The 500 samples were genotyped using the Globalfiler™ PCR amplification kit. This was accomplished after validation of the half PCR volume and of using 50 cm capillary/POP6. Although, the half volume achieved full profile with 0.5 ng DNA, the profile was less balanced than the manufacturer's protocol. The data of the 21 aSTRs were obtained and were evaluated for human identification applications in Saudi Arabia. Three of the additional STR loci (SE33, D12S391, and D1S1656) in this kit are more informative than any locus in the currently used kit (Identifiler® Plus). The kit provided a much higher discrimination power, by which CMP improved from $2.23E-18$ to $1.42E-26$ and the combined typical paternity index increased by 300-fold demonstrating the usefulness of adapting this kit in the forensic genetic laboratories of Saudi Arabia.

The data set examined here showed evidence of consanguinity in the population of Saudi Arabia that is supporting other studies either those conducted either by questionnaires or by aSTRs analysis.

Allele frequencies generated from this study can be used to estimate the profile frequencies in Saudi Arabia (Alsafiah *et al.* 2017).

The Saudi allele frequency data of the 21 aSTRs was used to measure the similarity with neighbouring populations. A general trend of more significant differences being detected as the populations become more geographically separated.

4 Chapter Four: Characterisation of STR allele variants detected in Saudi population.

4.1 Overview of experiment

Six allele variants were detected in the population of Saudi Arabia when the 500 samples were genotyped using the Globalfiler™ PCR amplification kit (AB) (Chapter 3). Four SE33 allele variants of 2 (Figure 3.8), 14.3, 20.3 and 38 (Figure 3.6) had not been reported in STRBase (STRBase 2017b) and two alleles 7 and 8, at D1S1656 (Figure 3.5), were reported in STRBase (STRBase 2017a), but no sequence data was available. Both STRs are within the three most informative loci for the population of Saudi Arabia (Alsafiah *et al.* 2018).

SE33 is the most polymorphic well-characterised STR that is commonly used in forensic genetics (Wiegand *et al.* 1993). The landscape of this locus is divided to three regions: repeat region, “local flanks”, and extended flanks (Borsuk *et al.* 2018). The sequence structure of the repeat region is based on tetra-nucleotides repeats of [(CTTT)_n] (forward strand) and the complexity of the structure increases as alleles become larger (Moller and Brinkmann 1994, Rolf *et al.* 1997) (Table 4.1 A). A recent study has classified the SE33 repeat motifs to 34 types (A0, A1, A2.... to D3) based on the structure of the repeat and the local flank regions, eleven of which were observed >1% of the tested populations (Borsuk *et al.* 2018) (Table 4.2) (Alsafiah *et al.* 2018).

The D1S1656 has a compound repeat structure of [(TAGA)_n (TAGG)] followed by [(TG)₅] in the 3′-local flank (Table 4.1 B). This locus was added to the European Standard Set (ESS) in 2006 (Gill *et al.* 2006) and to the Combined DNA Index System (CODIS) in 2015 (Hares 2015) (Alsafiah *et al.* 2018).

Based on the sequence-based nomenclature guidelines of the International Society for Forensic Genetics (ISFG), the local flank regions showed in Table 4.1 A and B, are not counted in allele calling system (Parson *et al.* 2016).

Table 4.1. Sequence structure of the SE33 and D1S1656 loci. (A) Shows the sequence structure of an SE33 allele that comprised of the 5'-local flank (15 bp), repeat region, and 3'-local flank (24 bp). (B) A typical sequence structure of a D1S1656 allele is shown. The sequence structure of reference alleles (SE33, allele 26.2 GenBank: V00481.1) and (D1S1656, allele 15.3 GenBank: G07820.1) are given for illustration. Based on the published guidelines of the International Society for Forensic Genetics (ISFG) (Parson *et al.* 2016), the local flank regions showed in A and B (greyed out sequences) are not counted in allele calling system (Alsafiah *et al.* 2018).

A. SE33 locus			
Reference allele	5'-local flank (15 bp)	Repeat region	3'-local flank (24 bp)
26.2 GenBank: V00481.1 (forward strand)	CT CTTT CTTT CCTT C	CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT TT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT CTTT	CT CTTT CTTT CTTT CT CTTT CTTT
B. D1S1656 locus			
Reference allele	Repeat region	3'-local flank (10 bp)	
15.3 GenBank: G07820.1	TAGA TAGA TAGA TAGA TGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGA TAGG	TGTGTGTGTG	

Table 4.2. Classification of the SE33 motifs. The table shows the eleven motif patterns that had >1% frequency in the tested populations. Most sequence-based alleles show A0 and A1 patterns (Borsuk *et al.* 2018).

Allele range	Motif ID	Motif
7 to 23	A0	CT [CTTT]3 C [CTTT]n CT [CTTT]3 CT [CTTT]2
19.2 to 33.2	A1	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2
15 to 23	A2	CT [CTTT]2 CCTT C [CTTT]n CT [CTTT]3 CT [CTTT]2
22.2 to 30.2	A3	CT [CTTT]2 CCTT C [CTTT]n CT [CTTT]n CT [CTTT]3 CT [CTTT]2
21.2 to 31.2	A4	CT [CTTT]2 [CCTT]2 C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2
16 to 23	A5	CT [CTTT]3 CCTT C [CTTT]n CT [CTTT]3 CT [CTTT]2
10.2 to 15.2	A6	CT [CTTT]3 C[CTTT]n [CTTT]3 CT [CTTT]2
30 to 36	A7	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]2
27.2 to 34.2	A8	CT [CTTT]2 CCTT C [CTTT]n TT [CTTT]n CT [CTTT]3 CT CTTT
15 to 20	A9	CT [CTTT]3 CCCTT [CTTT]n CT [CTTT]3 CT [CTTT]2
26.2 to 32.2	B0	CT [CTTT]2 [CCTT]2 C [CTTT]n CT [CTTT]3 CT [CTTT]2

Although SE33 is included in the primer mixes of the ForenSeq™ DNA Signature Prep (Verogen) (Novroski *et al.* 2016), it is not reported by the ForenSeq™ UAS (Verogen). This may be due to the high dropout rate that was observed when analysing the ForenSeq™ data using an independent software (Borsuk *et al.* 2018). The highly repetitive sequence of the extended flanking regions makes the size of amplicons large, which reduces the read quality (Gettings *et al.* 2015). In addition, thymine and cytosine represent more than 80% of the forward strand of the SE33 amplicons that makes sequencing more challenging (Borsuk *et al.* 2018). In contrast, the D1S1656 is already included and is reported in Precision ID GlobalFiler MPS Panel (AB) and in the ForenSeq™ DNA Signature Prep (Verogen) (Alsafiah *et al.* 2018), for example (Guo *et al.* 2017, Wang *et al.* 2017).

During a training course “Illumina Forensic Genomics Workshop, 14th – 15th November 2017 (Cambridge, UK), each participant could bring two samples to be sequenced and analysed using the Verogen system. Therefore, this chapter describes the characterisation of the D1S1656 variants using ForenSeq™ DNA Signature Prep (Verogen) and the characterisation the SE33 variants using the conventional Sanger sequencing.

4.2 Aims of the study

The aim of this chapter is to characterise the allele variants detected in the population of Saudi Arabia when the 500 samples were genotyped using the Globalfiler™ PCR amplification kit (AB), to allow more information about their sequence structure. This include a confirmation of that a deletion in the flanking region was the reason for the observation of allele 2 at SE33, which was suggested based on the presence of the

stutter artefact. Finally, reporting the sequence data of the alleles to be added to the STRBase database.

4.3 Objectives

- 1- Sequencing D1S1656 allele variants using the ForenSeq™ DNA Signature Prep kit and MiSeq FGx™ System, which was followed by an analysis using the ForenSeq™ UAS (Verogen).
- 2- Sequencing the SE33 allele variants using the BigDye™ Terminator v3.1 Cycle Sequencing Kit (AB). This included PCR amplification for the SE33 STR using a published primer set, loading the samples in agarose gel for the physical separation of the target band and Sanger sequencing.
- 3- Report the sequence structure of the six variants to STRBase.

4.4 Materials and Methods

The lab work in this chapter comprised of two different sequencing systems. Sequencing D1S1656 allele variants using the NGS part is described in Section 2.6 and the sequencing SE33 allele variants is described in Sections 2.4.1 and 2.8.2

4.5 Results and discussion

4.5.1 SE33 variants

The SE33 locus was successfully amplified from samples that exhibited the alleles 2, 14.3, 20.3 and 38. The physical separation of the alleles was successfully achieved when using the 20-cm-long agarose gel (Figure 4.1). DNA recovery from the targeted bands using PureLink™ Quick Gel Extraction Kit (AB) yielded adequate concentrations (0.25 to 0.78 ng/μl) to achieve successful direct sequencing for all alleles (Figure 4.2) (Alsafiah *et al.* 2018).

Based on size-based system, allele 2 could be due to a complete loss of the repeat region or due to sequence deletion within the flanking regions. However, the presence of a stutter artefact, which is associated with the repetitive regions, suggested sequence deletion in the flanking regions (Figure 3.8). Sequence data revealed that the allele 2 had B1 motif pattern and, as expected, consisted of 17 repeats in the repeat region with a 60 bp deletion (GRCh38, 6:88277290-88277349) in the extended 3'-flank (Table 4.3) (Alsafiah *et al.* 2018). This deletion was previously observed with alleles 14 and 16 (Hering *et al.* 2006, Lederer *et al.* 2008).

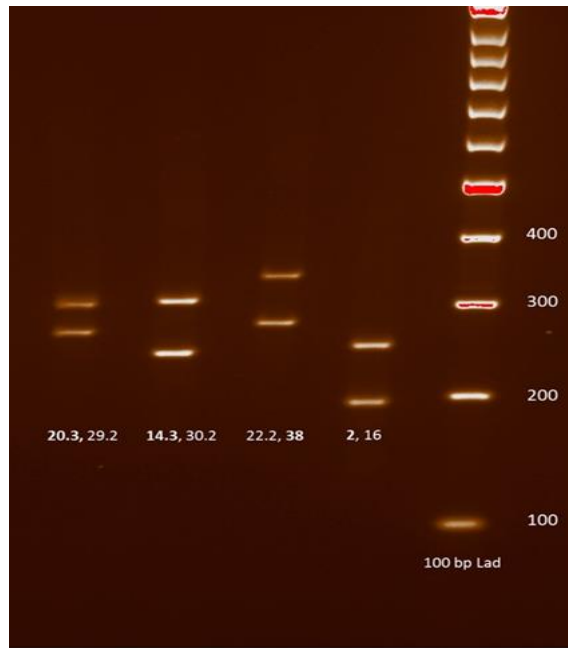


Figure 4.1. A 20-cm-long 3% agarose gel for the novel SE33 alleles; from the left side, alleles 20.3, 14.3, 38, 2 and a 100 bp ladder. It shows the separation of alleles 20.3 and 29.2 (35 bp) that could not be achieved with a shorter (10 cm) gel.

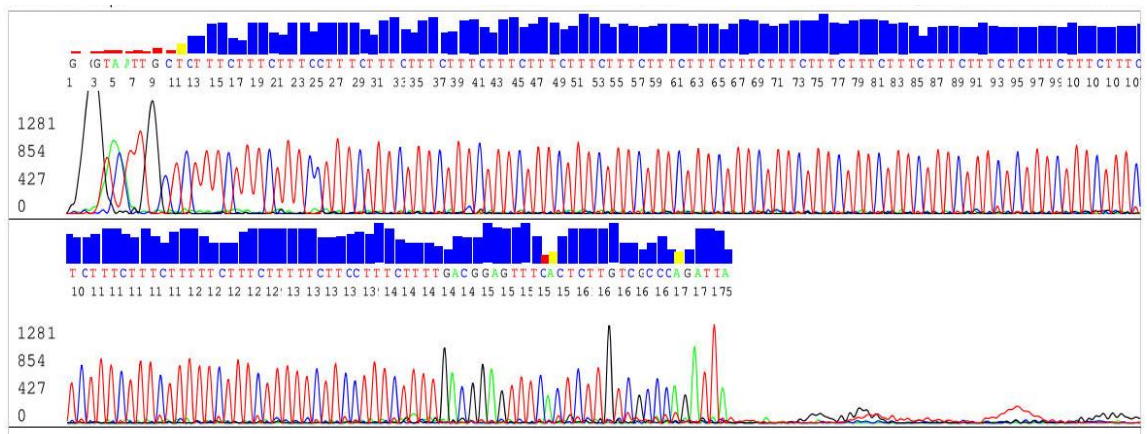


Figure 4.2. An example of the quality of sequencing results. This figure shows an electropherogram for the forward strand of allele 2 at the SE33 locus.

Table 4.3. Sequence data for the forward strand of 4 previously uncharacterized SE33 alleles: 2, 14.3, 20.3, and 38. The 5' uncounted sequence (15 bp) and 3' uncounted sequence (24 bp) of the local flank region; and the extended 3'-flank are shown. The amplicons sizes of the GlobalFiler kit and of primer pair (SE33-1 and SE33-2) used in this study are shown. It also shows allele names based on their sizes, based on the sequence data, and the motif pattern based on the classification of Borsuk *et al.*, (2018). Allele 2 had B1 motif and showed 17 repeats on the repeat region, but a 60 bp deletion in the extended 3'-flank led to the observation of the allele 2 based on the size. Allele 14.3 had 18.1 repeats on the repeat region, and the 14.3 size-based allele resulted from a 14 bp deletion in the extended 3'-flank. In addition, the motif pattern of this allele is novel (has not been reported in the classification of Borsuk *et al.* (2018)). The Allele 20.3 had D1 motif and showed a TTT within the repeat region. Allele 38 contained two hexanucleotide repeats (A7 motif) within the repeat region. ^(a) Represents rs1045867314 SNP at Location 6:88277260 in the allele 14.3 (Alsafiah *et al.* 2018).

Allele name (size-based system)	GlobalFiler Sizes (bp)	Amplicon sizes (bp) using SE33-1 and SE33-2	Allele name (sequence-based system)	Motif pattern	5'-Local flank (15 bp)				Repeat region						3'-Local flank (24 bp)				Extended 3' flank		
					CT	CTT	CCT	C	CTT	C	TTT	CTTTT	CTT	CTTTT	CTT	CT	CTT	CT		TTT	CTT
2	296.85	193	17	B1	1	2	0	1	17	0	0	0	0	0	0	1	3	1	0	2	CTTTT CTTT CTTTT C TTC <60 bp del> [CTTT] ₂ TGAC GGAG TT
14.3	348.21	244	18.1	Novel	1	2	0	1	3	1	0	0	15	0	0	1	3	0	1	2	<14 bp del> TT C [TTCC] ₃ TTT [CT] ₆ [CTTT] ₃ CTA [CT] ₂ CTTT GTCT [CTTT] ₄ TGAC GGAG TT
20.3	372.46	268	20.3	D1	1	2	1	1	9	0	1	0	11	0	0	1	3	1	0	2	CTTTT CTTT CTTTT C [TTCC] ₃ TTT [CT] ₆ [CTTT] ₃ CTA [CT] ₂ CTTT GTCT [CTTT] ₄ TGAC GGAG TT
38	441.70	337	38	A7	1	2	1	1	9	0	0	1	12	1	14	1	3	1	0	2	CTTTT CTTT CTTTT C [TTCC] ₃ TTT [CT] ₆ [CTTT] ₃ CTA [CT] ₂ CTTT GTCT [CTTT] ₄ TGAC GGAG TT

Allele 14.3 had a novel motif pattern of *CT CTTT₂ C CTTT_n C CTTT_n CT CTTT₃ TT CTTT₂* (italic bases are within the local flanks, underlined base is C>T variant), which have not reported in the classification of Borsuk *et al.* (2018). Although the allele had 18.1 repeats in the counted region, a 14 bp deletion (GRCh38, 6:88277269-88277283) in the extended 3'-flank led to the observation of allele 14.3, based on size (Table 4.3). In addition, the T variant at location 6:88277260 (GRCh38) in the 3'-local flank represents rs1045867314 SNP (C: > 99%, T: < 1%) (Auton *et al.* 2015) (Table 4.3) (Alsafiah *et al.* 2018).

Allele 20.3 had D1 motif pattern (Borsuk *et al.* 2018), and showed three T nucleotides within the repeat sequence that could have occurred due to a C deletion in a single repeat or due to an insertion of three T nucleotides (Table 4.3) (Alsafiah *et al.* 2018).

Allele 38 showed a A7 motif pattern that exhibits two hexanucleotide repeats within the repeat region (Table 4.3) (Alsafiah *et al.* 2018).

4.5.2 D1S1656 variants

Samples that showed alleles 7 and 8 at the D1S1656 were successfully sequenced using the ForenSeq DNA Signature Prep Kit (Primer Mix B) and the MiSeq FGx Forensic System. Both samples showed 100% concordance at 21 autosomal STRs and DYS391 loci overlapped with the GlobalFiler® PCR amplification kit (Alsafiah *et al.* 2018).

Allele 7 showed a typical sequence structure of TAGA₆ TAGG₁ TG₅ (Figure 4.3 A). However, allele 8 showed TAGA₈ TG₅ sequence where the TAGG repeat was absent (Figure 4.3 B) (Alsafiah *et al.* 2018). This absence was previously reported in (Kline *et al.* 2011, Gettings *et al.* 2016), which could be interpreted by the presence of an A variant of rs78443572 SNP (TAGG, G: 73%, A: 27%) (Auton *et al.* 2015).

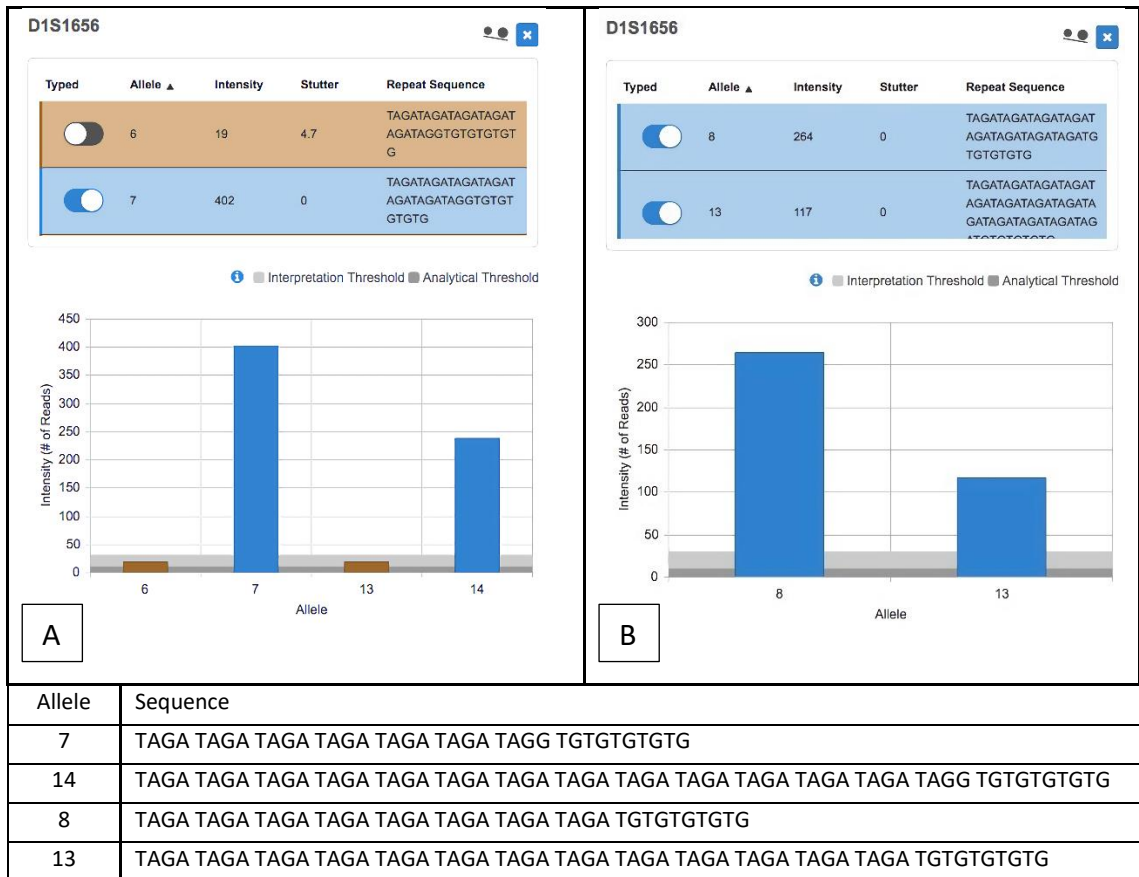


Figure 4.3. Sequencing data of the reverse strand of the alleles 7 and 8 at the D1S1656 locus. This data was generated using ForenSeq™ DNA Signature Prep (Primer Mix B) and MiSeq FGx System (Verogen). (A) Shows the sequence data of allele 7; (B) Shows the sequence data of allele 8. Due to the presence of the A variant of rs78443572 SNP (TAGG, G: 73%, A: 27%) in the alleles 8 and 13, these alleles ended with TAGA rather than TAGG (Alsafiah *et al.* 2018).

As the two samples were sequenced using Primer Mix B that includes 56 aiSNPs and 22 piSNPs, the ForenSeq™ UAS estimated biogeographical ancestry and predicted two phenotypic features (hair and eye colours). The software uses the principal component analysis (PCA) to estimate the biogeographical ancestry. Any sample can be classified to three main populations European, East Asian, and African and when a sample does not fit with any of the three populations, it will be assigned as ad-Mixed Americans (Verogen 2018b). In addition, the software uses the HirisPlex model (Walsh *et al.* 2014) (<https://hirisplex.erasmusmc.nl/>) to predict the eye and the hair colour (Verogen 2018b).

Both samples were within the ad-Mixed Americans classification, one of which showed more similarity to the European main population (Figure 4.4). The software calculates the estimation based on the main populations of the Phase I of the 1000 Genomes project (Verogen 2018b), and may be when subpopulation groups are added in the future phases, the estimation will be more specific.

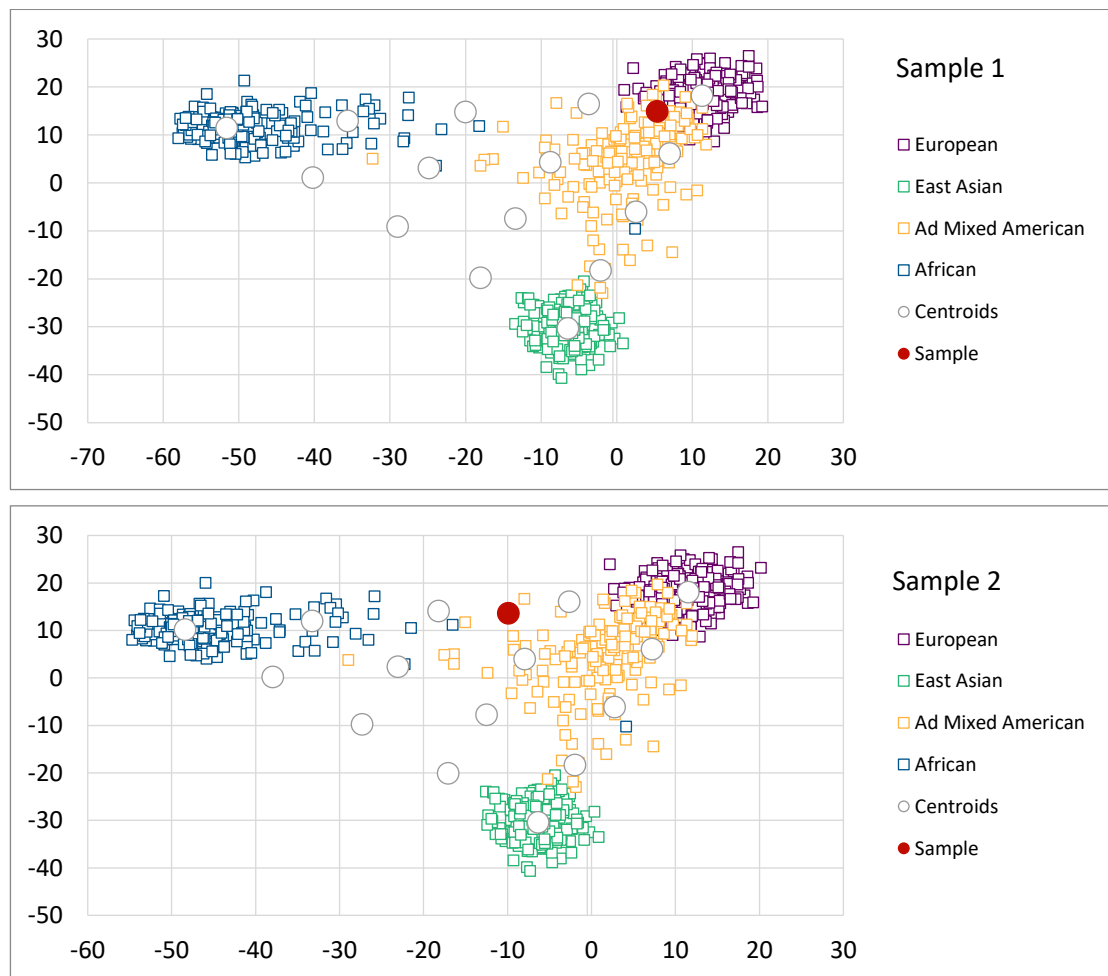


Figure 4.4. An estimation of the biogeographical ancestry. The figure shows the result of the biogeographical ancestry estimation of the two samples using the piSNPs and aiSNPs included in the Primer Mix B of the ForenSeq™ DNA Signature Prep kit. Both samples were classified as ad-Mixed Americans, one of which was more like the European main population.

The samples showed a higher probability of having brown or black hair, which are more likely in the population of Saudi Arabia than red hair. The eye colour was also predicted and both samples showed very high percentage of 94% and 100% that the source of the samples have brown eyes (Table 4.4).

Table 4.4. The results of hair and eye colour prediction. Both samples showed high probabilities of having brown eyes and brown or black hair. These features are more likely in the population of Saudi Arabia.

Sample 1		Sample 2	
Hair Colour Results		Hair Colour Results	
Brown	0.45	Brown	0.29
Red	0.00	Red	0.00
Black	0.46	Black	0.71
Blond	0.09	Blond	0.01
Eye Colour Results		Eye Colour Results	
Intermediate	0.05	Intermediate	0.00
Brown	0.94	Brown	1.00
Blue	0.01	Blue	0.00

4.6 Conclusion

Six allele variants, at SE33 (four) and D1S1656 (two), were detected in the population of Saudi Arabia when the 500 samples were genotyped using the Globalfiler™ PCR amplification kit (AB). The D1S1656 allele variants were sequenced using the ForenSeq™ DNA Signature Prep kit and MiSeq FGx™ System and analysed by the ForenSeq™ UAS, while the SE33 allele variants were sequenced using the conventional sequencing assay (Sanger sequencing).

This study has provided sequence data for six previously uncharacterized alleles at SE33 and D1S1656 loci. The SE33 alleles 2, 20.3 and 38 had B1, D1 and A7 motif pattern respectively, while allele 14.3 had a novel motif pattern. In addition, based on the sequence-based nomenclature guidelines of the ISFG, the alleles 2 and 14.3 at the SE33 should be called 17 and 18.1 respectively. The study confirmed the assumption that allele 2 at the SE33 was due to deletion in a flanking region (Alsafiah *et al.* 2018). The observation of alleles outside the designated windows of an allelic ladder may lead to misinterpretation of this allele that was resolved by analysing the sequence structure (Alsafiah *et al.* 2018).

The sequence data of all alleles were reported to STRBase and are now included in STRBase database (see https://strbase.nist.gov/var_SE33.htm and https://strbase.nist.gov/var_D1S1656.htm).

5 Chapter Five: An evaluation of 17 non-CODIS STRs for the population of Saudi Arabia using the SureID® 23comp Human Identification Kit.

5.1 Overview of experiment

The extended number of STR markers required for the CODIS and for the ESS (Gill *et al.* 2006, Hares 2015), has led to the development of GlobalFiler™ PCR Amplification Kit, VeriFiler™ Plus PCR Amplification Kit (AB), PowerPlex® Fusion 6C system (Promega Corporation) and Investigator® 24plex (Qiagen) (Table 1.1). The information obtained from these kits will be sufficient in most kinship cases; however, it is still possible to have inconclusive results in complex cases (Goodwin *et al.* 2004). Kinship testing can be further complicated when the level of consanguinity in the target population is relatively high (Phillips *et al.* 2012), or when the family pedigree is deficient (Poetsch *et al.* 2013), (Alsafiah *et al.* 2019a).

As mentioned in Section 1.5.2, It has been demonstrated that additional STRs can increase the power of genetic testing in determining the true relation between parent-child, sibling or half sibling (O'Connor *et al.* 2010). For example, Carboni *et al.* (2014) described four complex cases, including incest, which were inconclusive using 13-15 STRs, but that could be resolved using 39-41 STRs.

As most loci are shared between the commonly used kits, the maximum number of aSTRs that can be tested, when combining any two kits, is 24 STRs (e.g. VeriFiler™ Plus and PowerPlex® Fusion 6C), which necessitates the use of a supplementary STR kit when more loci need to be tested. A set of 25 supplementary STRs (26-plex including amelogenin) was suggested by the National Institute of Standards and Technology (NIST, USA) to increase the certainty in kinship testing (Hill *et al.* 2009) (Table 5.1); however, no multiplex combining these STRs is commercially available (Alsafiah *et al.* 2019a).

Table 5.1. Supplementary autosomal STRs included in 3 supplementary autosomal STR kits. The kits are Microreader™ 23sp ID (Li, J. *et al.* 2017) (Suzhou Microread Genetics), Goldeneye™ DNA ID 22NC (Fu *et al.* 2018) (Goldeneye® Technology Ltd.), AGCU 21+1 (Zhu *et al.* 2015) (AGCU ScienTech Incorporation). These kits are only commercially available in China (Phillips 2017). The table also shows a set of 25 supplementary STRs and amelogenin (26plex) recommended by the NIST, but no multiplex combining these STRs is commercially available.

Chr.	STRs	Microreader™23sp ID	Goldeneye™ DNA ID 22NC	AGCU 21+1	26plex (NIST)
1	D1S1656				
	F13B				
	D1S1677				
	D1S1627				
	D1GATA113				
2	D2S441				
	D2S1360				
	D2S1338				
	D2S1776				
3	D3S1744				
	D3S3045				
	D3S1358				
	D3S4529				
	D3S3053				
4	D4S2366				
	D4S2408				
	D4S2364				
5	D5S2800				
	D5S2500				
6	D6S474				
	D6S477				
	SE33 ^b				
	F13A01				
	D6S1017				
7	D7S3048				
	D7S1517				
8	D8S1132				
	D8S1115				
	LPL				
9	D9S1122				
	D9S2157				
	PentaC				
	D9S925				
10	D10S1248				
	D10S2325				
	D10S1435				
11	D11S2368				
	D11S4463				
12	D12S391				
	D12ATA63				
13	D13S325				
14	D14S1434				
	D14S608				
	D15S659				
15	FESFPS				
	PentaE				
16	D16S539				
17	D17S1301				
	D17S1290				
	D17S974				
18	D18S1364				
	D18S51				
	D18S535				
	D18S853				
19	D19S253				
	D19S433				
20	D20S482				
	D20S470				
	D20S1082				
21	D21S2055				
	PentaD				
	D21S1270				
22	D22GATA198B05				
	D22S1045				

Although, other supplementary Kits: Microreader™ 23sp ID (Li, J. *et al.* 2017) (Suzhou Microread Genetics, China), Goldeneye™ DNA ID 22NC (Fu *et al.* 2018) (Goldeneye® Technology Ltd., China), AGCU 21+1 (Zhu *et al.* 2015) (AGCU ScienTech Incorporation, China) have been developed (Table 5.1) (Alsafiah *et al.* 2019a), but they are only commercially available in China (Phillips 2017).

Massively parallel systems (MPS) allow simultaneous sequencing of multiple DNA markers. For example, Precision ID GlobalFiler™ NGS STR (Li, H. *et al.* 2017) (20 CODIS STRs and nine non-CODIS STRs) (AB), Promega PowerSeq™ Auto/Y system (Montano *et al.* 2018) (20 CODIS STRs, PentaD, PentaE, and 23 Y-STRs) (Promega Corporation), and ForenSeq™ DNA Signature Prep (Li, R. *et al.* 2019) (20 CODIS STRs, seven non-CODIS STRs, 24 Y-STRs, 7 X-STRs and 94 iSNPs) (Verogen). These can be utilised to increase the power of kinship testing. However, the systems are expensive to establish and are not yet commonly used in many laboratories (Alsafiah *et al.* 2019a).

SureID® 23 comp Human Identification kit (Health Gene Technologies, China), combines amelogenin and 22 autosomal STRs: D1S1656, D2S441, D10S1248, D12S391, D16S539 and 17 non-CODIS STRs (D3S1744, D4S2366, D5S2800, D6S474, D7S3048, D8S1132, D9S1122, D11S2368, D13S325, D14S1434, D15S659, D17S1301, D18S1346, D19S253, D20S482, D21S2055, and D22GATA198B05). Twelve of these STRs are not included in other available supplementary kits, such as Investigator® HDplex Kit (Qiagen 2012a) and PowerPlex® CS7 System (Promega Corporation 2016) (Table 5.2). The kit is now available in the UK and Europe (Alsafiah *et al.* 2019a).

Table 5.2. STR Markers included in the SureID® 23comp kit. This table shows the locations (GRCh38) and repeat structures of the 22 STRs included in the SureID® 23comp kit. Five loci are common with the CODIS and the ESS. Twelve loci are not included in other available supplementary kits (Investigator® HDplex and PowerPlex® CS7). All information was adapted from (Qiagen 2012a, Promega Corporation 2016, Phillips *et al.* 2018b) (Alsafiah *et al.* 2019a).

Chr.	SureID® 23 comp			Investigator® HDplex STRs ^(a and b)	PowerPlex® CS7 STRs
	STRs ^(a)	Location (GRCh38)	Repeat structure *		
1	D1S1656 ^a	230769616-230769683	CCTA [TCTA] _n		F13B
2	D2S441 ^a	68011947-68011994	[TCTA] _n	D2S1360	
3	D3S1744	147374752-147374828	[ATAG] _n atg [ATAG] _n at [ATAG] _n	D3S1744	
4	D4S2366	6483114-6483172	[GATA] _n [GATT] _n [GATA] _n gac [GATA] _n	D4S2366	
5	D5S2800	59403132-59403199	[GGTA] _n [GACA] _n [GATA] _n [GATT] _n	D5S2500	
6	D6S474	112557951-112558018	[AGAT] _n [GATA] _n	D6S474, SE33 ^b	F13A01
7	D7S3048	21227099-21227174	[TATC] _n [TACC] _n [CACC] _n	D7S1517	
8	D8S1132	106316692-106316774	[TCTA] _n tca [TCTA] _n	D8S1132	LPL
9	D9S1122	77073826-77073873	[TAGA] _n		PentaC
10	D10S1248 ^a	129294244-129294295	[GGAA] _n	D10S2325	
11	D11S2368	19259601-19259684	[TATC] _n [TGTC] _n [TATC] _n		
12	D12S391 ^a	12297020-12297095	[AGAT] _n [AGAC] _n AGAT	D12S391 ^a	
13	D13S325	42599304-42599382	[TCTA] _n tca [TCTA] _n		
14	D14S1434	94842054-94842105	[CTGT] _n [CTAT] _n		
15	D15S659	46081911-46081966	[TATC] _n		FESFPS, PentaE
16	D16S539 ^a	86352702-86352745	[GATA] _n		
17	D17S1301	74684855-74684902	[AGAT] _n		
18	D18S1364	65732998-65733056	[TAGA] _n TACA [TAGA] _n	D18S51 ^a	
19	D19S253	15617484-15617531	[ATCT] _n		
20	D20S482	4525692-4525747	[AGAT] _n		
21	D21S2055	39819508-39819649	[CTAT] _n CTA [CTAT] _n (30N) [TATC] _n	D21S2055	PentaD
22	D22GATA198B05	17169811-17169882	CTCT [ATCT] _n [ACCT] _n		

^a: CODIS and ESS locus

^b: Germany core locus

* Nucleotides in red are not counted in allele nomenclature system (Phillips *et al.* 2018b).

The SureID 23 comp was used to generate population genetic data for three main populations European, South Asian and African (Iyavoo *et al.* 2019), but it is believed that the kit has not been validated as no publications currently exist, either independently or from the manufacturer. Therefore, the kit should be validated using the minimum criteria for validation recommended by the European Network of Forensic Science Institutes (ENFSI) and by the scientific working group on DNA analysis Methods (SWGDM), in order to be used in forensic laboratories (Alsafiah *et al.* 2019a). The minimum criteria for validation of new kits for forensic applications includes repeatability, reproducibility, sensitivity stochastic effect, heterozygote peak balances, stutter/corresponding allele ratios, concordance with other kits for the same STRs, and performance when PCR inhibitors are present. The SWGDM guideline demand testing the precision and accuracy of the kit. Although both guidelines include mixture studies, they were not carried out as the kit is specifically designed to be used in complex kinship testing.

5.2 Aims of the study

This study aimed to carry out an internal validation of the SureID®23 comp Human Identification kit following the minimum criteria for validation recommended by the ENFSI and by the SWGDM (Section 1.4). This is to aid the forensic DNA laboratories and the manufacturer by highlighting the benefits and the drawbacks of the kit.

It also aimed to generate allele frequency data for the 17 non-CODIS loci using the 500 samples to facilitate the estimation of match probabilities of DNA profiles in Saudi Arabia and assess HWE and forensic statistical parameters and compare the data with other populations.

The kit will also be assessed for kinship testing as a supplementary STR kit in Chapter

7.

5.3 Objectives

- 1- Preparing the ABI 3500 DNA Genetic Analyser that included:
 - a. Undertake spectral calibration for the genetic analyser.
 - b. Optimisation the use of 50 cm capillaries and POP-6™ polymer.
 - c. Install the panels, bins and the analysis method to the GeneMapper™ ID-X Software v1.2 (AB).
- 2- Genotyping the 500 samples.
- 3- Analysing the raw data using the GeneMapper™ ID-X Software v1.2 (AB) and transporting the result using the export option.
- 4- Internal validation of the SureID® 23 comp kit following the ENFSI and the SWGDAM minimum criteria that included:
 - a. Confirmation of the identity of the D5 locus included in the kit (i.e. is it D5S2800 or D5S2500) by testing the 9947A control DNA.
 - b. Repeatability and reproducibility
 - c. Sensitivity and stochastic effect.
 - d. An evaluation of the kit's performance against common PCR inhibitors
 - e. Further assessment of the kit's performance using the bone samples.
 - f. Heterozygote peak balances study.
 - g. Stutter/corresponding allele ratios.
 - h. Precision and accuracy study.

- i. Concordance study of five common loci (D1S1656, D2S441, D10S1248, D12S391, D16S539) with the Globalfiler™ PCR amplification kit (AB) (Chapter 3).

5- Population genetic data for the 17 non-CODIS loci.

6- Evaluating the data for HWE, and the forensic statistical parameters.

7- Carry out the population comparison test to compare the data of the Saudi population with other published data.

8- Submit the data of the 17 non-CODIS loci to STRidER for quality control (Bodner *et al.* 2016).

5.4 Materials and Methods

All experimental work and analysis were described in Sections 2.5 and 2.7.

5.5 Results and discussion

5.5.1 Preparation ABI 3500 DNA Genetic Analyser

Before starting the validation study, the Genetic Analyser was successfully calibrated. The use of 50 cm capillaries and POP-6™ polymer was optimised by increasing the run time to 3900 s (36 cm capillaries and POP-4 uses 1,210 – 1,500 s). Increasing the run time to 3900 s allowed detection of amplicons up to 455 bp that included the designated area for all loci and at least two size markers that were larger than the largest allele allowing the local Southern method to be used. All alleles were successfully called after installing the panels, and the bins (Figure 5.1) (Alsafiah *et al.* 2019a).

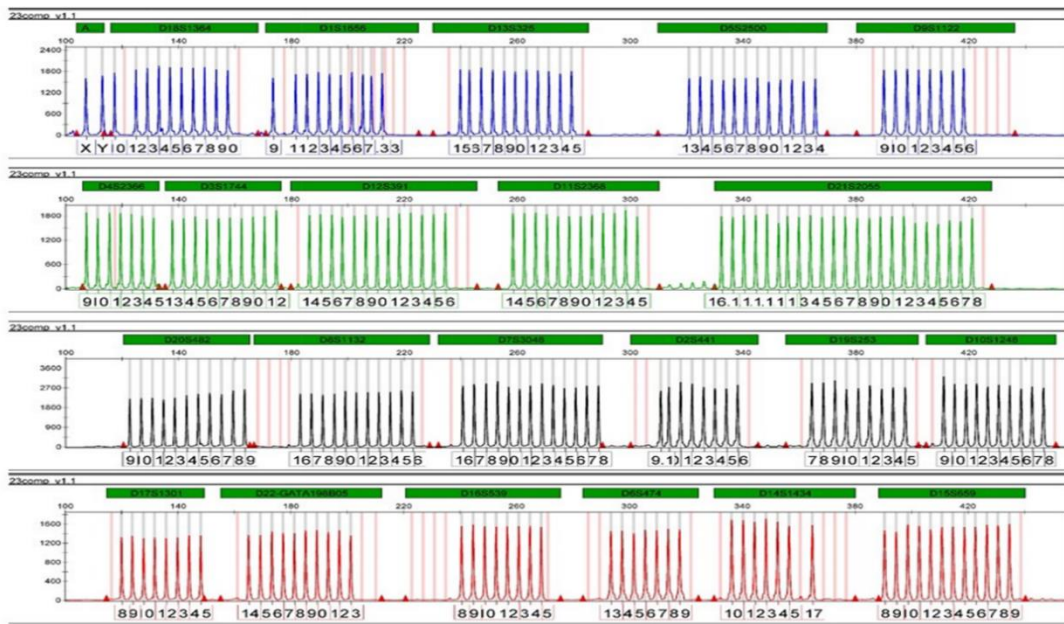


Figure 5.1. Allelic ladder of the SureID® 23comp kit. This figure shows the allelic ladder provided with the SureID® 23 comp kit. It represents 232 alleles that are supported by 53 additional bins for variant alleles (pink bins). It also shows the successful calibration and optimisation of the ABI 3500 DNA Genetic Analyser (Alsafiah *et al.* 2019a).

5.5.2 D5 locus confirmation

It is important to note that the D5 locus included in this kit was named as D5S2500 in the panels and in the supporting documents. Two different loci, which are 1643 bp apart and have different sequence structure, both had the D5S2500 name. One locus is a part of the Investigator® HDplex Kit (Qiagen) and the other one is a part of the AGCU 21-plex (AGCU ScienTech Incorporation). This duplication was detected when this locus showed different genotypes for the 9947A control DNA using both kits (15, 16 for the Investigator® HDplex and 14, 23 for the AGCU 21-plex) (Phillips *et al.* 2016). Therefore, the name of D5S2800 was proposed for the STR marker included in the AGCU 21-plex to be differentiated from the other one included in the Investigator® HDplex Kit (Phillips *et al.* 2016) (Alsafiah *et al.* 2019a).

The Health Gene Technologies has confirmed that the D5 locus included in SureID® 23comp kit is the same locus in the AGCU 21-plex (personal communication). This was

additionally confirmed by genotyping the 9947A control DNA included in the kit as a positive control that showed alleles 14, 23 (Figure 5.2). Based on this confirmation, the Health Gene Technologies has updated the name of the locus to D5S2800 in the panels and in the supporting documents (Alsafiah *et al.* 2019a).

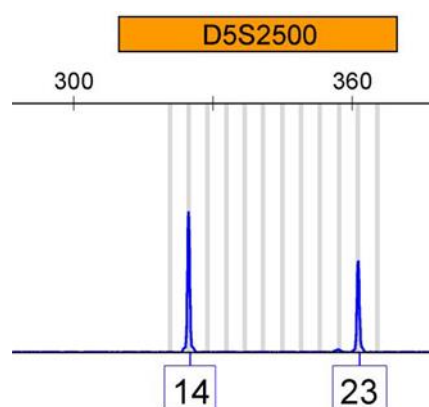


Figure 5.2. The genotype of the 9947A control DNA at the D5 locus included in the SureID® 23 comp kit. The locus had 14, 23, which is the genotype of D5S2800, confirming the correct name. The locus name is now updated by the manufacturer from D5S2500 (as shown in the locus name) to D5S2800.

5.5.3 Repeatability and Reproducibility.

In the initial tests of the SureID® 23comp, the two reaction volumes (25 and 10 µl) optimised by the manufacturer were validated by two independent operators. A total of 0.5 ng of the 2800M control DNA was amplified in 20 replicates using both reaction volumes (5 replicates per reaction volume per operator). All replicates were successfully profiled and showed full profiles that were fully concordant demonstrating repeatability and reproducibility (Alsafiah *et al.* 2019a).

5.5.4 Sensitivity stochastic effect.

The five replicates of dilution series were profiled using the 25 µl and 10 µl volumes with 28 and 30 cycles. Full profiles were generated from the 125 pg samples when using the 10 µl volume (28 and 30 cycles), while 25 µl volume was able to generate full profile

with 30 PCR cycles only. For the 62 pg samples, the 25 μ l and 10 μ l volumes with 30 cycles, allow detection of 90.24% and 95.12% of alleles, respectively. The remaining alleles were visible and could be detected with a reduced RFU threshold of 30. With 31 pg, 85.3% of alleles were detected when using the 10 μ l volume with 30 cycles, while allele dropout was observed with the 25 μ l volume (28 and 30 cycles) and with the 10 μ l volume (28 cycles) (Figure 5.3). The sensitivity results are comparable to other commonly used kits, for example, Identifiler Kit (Collins *et al.* 2004), Investigator HDplex Kit (Westen *et al.* 2012) and PowerPlex Fusion 6C System (Ensenberger *et al.* 2016) where the profile percentage ranged from 82% to 94% for the 62 pg and from 37% to 72% for the 31 pg (Alsafiah *et al.* 2019a).

Reaction volume	PCR cycles	DNA Con. (pg)	AMEL	D18S1364	D1S1656	D13S325	D5S2800	D9S1122	D4S2366	D3S1744	D12S391	D11S2368	D21S2055	D20S482	D8S1132	D7S3048	D2S441	D19S253	D10S1248	D17S1301	D22GATA19	D16S539	D6S474	D14S1434	D15S659		
25 µl	30	500	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	
		250	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		125	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		62	Green	Green	Yellow	Green	Red	Green	Yellow	Yellow	Red	Green	Red	Yellow	Yellow	Red	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Green	Green
		31	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Yellow	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
	28	500	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		250	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		125	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
		62	Red	Red	Red	Green	Red	Green	Red	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
		31	Red	Red	Red	Green	Red	Green	Red	Red	Red	Red	Red	Yellow	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red
10 µl	30	500	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	
		250	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	
		125	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		62	Green	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Red	Green	Yellow	Green	Yellow	Red	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Green
		31	Green	Green	Yellow	Green	Red	Green	Red	Red	Red	Red	Yellow	Yellow	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Red
	28	500	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		250	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		125	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
		62	Red	Red	Green	Green	Red	Green	Red	Red	Red	Red	Yellow	Red	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Green	Yellow	Red
		31	Red	Green	Yellow	Red	Red	Green	Red	Red	Red	Red	Red	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Red	Yellow

Figure 5.3. Sensitivity and stochastic tests for the SureID® 23comp kit. Serial dilutions (500, 250, 125, 62, and 31) pg were prepared from the 2800M control DNA (Promega Corporation). Each test was done in five replicates and the highest number of detected alleles are shown. Each cell represents an allele and merged cells represent homozygote loci in 2800M. Green cells identify detected alleles with ≥ 60% peak balance ratios. Yellow cells identify detected alleles with < 60% peak balance ratios. Red cells represent non-detected alleles with threshold of 50 RFU/150 RFU for heterozygotes/homozygotes (Alsafiah *et al.* 2019a).

5.5.1 Performance against common PCR inhibitors

The performance of the SureID® 23comp kit with different concentrations of two common PCR inhibitors was tested. Full profiles were generated in the presence of ≤ 120 ng/ μ l of tannic acid and of ≤ 75 ng/ μ l of humic acid (Figure 5.4-5.5 and Figure 5.6-5.7). Although these levels are similar to those reported for the SureID® PanGlobal system (Health Gene Technologies) (Liu *et al.* 2017), other commonly used kits are more robust in the presence of higher concentrations of inhibitors (Lin *et al.* 2017) (Figure 5.8) (Alsafiah *et al.* 2019a).

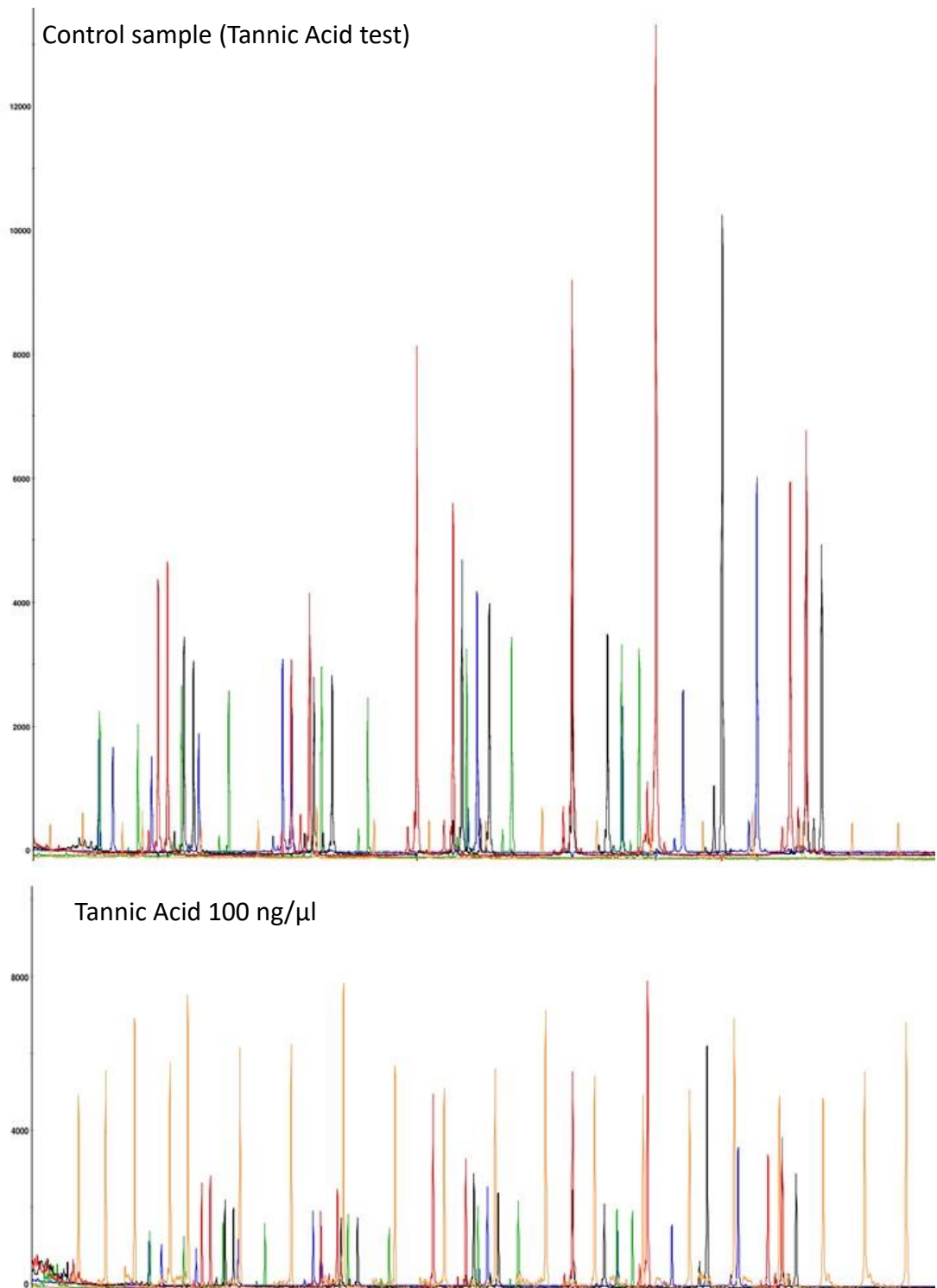


Figure 5.4. Testing of the SureID® 23comp kit with tannic acid. Three different concentrations of 100 ng/μl, 120 ng/μl and 150 ng/μl were tested. This figure shows the results of the control sample (no inhibition) and of the 100 ng/μl (tannic acid) sample. Figure 5.5 shows the results of 120 and 150 ng/μl of tannic acid.

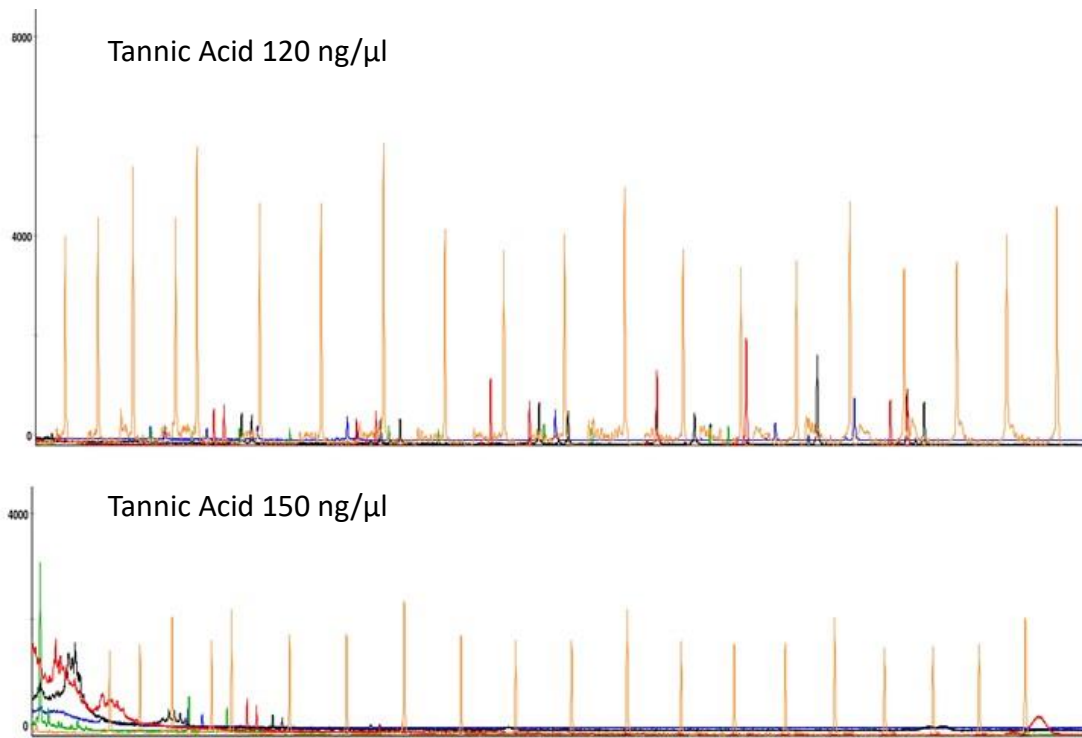


Figure 5.5. Testing of the SureID® 23comp kit with tannic acid. Three different concentrations of 100 ng/μl, 120 ng/μl and 150 ng/μl were tested. This figure shows the results of the 120 ng/μl (tannic acid) sample and of the 150 ng/μl (tannic acid) sample. Full profiles were achieved with ≤ 120 ng/μl of tannic acid.

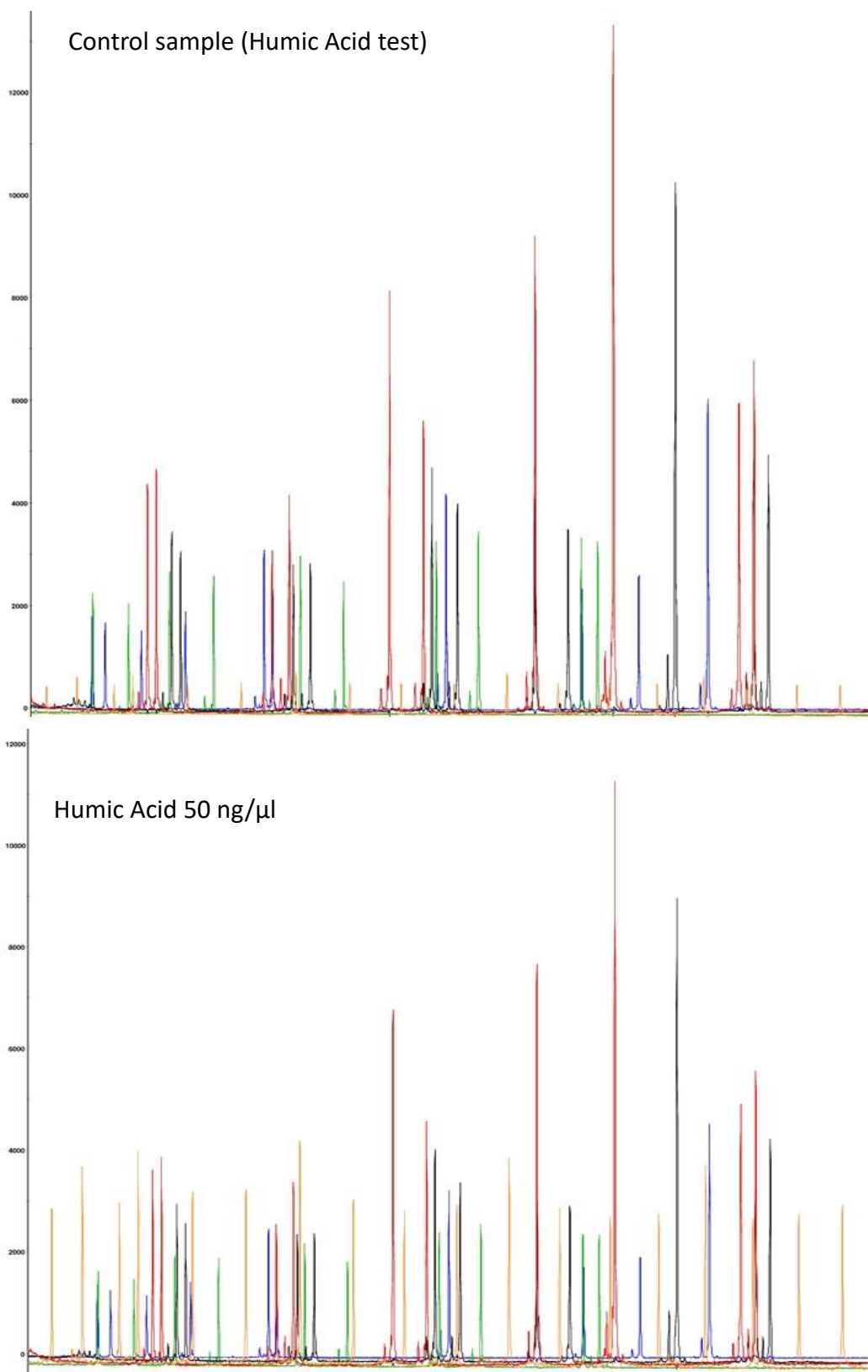


Figure 5.6. Testing of SureID® 23comp kit with humic acid. Three different concentrations of 50 ng/μl, 75 ng/μl and 100 ng/μl were tested. This figure shows the results of the control sample (no inhibition) and of the 50 ng/μl (humic acid) sample. Figure 5.7 shows the results of 75 and 100 ng/μl of humic acid.

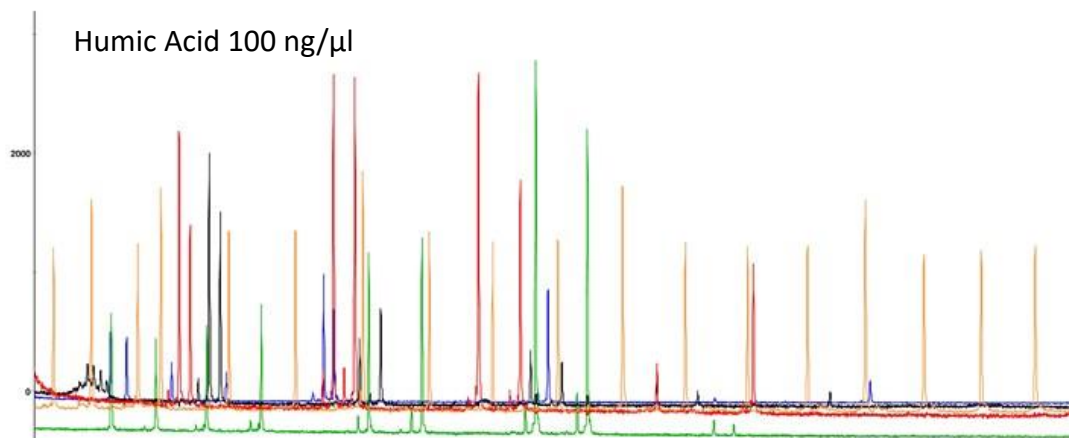
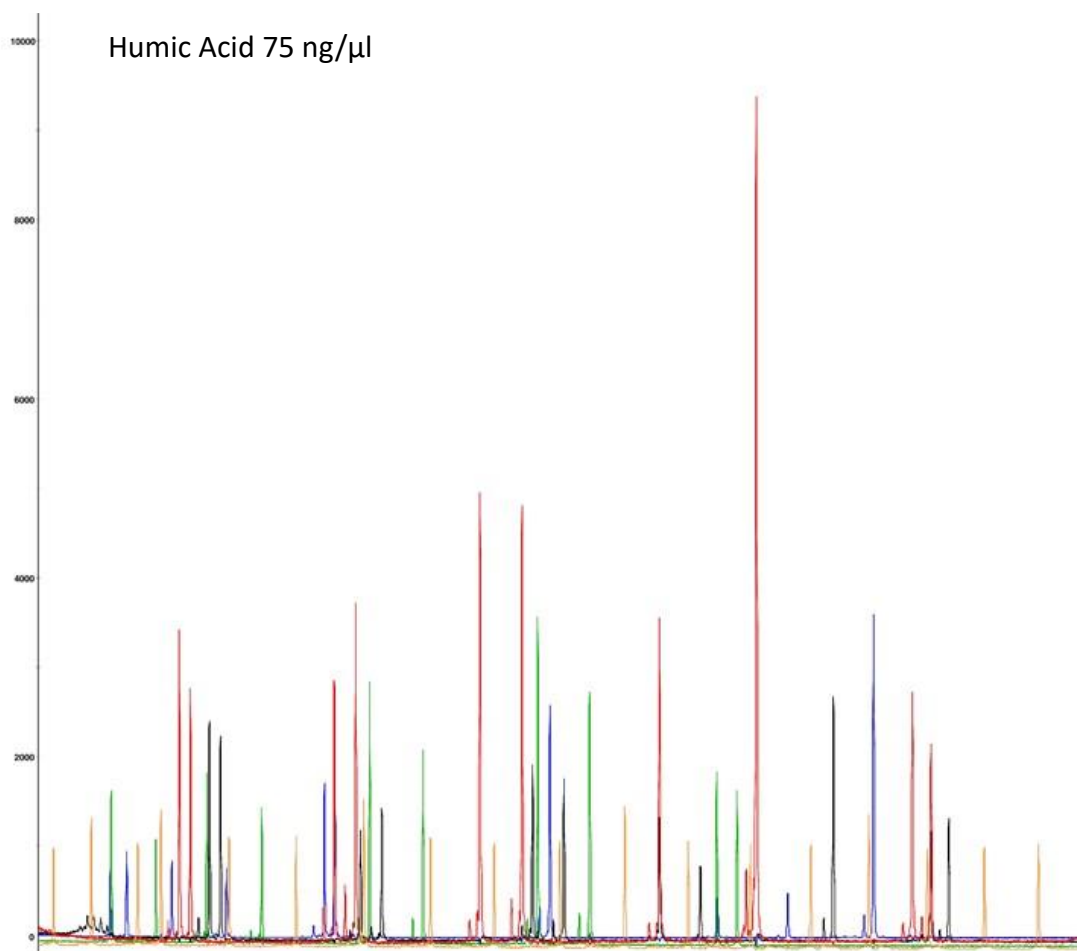


Figure 5.7. Testing of SureID® 23comp kit with humic acid. Three different concentrations of 50 ng/μl, 75 ng/μl and 100 ng/μl were tested. This figure shows the results of the 75 ng/μl (humic acid) sample and of the 100 ng/μl (humic acid) sample. Full profiles were achieved with ≤ 75 ng/μl of humic acid.

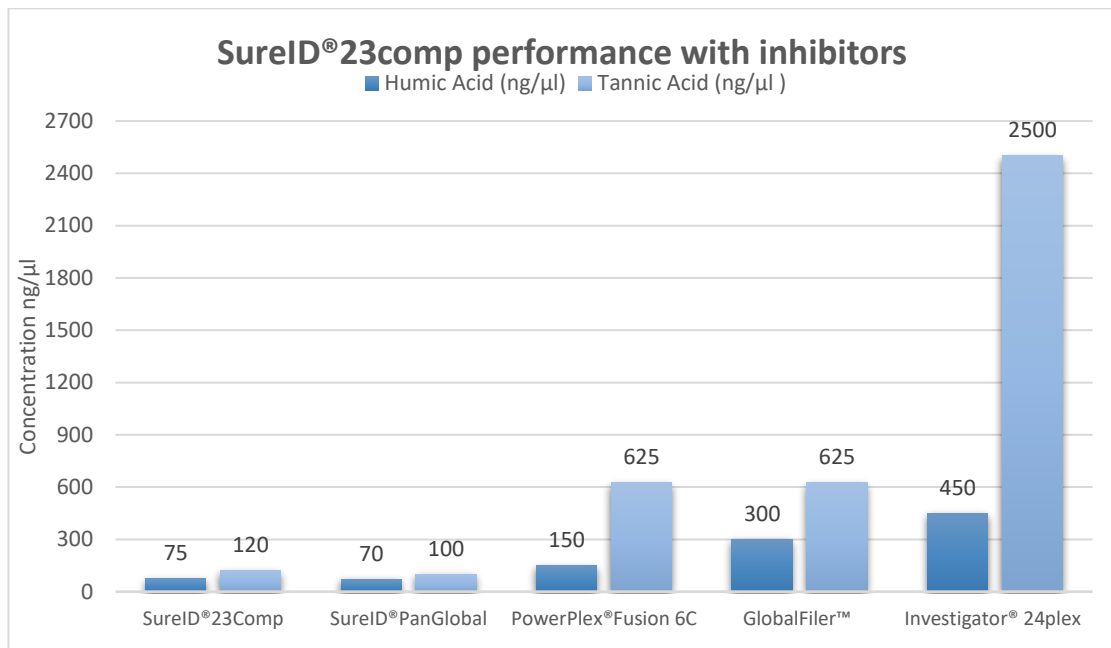


Figure 5.8. SureID® 23comp kit performance with two common PCR inhibitors. Full profiles were generated in the presence of 75 ng/μl of humic Acid and 120 ng/μl of tannic acid. These figures are similar to those reported for the SureID® PanGlobal (Liu *et al.* 2017). However, the kit was not as robust with inhibitors as PowerPlex® Fusion 6C, GlobalFiler™, and Investigator® 24plex (Lin *et al.* 2017) (Figure from (Alsafiah *et al.* 2019a).

5.5.2 Further performance assessment

The performance of the SureID® 23comp was further evaluated using the nine bone samples. The bone samples were profiled using the 25 μl volume to increase the capacity of the DNA input, in the PCRs, to 6.25 μl. Seven samples, where the total DNA input ranged from 0.2575 ng to 2.0444 ng/reaction, showed similar percentage of detected alleles to other kits previously used (Table 5.3). However, in two samples, which had lower concentrations and higher degradation indexes (DIs) of 0.0173 ng/μl (DI: 57.7) and 0.0194 ng/μl (DI: 16.2) (total DNA input 0.1081 ng and 0.1213 ng); the performance deteriorated, both in absolute terms, and in comparison to other kits. The capacity of DNA quantity in the other kits (15 μl) allowed 2.4-fold more DNA to be added to the reaction compared to the SureID® 23comp (6.25 μl) (Alsafiah *et al.* 2019a).

Table 5.3. The results of the bone samples used in the validation tests of the SureID® 23 comp kit. Nine samples, collected from a mass grave in Iraq, were extracted using PrepFiler™ BTA Forensic DNA Extraction Kit (AB), and were quantified using Quantifiler™ Trio DNA Quantification Kit (AB). This table shows Quantifiler™ Trio small fragment concentrations (ng/μl), total DNA quantities added to the PCRs of SureID®23 kit and other kits. The percentages of detected alleles of autosomal STRs (aSTRs) when using different STR kits, are also shown. The two samples that showed lower detection rate are shaded (Alsafiah *et al.* 2019a).

Sample #	Quantifiler™ Trio		Total DNA added to PCRs: SureID®23 PCR (ng/6.25 μl) / other kits (ng/15 μl)	% of detected alleles using different kits			
	Small fragment concentration (ng/μl)	Degradation Index (DI)		SureID®23 (22 STRs)	PowerPlex®21 (20 STRs)	GlobalFiler™ (21 aSTRs)	PowerPlex® Fusion 6C (23 aSTRs)
76 c	0.0173	57.666	0.1081/0.2595	27.30%	60%	66.60%	60.80%
78 a	0.0194	16.166	0.1213/0.2910	54.50%	90%	95.20%	82.60%
93 b	0.3271	2.7464	2.0444/4.9065	100%	100%	N/A	N/A
76 e	0.093	2.2962	0.5813/1.3950	100%	100%	N/A	N/A
81 a	0.0571	1.929	0.3569/0.8565	100%	100%	76.20%	N/A
97 b	0.0548	1.6758	0.3425/0.8220	100%	100%	N/A	N/A
94 a	0.0685	2.4204	0.4281/1.0275	100%	100%	N/A	N/A
25 a	0.0463	4.9784	0.2894/0.6945	86.30%	95%	N/A	N/A
46 b	0.0412	3.1937	0.2575/0.6180	100%	100%	N/A	N/A

N/A: sample was not profiled using this kit.

Overall, the sensitivity tests, when using the DNA control, demonstrated the robustness of generating full profiles even below the recommended DNA concentrations and showed similar sensitivity to other commonly used STR kits. However, this kit was less sensitive with the bone samples, which is most likely due to the limited capacity of DNA input compared to the other kits (Table 5.3 and Table 5.4). Although this kit was designed as a supplementary kit for forensic genetics laboratories, some cases may involve human remains, e.g. disaster victim identification (DVI). Therefore, increasing the concentration of the master and primer mixes (e.g. to 2X) would permit additional space for more DNA input especially for highly degraded samples (Alsafiah *et al.* 2019a).

Table 5.4. PCRs contents for the SureID® 23comp, PowerPlex® 21, GlobalFiler™, PowerPlex® Fusion 6C. The table shows the contents of the 25 µl volume PCRs for four kits used to genotype the bone samples. The SureID® 23comp has less space (6.25 µl) for DNA input compared to the other three kits (15 µl). Increasing the concentration of the master and primer mixes will increase the space for the DNA input (Alsafiah *et al.* 2019a).

Kit	PCRs total volume	Master Mix	Primer Mix	Maximum DNA input
SureID® 23comp	25 µl	12.5 µl	6.25 µl	6.25 µl
PowerPlex® 21	25 µl	5 µl	5 µl	15 µl
GlobalFiler™	25 µl	7.5 µl	2.5 µl	15 µl
PowerPlex® Fusion 6C	25 µl	5 µl	5 µl	15 µl

5.5.3 Heterozygote peak balances.

Peak balances study started with measuring of the optimal DNA quantity for the 10 µl reaction volume. The first 90 samples of the 500 samples were tested using three different DNA quantities (0.5 ,0.35, and 0.25) ng. With all template amounts, the minimum peak balance ratios were > 68%, which meets the criteria set out in the ENFSI guidelines (> 60%). The DNA input of 0.5 ng achieved the most balanced heterozygous peaks, with an average of 88.31% (Table 5.5), which are similar to ratios observed when testing other kits, for example Investigator® HDplex Kit (Westen *et al.* 2012). The

D21S2055 showed the lowest degree of balance at all template concentrations, with ratios of 73.11% at 0.5 ng, 79.75% at 0.35 ng, and 68.12% at 0.25 ng (Table 5.5) (Alsafiah *et al.* 2019a).

Table 5.5. Peak balance ratios study for the SureID® 23comp kit. The table shows the averages of peak balance ratios calculated for the amelogenin (AMEL) and 22 STRs included in the SureID® 23comp kit. A total of 90/500 samples were used to study balance ratios. The 10 µl reaction volume was evaluated using three DNA quantities 0.5, 0.35, and 0.25 ng. The 0.5 ng showed the highest peak ratios average. The D21S2055 showed the lowest ratio at all DNA quantities (shaded row) (Alsafiah *et al.* 2019a).

Marker	Average peak balance ratios (%) of the 10 µl reaction volume		
	0.5 ng	0.35 ng	0.25 ng
AMEL	94.29	83.90	81.83
D18S1364	90.90	86.11	83.90
D1S1656	86.62	87.43	85.21
D13S325	91.35	81.69	84.15
D5S2800	85.95	88.85	80.28
D9S1122	90.65	85.88	84.80
D4S2366	91.42	86.94	85.15
D3S1744	90.08	87.97	87.23
D12S391	85.38	83.09	78.35
D11S2368	91.57	84.19	83.81
D21S2055	73.11	79.75	68.12
D20S482	94.02	91.62	87.02
D8S1132	88.96	82.64	84.44
D7S3048	85.98	86.05	80.80
D2S441	90.83	84.08	87.61
D19S253	82.62	87.44	79.48
D10S1248	90.53	84.86	85.27
D17S1301	92.58	90.00	87.64
D22GATA198B05	86.48	85.54	80.41
D16S539	87.54	85.94	87.94
D6S474	85.07	87.73	84.22
D14S1434	88.41	87.93	85.99
D15S659	86.74	85.12	82.72
All markers' average (%)	88.31	85.86	83.32

The remaining 410 samples were successfully profiled using the 10 µl volume and 0.5 ng of DNA input. Overall, the intra-locus balances were 81.8 % (D21S2055) - 96.9 % (D16S539), the intra-dye balances 71.9 % (TAMRA) – 82.6 % (JOE), and the inter-dye balances >43 % (Alsafiah *et al.* 2019a). These figures are consistent with the recommended standard of PCR performance that are >70% for intra-locus balance, 50% for intra-dye balance, and >30% for inter-dye balance (Liu *et al.* 2017).

The peak imbalances of the D21S2055 became less than 50% when the size difference between heterozygous alleles was more than ten repeats (40 nt) (Figure 5.9). The peak

balances further decreased to < 45% when the size difference became > 50 nt (> 12 repeats). For example, an average of 43.5% for the genotypes (16.1, 34) (two samples), and 31.8% for the genotype (16.1, 36) (one sample) (Figure 5.9). This locus is the longest marker in this kit (332 bp to 420 bp) and has the highest number of possible alleles (23 alleles: 16.1 to 38) (Alsafiah *et al.* 2019a).

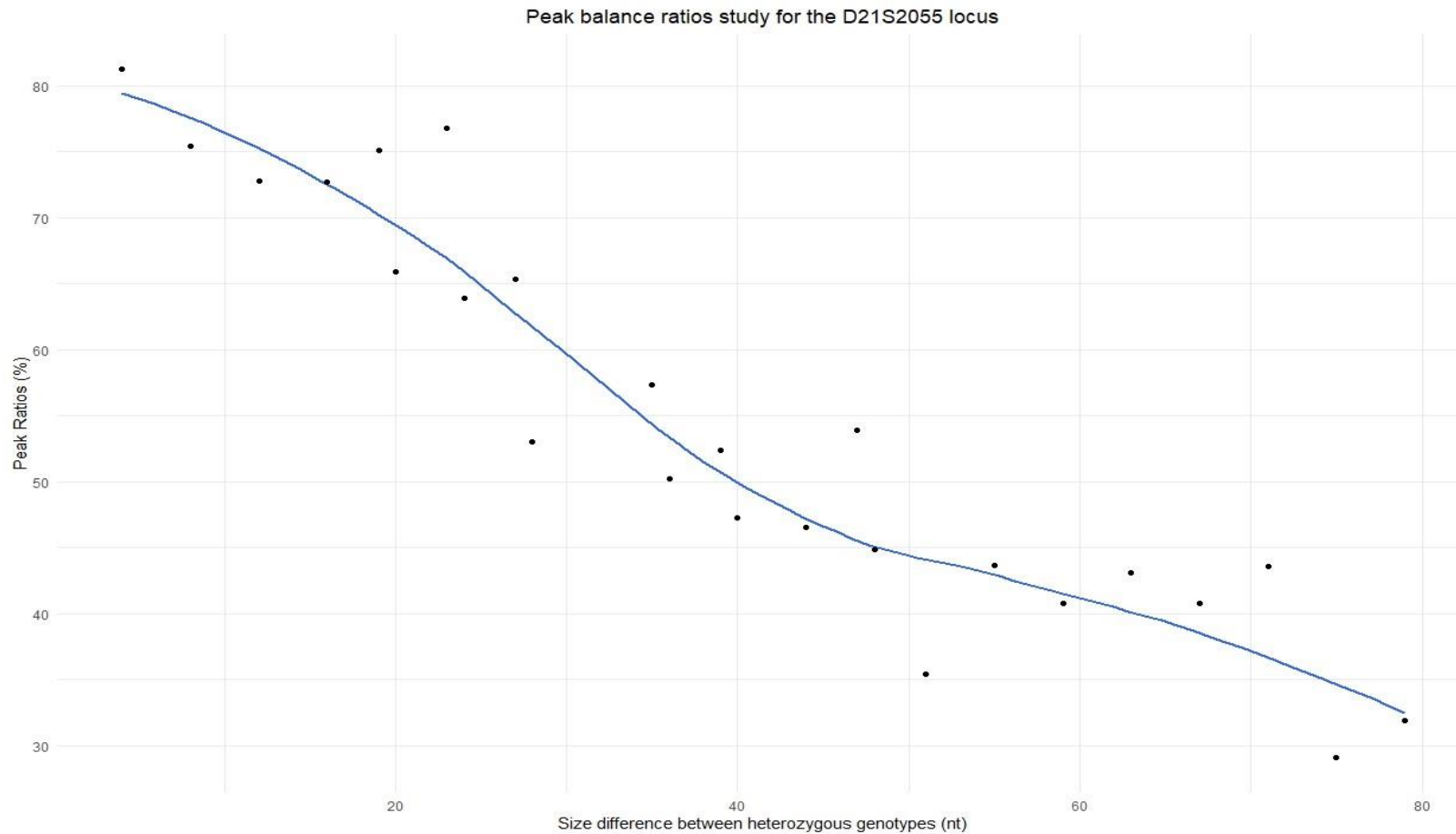


Figure 5.9. Peak balance ratios study for the D21S2055 locus. This figure shows a study of the correlation between the size difference between heterozygous alleles and the peak balance ratio for the D21S2055 locus using data of 500 samples. The peak ratios of all genotypes that have the same size difference (nt) (e.g. the genotypes 13, 17; 14, 18; and 15, 19 have the same size difference of 4 nt) were averaged and are represented by the black dots. The blue line shows the smoothed mean of the peak ratios. Heterozygote alleles with >50 nucleotides difference showed peak ratios <45% (Alsafiah *et al.* 2019a).

5.5.4 Stutter/corresponding allele ratios.

Stutter artefacts are common to all PCR-based STR analysis and the most common type of stutter is a peak with one repeat smaller than the true allele (Krenke *et al.* 2005). In this study, the average of the stutter peak ratios was 9.18% and the average range was from 3.8% for D2S441 to 16.15% for D12S391 (Figure 5.10). In addition, allele variants of x.1, x.2 and x.3 had lower stutter ratios than alleles x-1, x-2 and x-3, respectively. These figures were as expected that showed the correlation between the size of an allele and the complexity of a locus with the stutter ratios and were below the stutter filter provided by the manufacturer (Alsafiah *et al.* 2019a).

5.5.5 Precision and accuracy.

For the precision study, the data of 22,975 alleles (23,000 alleles from 500 samples excluding 25 alleles with a single observation) were used to calculate the standard deviation (s.d.) of the fragment sizes of each allele at a locus. Overall, the maximum s.d. was 0.1048 nucleotide (nt) observed in allele 21 at D7S3048 and the minimum was 0.0071 nt observed in allele 22 at D3S1744 (Figure 5.11) (Alsafiah *et al.* 2019a).

To measure the accuracy of the kit, the average sizes of each allele in the data of the 500 samples and in 21 allelic ladders was compared to the actual size values of the corresponding allele (actual sizes provided by the manufacturer). All alleles fell within the range of ± 0.41 nt, where allele 17 at D6S474 (0.4096 nt) and allele 26 at D7S3048 (-0.4084 nt) recorded the highest difference compared to the actual sizes (Figure 5.12) (Alsafiah *et al.* 2019a).

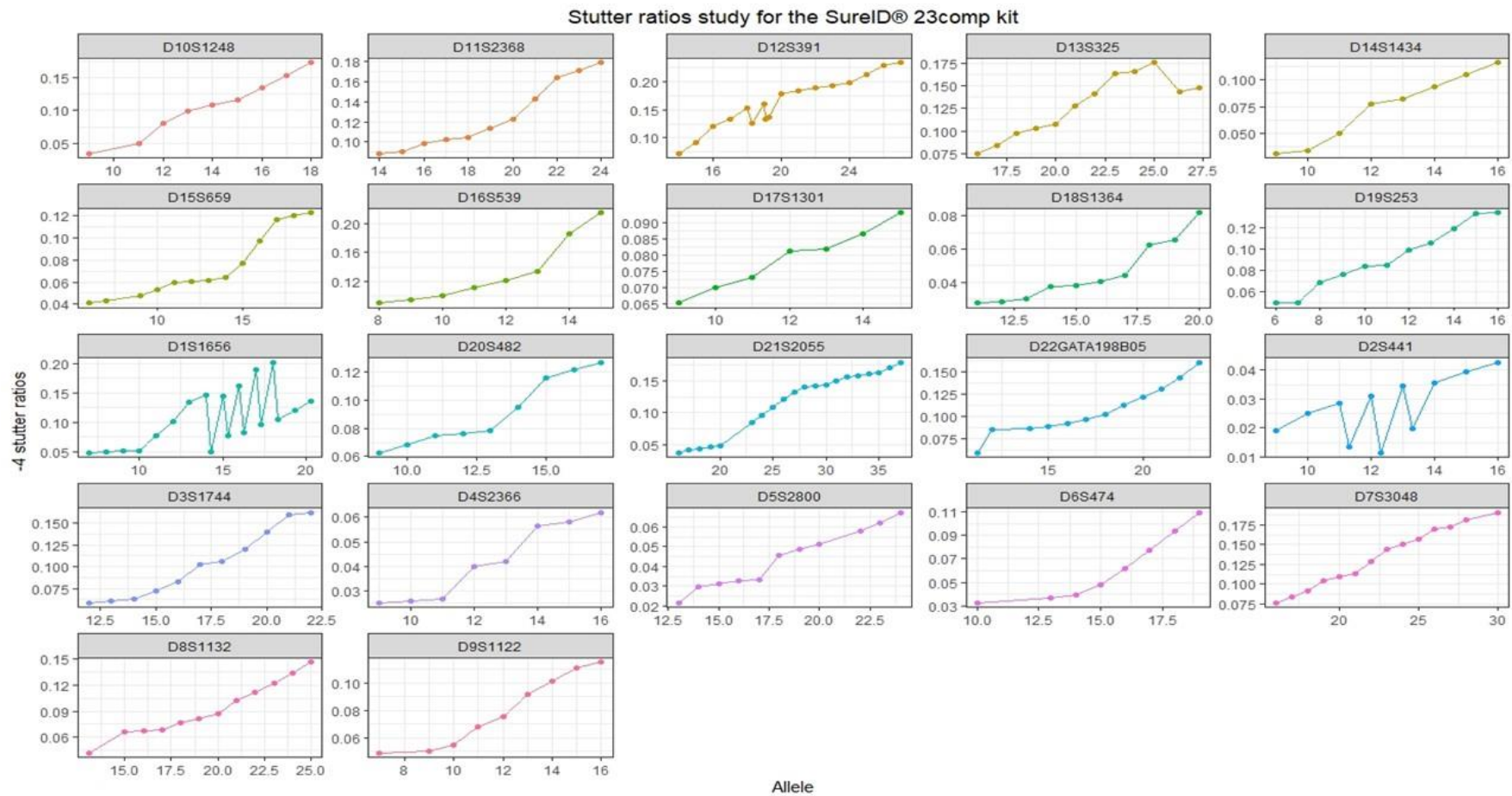


Figure 5.10. Stutter ratios study for the SureID 23comp Kit. The figure shows the average of - 4 stutter ratios for STRs included in the SureID 23comp kit. Each box represents the stutter ratios of an STR. The x-axis represents alleles and the y-axis represents the stutter ratios. The line was drawn based on the average ratios of observed stutters. Alleles of x.1, x.2 and x.3 are plotted at x.25, x.5, and x.75 respectively. The average of stutter ratios ranged from 3.8% for D2S441 to 16.15% for D12S391 (Alsafiah *et al.* 2019a).

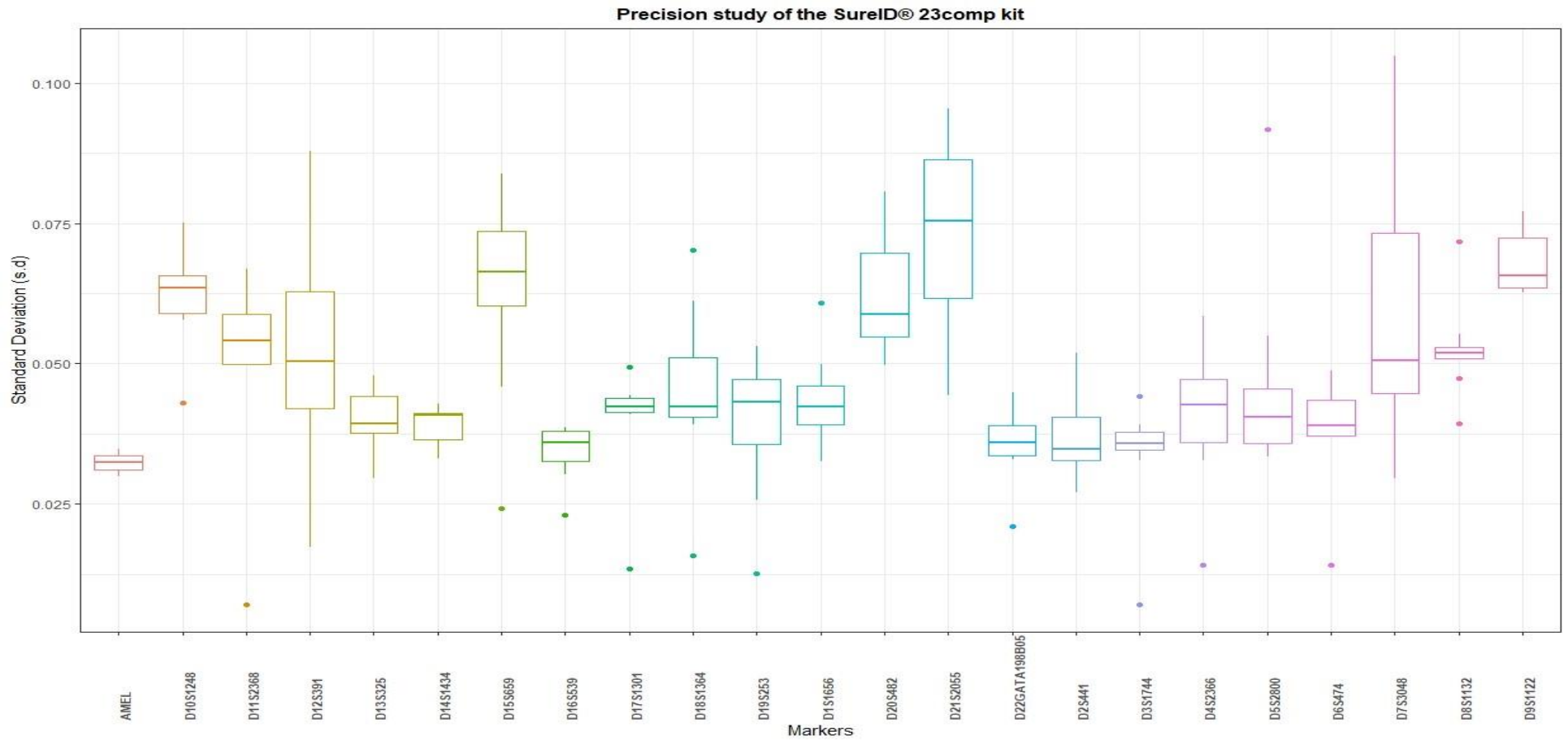


Figure 5.11. Precision study of the SureID 23comp Kit. The figure shows standard deviation (s.d) values of the fragment sizes of 22,981 alleles generated from 500 samples tested by the SureID 23comp. The highest s.d. was observed in allele 21 at D7S3048 (0.1048 nt) (Alsafiah *et al.* 2019a). In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).

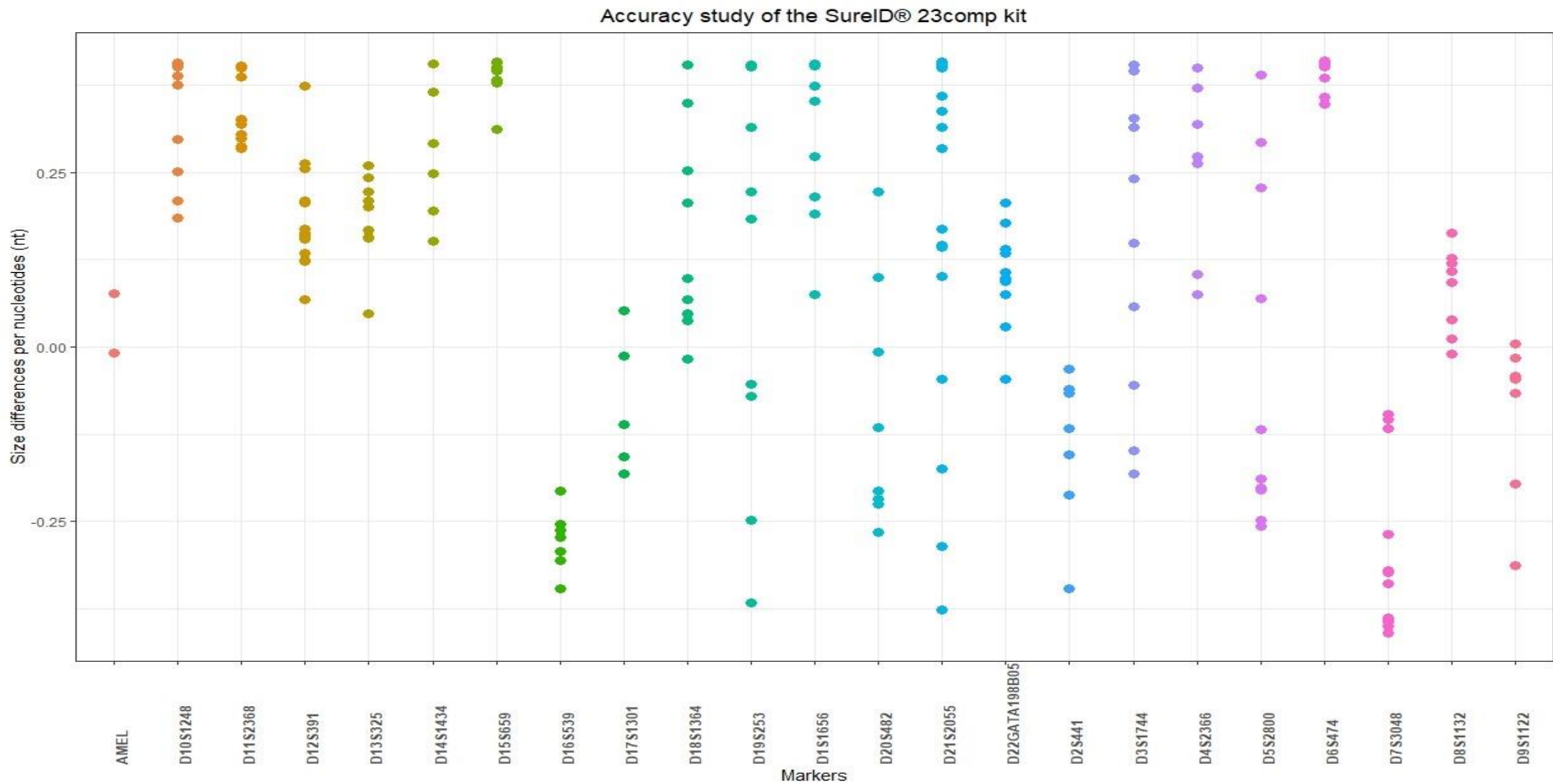


Figure 5.12. Accuracy study of the SureID 23comp Kit. The average of the size values of each allele in the data of the 500 samples and in 21 allelic ladders were compared to the actual sizes of the corresponding allele (actual sizes provided by the manufacturer). The size differences per nucleotides were calculated and are represented by the coloured dots. All alleles fell within the range of ± 0.41 nt of the allelic window; the largest differences were seen at D6S474 allele 17 (0.4096 nt) and D7S3048 allele 26 (-0.4084 nt) (Alsafiah *et al.* 2019a).

The precision and the accuracy tests demonstrated the capability of detecting heterozygous alleles that differ by a single nucleotide and demonstrated that it is unlikely for any allele to be sized out of the designated window (± 0.5 nt). The SureID[®] 23comp was reliably able to detect genotypes where the difference between the alleles was a single nucleotide, for example 11.3, 12 at D2S441, 15.3, 16 and 16.3, 17 at D1S1656 (100% concordant with GlobalFiler[™] genotypes) (Alsafiah *et al.* 2019a).

5.5.6 Concordance study

The concordance study was also carried out by comparing data of the 500 samples obtained from this study and that generated using the GlobalFiler[™] kit (Chapter 3) (Alsafiah *et al.* 2017). The five common loci (D1S1656, D2S441, D10S1248, D12S391 and D16S539) showed 100% concordance. In addition, alleles generated from the bone samples using the SureID[®] 23comp kit at the common loci were concordant with alleles generated using the other kits. In addition, the amelogenin showed concordant genotypes to that generated by the GlobalFiler[™] kit (Alsafiah *et al.* 2019a).

5.5.7 Allelic ladder and rare alleles.

This kit provides an allelic ladder representing 232 alleles that are supported by 53 additional bins for variant alleles (Figure 5.1). After analysing the 500 samples, 34 alleles in 15 loci were not represented by the allelic ladder; three of which had been observed ≥ 40 times (Table 5.6). In addition, ten of these alleles were situated outside the designated window of their loci: alleles 7 and 8 at D1S1656, 26.3 and 27.3 at D13S325, allele 16 at D4S2366, allele 12 at D3S1744, allele 30 at D7S3048, allele 10 at D6S474 and alleles 6 and 7 at D15S659 (Figure 5.13). The allele 7 at D1S1656 was situated under the designated area of D18S1364 locus (Figure 5.13 a). Although this allele could belong to D18S1364, forming triplet allele genotype, it was confirmed by sequencing that it

belongs to D1S1656 (Alsafiah *et al.* 2018) (Chapter 4). It is not necessary for an allelic ladder to represent all rare alleles; however, alleles outside the designated window of a locus may be misinterpreted especially when adjacent loci are homozygous. Examining data of 256 samples collected from the population of Ningbo, China (data provided by the Health Gene Technologies) (Table 5.6), most alleles present in the Saudi Arabian population but not present in the allelic ladder were absent in the Ningbo population (Alsafiah *et al.* 2019a).

Table 5.6. Alleles not represented by the allelic ladder of SureID® 23comp kit detected in the population of Saudi Arabia; 34 alleles were detected at 15 STRs. It shows also the frequency of these alleles in Ningbo population (data provided by the Health Gene Technologies). The frequencies of detected alleles ranged from 0.001 (one observation) to 0.066 (66 observations). Shaded rows indicate alleles observed ≥ 40 times (Alsafiah *et al.* 2019a).

STRs	Allele	frequency		STRs	Allele	frequency	
		Saudi	Ningbo			Saudi	Ningbo
D18S1364	11	0.001	0.002	D13S325	26.3	0.002	0
D1S1656	7	0.001	0		27.3	0.001	0
	8	0.001	0	D8S1132	13.1	0.001	0
	10	0.004	0.001		15	0.003	0
	14.3	0.002	0	D7S3048	30	0.001	0
	15.3	0.040	0	D2S441	8.3	0.001	0
	16.3	0.061	0.007		9	0.005	0
	18	0.003	0.011		11.3	0.066	0
	19.3	0.006	0.003		13.3	0.001	0
	20.3	0.001	0.002	D19S253	6	0.004	0
D9S1122	7	0.001	0		16	0.001	0
D4S2366	16	0.002	0	D22GATA198B05	11.2	0.001	0
D3S1744	12	0.001	0		12	0.004	0
D12S391	18.3	0.005	0	D6S474	10	0.001	0
	19.1	0.001	0	D14S1434	16	0.004	0.004
	19.3	0.004	0	D15S659	6	0.001	0
	27	0.003	0.003		7	0.003	0

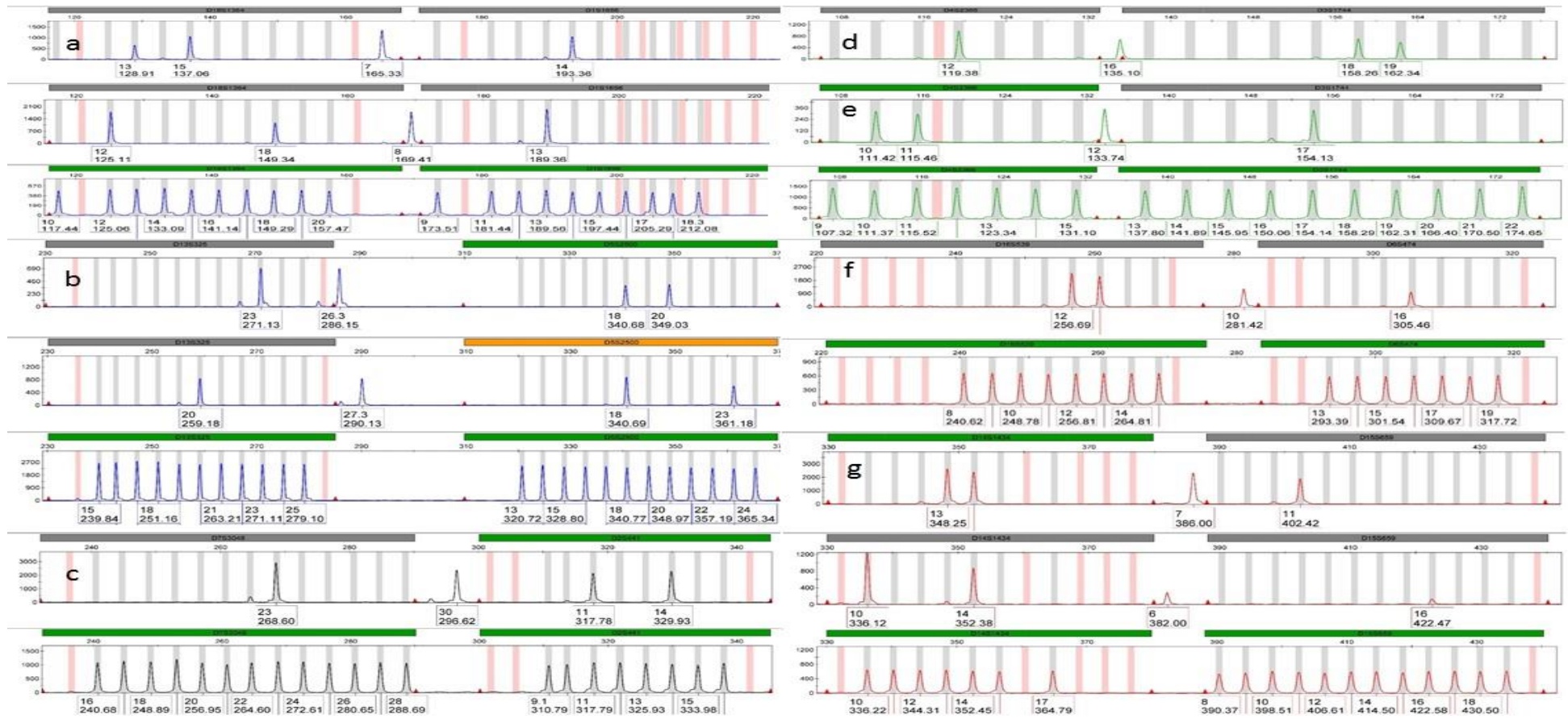


Figure 5.13. Alleles outside the windows of the allelic ladder of the SureID® 23comp kit. This figure shows ten alleles observed in the population of Saudi Arabia that are not represented and were situated outside the designated widow of their loci. a) Alleles 7 and 8 at D1S1656. B) Alleles 26.3 and 27.3 at D13S325. c) Allele 30 at D7S3048. d) Allele 16 at D4S2366. e) Allele 12 at D3S1744. f) Allele 10 at D6S474. g) Alleles 6 and 7 at D15S659. Allele 7 at D1S1656 (a) was situated under the designated area of D18S1364 (Alsafiah *et al.* 2019a).

5.5.8 Population study and excess of homozygosity.

In Chapter 3, the 21 aSTRs included in the GlobalFiler™ kit have shown excess of homozygosity in 20/21 aSTRs (TPOX was the exception) with an inbreeding coefficient of 0.03560, but none of the loci showed significant deviation from HWE. Here, 14/17 non-CODIS loci showed fewer than expected heterozygotes (D9S1122, D4S2366 and D8S1132 were the exception). In addition, D20S482 was the only locus that showed significant deviation (P value = 0) (Table 5.7). This also revealed some level of consanguinity in the population of Saudi Arabia which was supported by an inbreeding coefficient (F_{IS}) of 0.02977.

Table 5.7. Results of the expected heterozygosity calculation and of Hardy-Weinberg equilibrium exact test, conducted by Arlequin v3.5.2.1 Software for the 17 non-CODIS loci included in the SureID® 23comp kit. The P values is significant if < 0.05 . The five common loci with the GlobalFiler™ kit were not included in this table as they had the same results in Table 3.2.

Locus	Alleles No	Observed Heterozygosity	Expected Heterozygosity	Exact test P value	Standard Deviation	Steps done
D18S1364	1000	0.816	0.84136	0.56136	0.00041	1001000
D13S325	1000	0.766	0.79728	0.39205	0.00043	1001000
D5S2500	1000	0.726	0.77991	0.11518	0.00028	1001000
D9S1122	1000	0.724	0.71099	0.05367	0.00021	1001000
D4S2366	1000	0.818	0.7953	0.25306	0.00036	1001000
D3S1744	1000	0.754	0.80825	0.23985	0.0003	1001000
D11S2368	1000	0.77	0.80282	0.16433	0.00025	1001000
D21S2055	1000	0.888	0.91181	0.16668	0.00024	1001000
D20S482	1000	0.628	0.6915	0	0	1001000
D8S1132	1000	0.87	0.85408	0.21807	0.00028	1001000
D7S3048	1000	0.848	0.88177	0.48789	0.0003	1001000
D19S253	1000	0.774	0.77516	0.88958	0.00022	1001000
D17S1301	1000	0.662	0.67583	0.18288	0.00038	1001000
D22GATA198B05	1000	0.806	0.84142	0.33665	0.00044	1001000
D6S474	1000	0.726	0.75407	0.11216	0.00029	1001000
D14S1434	1000	0.698	0.69978	0.13021	0.0003	1001000
D15S659	1000	0.802	0.83976	0.34694	0.00048	1001000

However, it is not clear whether the D20S482's deviation was due to the consanguinity detected in the population of Saudi Arabia or due to null alleles.

Examining SNP variants with > 1% frequency at the flanking regions (100 bp each side) of the locus using the 1000 Genome browser (Auton *et al.* 2015), two SNPs in the 5' flanking region: rs151133985 (all populations C: 99%, G: 1%; Africans C: 98%, G: 2%) and rs77560248 (all populations C: 94%, T: 6%; Europeans and South Asians C: 91%, T: 9%); and one SNP at 3' flanking region: rs551422781 (Africans G: 99%, A: 1%), were found. These SNPs may cause null alleles if any of them was at a critical annealing region of the primer pair. However, none of the three populations European, South Asian and African that were studied using the same kit has shown deviation from HWE at this locus (Iyavoo *et al.* 2019). Sequencing homozygotes samples or using a different kit to genotype the locus may reveal more information. Therefore, based on our results for the population of Saudi Arabia, D20S482 cannot be included in the product rule to calculate the probability of a DNA profile.

Assuming no deviation from HWE, the CMP for the 22 STRs was 7.2×10^{-27} , the CPE was 0.999999037259, and the CPD was 0.999999999999999999999999999928. The non-CODIS loci alone had 1.2×10^{-20} CMP, 0.9999747848 CPE and 0.999999999999999999999999999988164 CPD (Table 5.8). D21S2055 was the most informative locus, with a MP of 0.016, and D17S1301 was the least informative locus, with a MP of 0.162. Heterozygosity ranged from 0.624 (D20S482) to 0.89 (D21S2055). The number of observed alleles per locus varied from 7 alleles in D17S1301 to 20 alleles in D21S2055. Three alleles, allele 14 in D20S482, allele 12 in D17S1301 and allele 12 in D9S1122; showed very high frequencies of 0.477, 0.449 and 0.405 respectively (Table 5.8). The frequency of the theoretical most common SureID® DNA profile, generated based on the frequencies of the 22 STRs (and assuming heterozygosity), was 3×10^{-21} that equates to 1 in 3.3×10^{20} . The CMP of the 22 STRs is ten times higher than CMP calculated when using the 21 loci of GlobalFiler™ kit (1.421×10^{-26}) (Alsafiah *et al.* 2017). Apart of SE33, the

SureID®23comp kit includes the four most informative loci that have been studied for the population of Saudi Arabia (D21S2055, D12S391, D7S3048, and D1S1656) (Alsafiah *et al.* 2019a).

Table 5.8. Allele frequency of the 17 non-CODIS loci. The table shows the allele frequency and statistical parameters for the 17 non-CODIS included in the SureID® 23comp kit. Allele frequencies for D1S1656, D2S441, D10S1248, D12S391 and D16S539 are not shown as they are presented in Table 3.4.

allele	D18S1364	D13S325	D5S2800	D9S1122	D4S2366	D3S1744	D11S2368	D21S2055
7				0.001				
9				0.008	0.299			
10				0.054	0.107			
11	0.001			0.237	0.172			
12	0.045			0.405	0.243	0.001		
13	0.247		0.021	0.256	0.106	0.008		
14	0.142		0.238	0.031	0.066	0.117	0.001	
15	0.185		0.003	0.007	0.005	0.079	0.002	
16	0.126	0.004	0.001	0.001	0.002	0.133	0.016	
16.1								0.082
17	0.043	0.012	0.239			0.34	0.048	
17.1								0.013
18	0.143	0.045	0.28			0.167	0.143	
18.1								0.013
19	0.062	0.167	0.006			0.096	0.241	
19.1								0.108
20	0.006	0.308	0.025			0.045	0.266	
20.1								0.01
21		0.231				0.012	0.206	
22		0.141	0.003			0.002	0.059	
23		0.071	0.165				0.016	0.001
24		0.017	0.019				0.002	0.026
25		0.001						0.131
26								0.121
26.3		0.002						
27								0.014
27.3		0.001						
28								0.012
29								0.056
30								0.032
31								0.041
32								0.102
33								0.124
34								0.07
35								0.034
36								0.008
37								0.002
Total Alleles	1000	1000	1000	1000	1000	1000	1000	1000
Matching Probability	0.046	0.070	0.079	0.141	0.076	0.060	0.068	0.016
Expressed as 1 in ...	21.868	14.346	12.579	7.093	13.176	16.611	14.685	61.516
Power of Discrimination	0.954	0.930	0.921	0.859	0.924	0.940	0.932	0.984
Polymorphic Information Content	0.821	0.768	0.744	0.661	0.765	0.785	0.773	0.904
Power of Exclusion	0.629	0.538	0.470	0.466	0.637	0.517	0.541	0.775
Typical Paternity Index	2.717	2.137	1.825	1.812	2.778	2.033	2.155	4.545

The data of the 22 loci of the SureID® 23comp was uploaded to the R studio to find out the maximum number of matched loci between any two DNA profiles using the DNA Tools package. In the 500 samples, the maximum number of loci matching between any two samples was 9 out of 22 loci (40% of the 22 loci), which was observed in two sample pairs. One pair of sequences showed partial matching (i.e. one of the two alleles) at 20 out of 22 loci. This illustrates the power of the additional loci for human identification and kinship testing (Table 5.9) (Alsafiah *et al.* 2019a).

To assess the SureID®23comp for kinship testing, a typical paternity case (an alleged father, a child and a known mother) was assumed, and the combined typical paternity index (CPI) of 93,835,307.21 was used to calculate the paternity probabilities with different prior probabilities (Pr): 0.90, 0.50 and 0.10. Assuming that all loci are within HWE, the probabilities of paternity were 99.9999988% (Pr = 0.90), 99.9999893% (Pr = 0.50) and 99.9999041% (Pr = 0.10), which are higher than those probabilities calculated when using the GlobalFiler™ kit and the currently used kit in Saudi Arabia (Table 3.5)(Alsafiah *et al.* 2019a).

The ForenSeq™ DNA Signature Prep (Verogen), when combining the 94 SNPs and the 27 STRs, has shown much higher CMP of 10E-67 to 10E-69 (length-based STRs calls) and of 10E-71 to 10E-74 (sequence-based STRs calls), where the CMP of the 94 SNPs alone were (10E-38 to 10E-35) (Churchill *et al.* 2017). In addition, using MPS systems in kinship testing could help in tracking mismatches between tested individuals that have occurred due to mutation in the binding sites of primers (Li, R. *et al.* 2019). However, this requires additional technology to be implement and is currently not available in many countries (Alsafiah *et al.* 2019a).

Table 5.9. The maximum of matched loci per any sample pair within the 500 samples. In the 500 samples, only two pairs of samples showed full matching in 9 loci (i.e. both alleles). This was the maximum number of matched loci (shaded row). One pair of sequences showed partial matching (i.e. one of the two alleles) at 20 out of 22 loci (shaded column). This table was generated by the R studio using the package of DNA tools (Alsafiah *et al.* 2019a).

		No. of partial match per any sample pair																						
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
No. of matched loci per any sample pair	0	0	0	2	11	29	114	320	730	1395	2219	3194	3689	3493	2820	2032	1065	494	194	57	19	1	0	0
	1	0	0	0	19	90	309	768	1816	3296	5036	6509	6769	6069	4408	2663	1264	448	133	24	5	0	0	
	2	0	1	2	26	114	389	1060	2212	3588	5336	6016	5716	4473	2852	1450	565	189	39	7	0	0		
	3	0	0	4	22	98	333	792	1660	2456	3269	3452	2811	1927	1019	469	158	35	8	2	0	0		
	4	0	0	3	18	68	179	409	757	1081	1299	1183	899	517	268	76	20	4	0	0				
	5	0	0	0	12	20	83	179	272	348	364	278	194	124	26	7	0	1	0					
	6	0	0	1	6	8	26	53	63	89	85	49	25	9	0	1	0	0						
	7	0	0	1	0	1	8	11	17	24	15	6	1	1	1	0	0							
	8	0	0	0	1	1	3	2	1	2	0	0	0	0	0	0								
	9	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0								
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0									
	11	0	0	0	0	0	0	0	0	0	0	0	0	0										
	12	0	0	0	0	0	0	0	0	0	0	0	0											
	13	0	0	0	0	0	0	0	0	0	0	0												
	14	0	0	0	0	0	0	0	0	0	0													
	15	0	0	0	0	0	0	0	0															
	16	0	0	0	0	0	0	0																
	17	0	0	0	0	0	0																	
	18	0	0	0	0	0																		
	19	0	0	0	0																			
	20	0	0	0																				
	21	0	0																					
22	0																							

5.5.9 Population comparison

Recently, the kit was used to generate population genetic data for three main populations European, South Asian and African (Iyavoo *et al.* 2019). The data of the three populations, and the Ningbo population (China, data provided by the Health Gene Technologies) were compared to the data generated by this study. Arlequin v 3.5.2 was used to estimate the distance between all populations by calculating the F_{ST} values and to carry out the population differentiation exact test.

The F_{ST} values showed that the European and South Asian populations were more similar to Saudi population than the African and Ningbo populations (Figure 5.14).

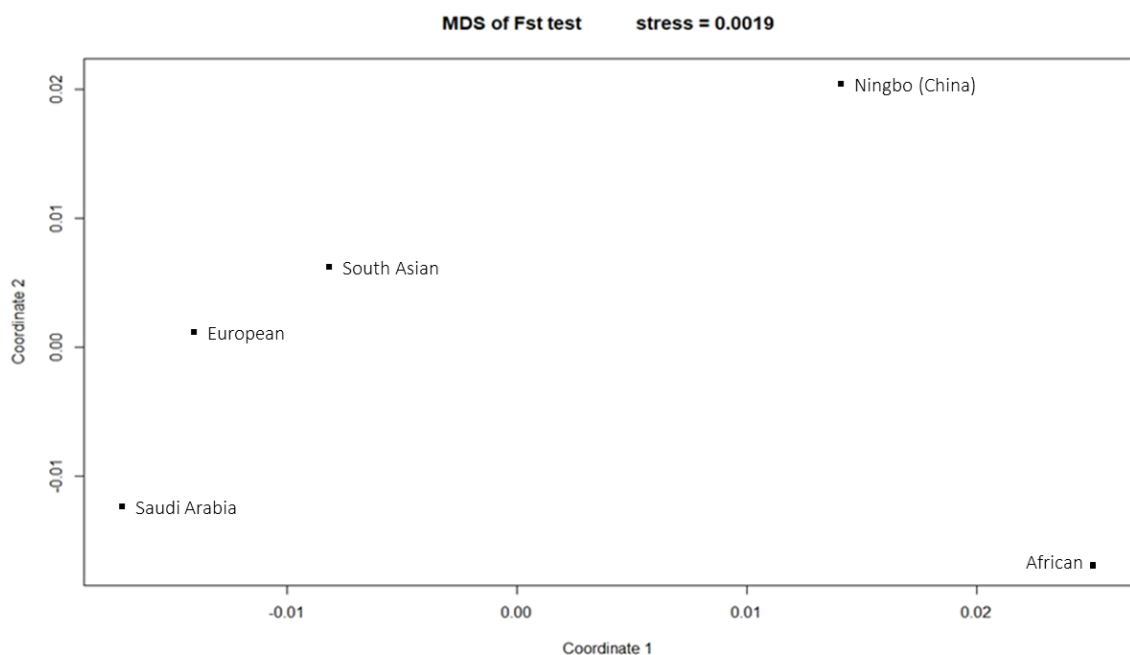


Figure 5.14. Multi-dimensional scaling for the average F_{ST} values. Five populations were included in the comparison and each number represent a population, Saudi Arabia (this study), European (Iyavoo *et al.* 2019), African (Iyavoo *et al.* 2019), South Asian (Iyavoo *et al.* 2019) and Ningbo population (data provided by the Health Gene Technologies). The European and South Asian populations were more similar to Saudi population than the African and Ningbo populations. The cmdscale function was used in R software to generate a multi-dimensional scale (MDS).

The P values of the exact test showed concordant result with the F_{ST} value estimates.

The European population (Figure 5.14) had the lowest number of STRs with significant difference 15/22 loci (P value < 0.05) and South Asian population (Figure 5.14) had 16/22 loci. The African (Figure 5.14) and Ningbo population (Figure 5.14) had more loci with significant difference of 21/22 and 18/22 respectively.

Table 5.10. Population differentiation exact test results using the Arlequin v3.5.2 Software. Shaded data indicates significant differences (P value < 0.05). European and South Asian populations showed lower number of STRs with significant difference.

	D18S1364	D1S1656	D13S325	D5S2800	D9S1122
European	0.00000+-0.0000	0.00000+-0.0000	0.00015+-0.0001	0.00000+-0.0000	0.01569+-0.0034
African	0.00000+-0.0000	0.00000+-0.0000	0.00689+-0.0029	0.00000+-0.0000	0.00000+-0.0000
South Asian	0.03643+-0.0041	0.00000+-0.0000	0.02873+-0.0121	0.00002+-0.0000	0.15175+-0.0248
Ningbo	0.00000+-0.0000	0.00000+-0.0000	0.60367+-0.0282	0.00000+-0.0000	0.00000+-0.0000
	D4S2366	D3S1744	D12S391	D11S2368	D21S2055
European	0.00000+-0.0000	0.51861+-0.0200	0.00000+-0.0000	0.19631+-0.0167	0.00000+-0.0000
African	0.00000+-0.0000	0.00059+-0.0004	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
South Asian	0.00184+-0.0008	0.00195+-0.0009	0.00000+-0.0000	0.20596+-0.0118	0.00000+-0.0000
Ningbo	0.00000+-0.0000	0.21489+-0.0195	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
	D20S482	D8S1132	D7S3048	D2S441	D19S253
European	0.61243+-0.0130	0.35328+-0.0362	0.01699+-0.0085	0.00018+-0.0002	0.10409+-0.0086
African	0.11142+-0.0199	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
South Asian	0.83443+-0.0185	0.00889+-0.0018	0.00000+-0.0000	0.00000+-0.0000	0.53261+-0.0272
Ningbo	0.00201+-0.0008	0.08212+-0.0123	0.00000+-0.0000	0.00000+-0.0000	0.00111+-0.0003
	D10S1248	D17S1301	D22GATA198B05	D16S539	D6S474
European	0.00000+-0.0000	0.11679+-0.0122	0.00002+-0.0000	0.00452+-0.0012	0.00000+-0.0000
African	0.00000+-0.0000	0.00000+-0.0000	0.00196+-0.0005	0.02100+-0.0044	0.00000+-0.0000
South Asian	0.00000+-0.0000	0.00164+-0.0008	0.00000+-0.0000	0.07688+-0.0085	0.00000+-0.0000
Ningbo	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
	D14S1434	D15S659			
European	0.00025+-0.0001	0.07551+-0.0113			
African	0.00001+-0.0000	0.00000+-0.0000			
South Asian	0.00000+-0.0000	0.09918+-0.0144			
Ningbo	0.00000+-0.0000	0.00000+-0.0000			

5.5.10 STRidER quality control

The data of the 17 non-CODIS STRs was sent to STRidER (Bodner *et al.* 2016) for quality control check. The data was approved and was given a dataset reference number of STR000178 (Appendix 4).

5.6 Conclusion

The SureID[®] 23comp was validated following the minimum criteria for validation recommended by the ENFSI and the SWGDAM for forensic applications as a supplementary kit with an exception of mixture studies that were not carried out as the kit is specifically designed to be used in complex kinship testing. The kit is reproducible, precise, accurate and reliable for forensic application as a supplementary kit and for databasing. The validation included a clarification of the correct identity of the D5 locus which is D5S2800 not D5S2500 that is now updated in the panels and supporting documents of the kit (Alsafiah *et al.* 2019a). The sensitivity tests demonstrated the capability of generating a full profile below the recommended DNA input but showed that the kit was less sensitive compared to other commonly used kits with degraded samples, which was at least in part because of the lower volume of template that can be added. Therefore, the kit can benefit from increasing the concentration of the reaction mix allowing more space for DNA input to 15 µl rather than 6.25 µl (Alsafiah *et al.* 2019a). In addition, including additional alleles and allele variations in the available spaces of the allelic ladder will allow specific allele designation and will minimise the need to re-run undesignated alleles (Alsafiah *et al.* 2019a).

The kit was evaluated for the population of Saudi Arabia and showed that the 22 STRs provided a CMP of 7.4E-27; the 17 non-CODIS loci alone provided 1.2 E-20 CMP. When the kit is used with the GlobalFiler kit, the 38 loci combined provided 1.7E-46 CMP.

Apart of SE33, the kit includes the four most informative loci that have been studied for the population of Saudi Arabia (D21S2055, D12S391, D7S3048, and D1S1656), two of which are included in the GlobalFiler kit. The kit achieved a CPI of 93,835,307.21 that is two times higher the CPI recorded for the GlobalFiler™ kit and allowed higher paternity probability of 99.99999893% ($Pr = 0.5$) (Alsafiah *et al.* 2019a). The study provides allele frequency data for additional 17 STRs that can be used to estimate the profile frequencies in Saudi Arabia.

Four populations were included in the population comparison by which the European and South Asian populations were, as expected, more similar to the Saudi population than the African and Ningbo populations.

Overall, this study evaluates the utility of the SureID® 23comp as a supplementary kit for kinship testing and determined that the kit met the criteria commonly used in forensic genetics laboratories. The kit allows the analysis of 17 non-CODIS loci and increases likelihood ratios, and thereby has the potential to increase the level of confidence in conclusions in complex kinship tests (Alsafiah *et al.* 2019a).

6 Chapter Six: Population Genetic Data For 122 DNA Markers for The Saudi Arabian Population Using the ForenSeq™ DNA Signature Prep Kit.

6.1 Overview of experiment

Massively Parallel Sequencing (MPS) systems are now being adopted in many forensic laboratories generating detailed sequence data for different types of markers simultaneously. The ForenSeq™ DNA Signature Prep Kit allows sequencing >150 (Primer Mix A) or >230 markers (Primer Mix B) where users can decide which primer mix will be used (Table 1.5) (Alsafiah *et al.* 2019b). Libraries can be sequenced on the MiSeq FGx instrument and the data analysed using ForenSeq™ Universal Analysis Software (UAS). The system had been under extensive evaluation to measure reliability, reproducibility, sensitivity, mixture discrimination capability and to investigate concordance with CE systems (Xavier and Parson 2017, Almalki *et al.* 2017, Devesse *et al.* 2018, Köcher *et al.* 2018). In addition, the system provides higher degree of recovery from degraded samples (Almohammed *et al.* 2017), and improves the resolution in relationship testing (Ma *et al.* 2016, Li, R. *et al.* 2019) more than the CE systems do.

The system was employed to solve the first court Dutch case where the CE system used concluded inconclusive DNA evidence. In addition, the Institut National de Police Scientifique (INPS, France) has implemented the system for casework in 2017 and started to feed the national databased in 2018. In April 2019, the SWGDAM extended the guidelines to cover the interpretation of STR data generated by MPS systems (SWGDAM 2019). As a result, Verogen's MPS system is now approved by the FPI and the company has initiated a collaboration with Cellmark laboratories to establish an MPS centre in the UK.

Although MPS systems are well established for medical research in Saudi Arabia through the Saudi Human Genome Project (SHGP) launched in 2013 (Abedalthagafi 2019) and released about 109 publications up to date (<https://genomics.saudigenomeprogram.org/en>), the systems have not been used for forensic applications.

So far, little data has been published about the Middle East region (136 samples) (Phillips *et al.* 2018a), one study was about the population of Saudi Arabia (89 samples) (Khubrani *et al.* 2019b), and one study about the Qatari population (150 samples) (Almohammed and Hadi 2019). Despite that the Saudi population has been studied using this kit (Khubrani *et al.* 2019b), STRs like D12S391, D2S1338 and D21S11 have shown higher number sequence-based variants for the same sized-based allele that necessitates sequencing more samples to generate better allele frequencies estimates (Gettings *et al.* 2016, Gelardi *et al.* 2014). In addition, sequence-based data for SE33, which is included in the kit but not reported by the ForenSeq™ UAS, has not been studied (Alsafiah *et al.* 2019b).

6.2 Aims of the study

The main aim of this part is to generate sized-based Saudi population data for four additional STRs (PentaE, PentaD, D6S1043, and D4S2408), which is not provided by STR kits used in previous chapters (Chapters 3 and 5), and for 94 identity informative SNPs (iiSNPs) (Table 6.1). The second aim was to generate and sequence-based data for autosomal DNA markers combined in the ForenSeq™ DNA Signature Prep Kit including the SE33 locus. Both types of data will be statistically evaluated for forensic applications in Saudi Arabia, which included Hardy-Weinberg equilibrium (HWE), linkage

disequilibrium (LD) and other forensic parameters. Lineage makers included in the kit (7 X-STRs and 24 Y-STRs) were not part of the project and were not analysed.

Finally, reporting any novel allele sequences and novel variants in the flanking region that were observed in the Saudi population.

Table 6.1. Identity informative SNPs included in the ForenSeq™ DNA signature prep kit. The table shows the amplicon sizes and the chromosomes of 94 iSNPs included in this study (Verogen 2018a).

Locus	Amplicon Length (bp)	Chr.	Locus	Amplicon Length (bp)	Chr.	Locus	Amplicon Length	Chr.
rs10495407	109	1	rs917118	109	7	rs1528460	115	15
rs1294331	85	1	rs10092491	116	8	rs1821380	118	15
rs1413212	64	1	rs2056277	104	8	rs8037429	63	15
rs1490413	98	1	rs4606077	151	8	rs1382387	89	16
rs560681	90	1	rs763869	85	8	rs2342747	104	16
rs891700	115	1	rs1015250	117	9	rs430046	119	16
rs1109037	118	2	rs10776839	103	9	rs729172	104	16
rs12997453	100	2	rs1360288	119	9	rs740910	113	17
rs876724	119	2	rs1463729	99	9	rs8078417	143	17
rs907100	115	2	rs7041158	115	9	rs938283	98	17
rs993934	120	2	rs3780962	94	10	rs9905977	170	17
rs1355366	119	3	rs735155	170	10	rs1024116	98	18
rs1357617	120	3	rs740598	120	10	rs1493232	75	18
rs2399332	157	3	rs826472	153	10	rs1736442	153	18
rs4364205	98	3	rs964681	105	10	rs9951171	119	18
rs6444724	120	3	rs10488710	118	11	rs576261	76	19
rs1979255	102	4	rs1498553	111	11	rs719366	170	19
rs2046361	120	4	rs2076848	118	11	rs1005533	158	20
rs279844	167	4	rs901398	90	11	rs1031825	126	20
rs6811238	120	4	rs10773760	99	12	rs1523537	117	20
rs13182883	169	5	rs2107612	103	12	rs445251	119	20
rs159606	104	5	rs2111980	94	12	rs221956	97	21
rs251934	97	5	rs2269355	65	12	rs2830795	114	21
rs338882	157	5	rs2920816	157	12	rs2831700	79	21
rs717302	110	5	rs1058083	76	13	rs722098	101	21
rs13218440	170	6	rs1335873	109	13	rs914165	156	21
rs1336071	120	6	rs1886510	116	13	rs1028528	78	22
rs214955	120	6	rs354439	170	13	rs2040411	68	22
rs727811	115	6	rs1454361	118	14	rs733164	120	22
rs321198	165	7	rs4530059	170	14	rs987640	120	22
rs6955448	120	7	rs722290	101	14			
rs737681	120	7	rs873196	114	14			

6.3 Objectives

9- Prepare the samples for the library preparation stage by bringing the concentrations to 0.2 ng/μl.

10- Library preparation using the ForenSeq™ DNA Signature Prep Kit following the manufacturer's protocol (Verogen 2018a), except the volume of pooled

normalised library (PNL) that was increased to 12 µl as used in (Devesse *et al.* 2018).

11- Sequencing the libraries using a MiSeq FGx™ instrument following the manufacturer's protocol (Verogen 2018c).

12- Use the ForenSeq™ UAS following the manufacturer's default setting (Verogen 2018b) for the data analysis and for generating the samples' reports and the Flanking Region Report.

13- Additional analysis using the STRait Razor v3.0 (SR) (Woerner *et al.* 2017) for bioinformatical concordance, flanking region variants not highlighted by the ForenSeq™ UAS, and for SE33 sequence-based data.

14- Study the sequence variants generated by this study and compare them with previously reported variants in the Middle East region (Phillips *et al.* 2018a), in the Saudi Arabian population (Khubrani *et al.* 2019b) and in the Qatari population (Almohammed and Hadi 2019). This is for reporting any specific population variants and to assess the novelty of any sequence-based allele.

15- A statistical evaluation for the size-based, sequence-based data (repeat region sequences and repeat + flanking regions) for the population of Saud Arabia.

6.4 Materials and Methods

This chapter focused on aSTRs (Table 1.5) and iiSNPs (Table 6.1) included in the ForenSeq™ DNA Signature Prep Kit. As SE33 data can be obtained from the FASTAQ files using the STRait Razor (SR), the SE33 data was analysed too. However, the results of the SE33 analysis will be in a separate sub-heading (Section 6.5.7). Materials and methods used in this part are described in Sections 2.6 and 2.7.

6.5 Results

A total of 94 male samples from the population of Saudi Arabia were diluted to the appropriate concentrations. A positive and negative controls were also added (96 samples in total), and libraries were prepared for the pooling step. Libraries were then pooled, denatured, transferred to the reagent cartridge, and were sequenced on the MiSeq FGx instrument.

6.5.1 Run metrics, sequencing results, and depth of coverage (DoC)

The run metric indicators showed 958 K/mm² cluster density, 91.98% of clusters passed the Illumina chastity filter, 0.188% for phasing, and 0.097% for pre-phasing. These figures are within the recommended values that are 400–1650 K/mm², ≥ 80%, ≤ 0.25%, and ≤ 0.15% respectively (Verogen 2018b). In addition, quality metric of the run showed that all indicators (read 1, read 2, index 1 and index 2) passed the quality filter (Figure 6.1).



Figure 6.1. Run metric indicators of the sequencing results. The indicators of the sequencing showed that the average quality of the generated reads is within the optimal ranges.

The 96 samples were sequenced, and 121 autosomal DNA markers were analysed using the ForenSeq™ UAS. The positive control showed a full profile for the 121 loci analysed and the negative control sample performed as expected. Seven out of the 94 samples were eliminated after the primary analysis due to poor coverage, while the rest (87 samples) were further analysed.

This study was able to achieve full profiles in 76/87 samples using the default setting of analytical threshold (AT) and interpretation threshold (IT). Table 6.2 summaries the 11 samples that showed incomplete profiles.

Table 6.2. Samples with partial DNA profiles for the 27 aSTRs and the 94 iiSNPs. Shaded cells represent sequences below the default thresholds. All samples presented here had lower average reads count comparing to other sample. This has led to allele drop out in PentaE, rs1357617, rs2920816, and rs1736442. The D22S1045 was previously genotyped in Chapter 3 and all samples presented in the table had heterozygous genotypes. Due to the lower coverage of samples presented here and the lower allele count ratio (ACR) feature of D22S1045, the absence of the second allele in samples 4,7 and 10 was considered as alleles drop out not discordance.

sample	PentaE		D22S1045		rs1357617		rs2920816		rs1736442	
	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										

The average of total number of reads for autosomal markers analysed in this study was 72,166 per sample. The average reads count for aSTRs ranged from 2936 for the TH01 to 173 reads for the D5S818 (Figure 6.2) and for iiSNPs it ranged from 1320 for rs1109037 to 36 for rs1736442 (Figure 6.3).

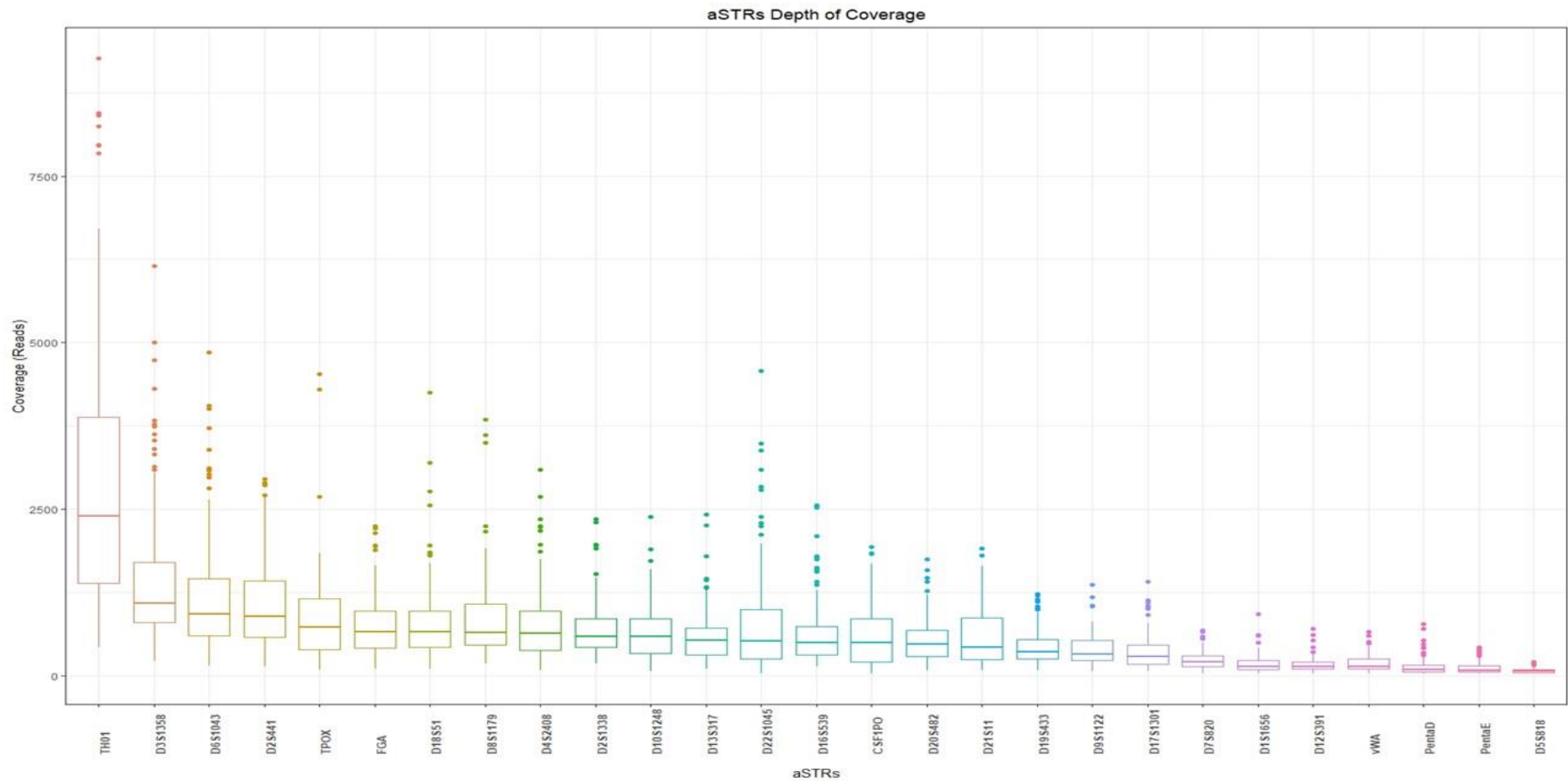


Figure 6.2. Depth of coverage for 27 aSTRs analysed in this study. The average reads count was 673 for all aSTRs that ranged from 173 reads for D5S818 to 2936 reads for TH01. In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).

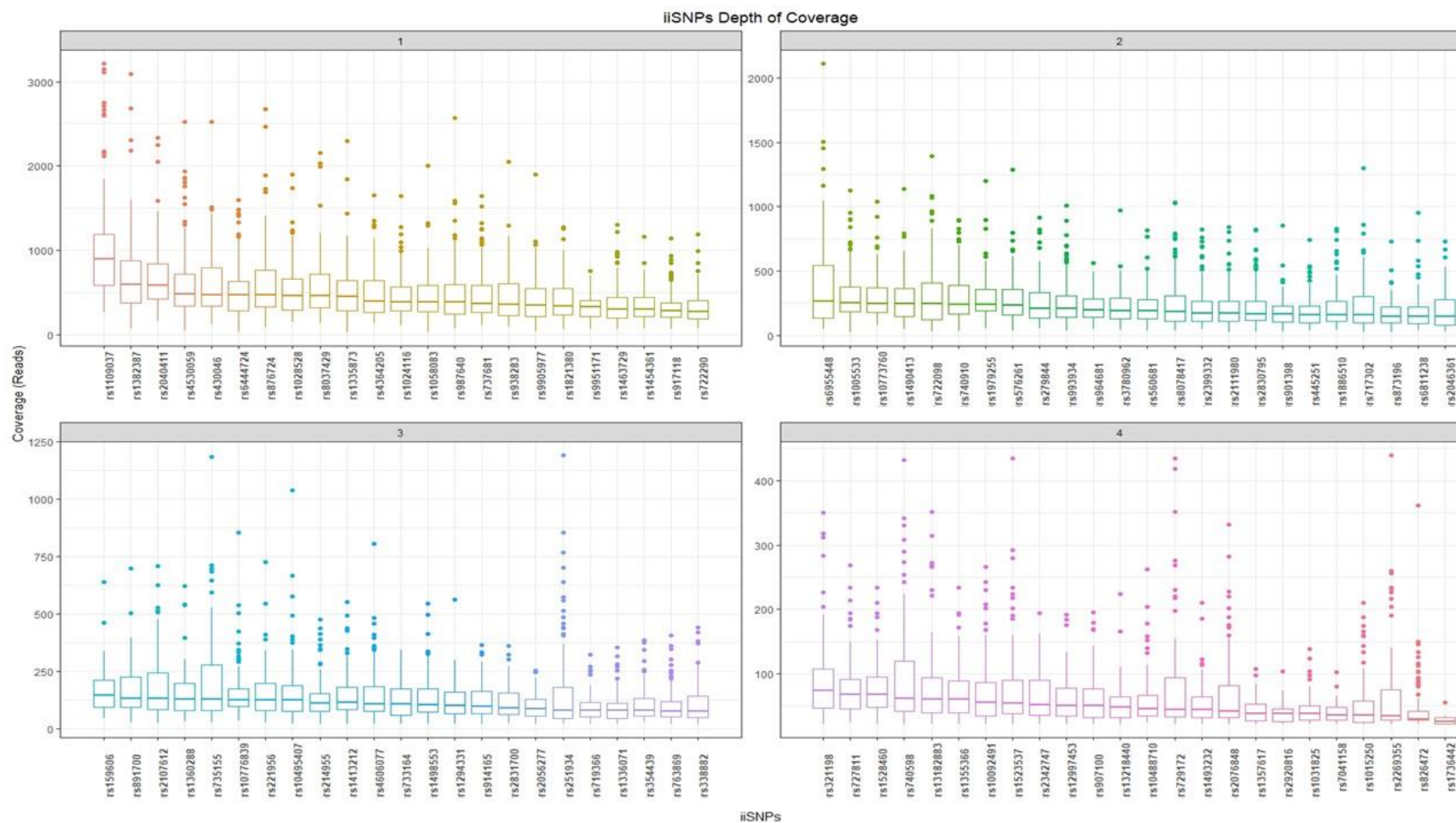


Figure 6.3. Depth of coverage for 94 iiSNPs analysed in this study. The average was 120 reads for all iiSNPs that ranged from 36 for rs1736442 to 1320 reads for rs1109037. In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).

Allele count ratio (ACR) is an alternative description of heterozygous balance in CE systems. All aSTRs showed >60% ACR where the D17S1301 showed 92.5% as the highest ACR average, D22S1045 had the least ACR average of 65.5% and the rest of aSTRs were from 73.7% to 90.7% ACR (Figure 6.4). Remarkably, four heterozygous samples at D22S1045 showed lower ACR (2.78% to 13.90%), two of which had higher stutter ratios of the smaller allele than the ACRs of the true alleles (Table 6.3).

Table 6.3. The four samples that showed lower ACRs at D22S1045. The table shows the CE data, ForeSeq data (including the true alleles, coverage and the ACRs) and the n-4 stutter of allele 1 (including coverage of the -4 stutter and stutter ratios). The four samples showed relatively lower ACRs, two of which (shaded rows) had stutter ratios of the n-4 stutter of allele 1 greater than the ACR of the second true allele (allele 2).

CE data		ForenSeq data					n-4 Stutter of Allele 1	
Allele 1	Allele 2	Allele 1	Coverage 1	Allele 2	Coverage 2	ACR	Coverage	Stutter Ratio
11	15	11	1149	15	32	2.78%	39	3.40%
11	16	11	1330	16	40	3%	50	3.80%
11	16	11	3486	16	486	13.90%	116	3.30%
11	16	11	2838	16	312	10.90%	94	3.30%

The average ACR of all iSNPs were >60% with an exception of the rs6955448 SNPs that showed an average of 40% ACR (Figure 6.5).

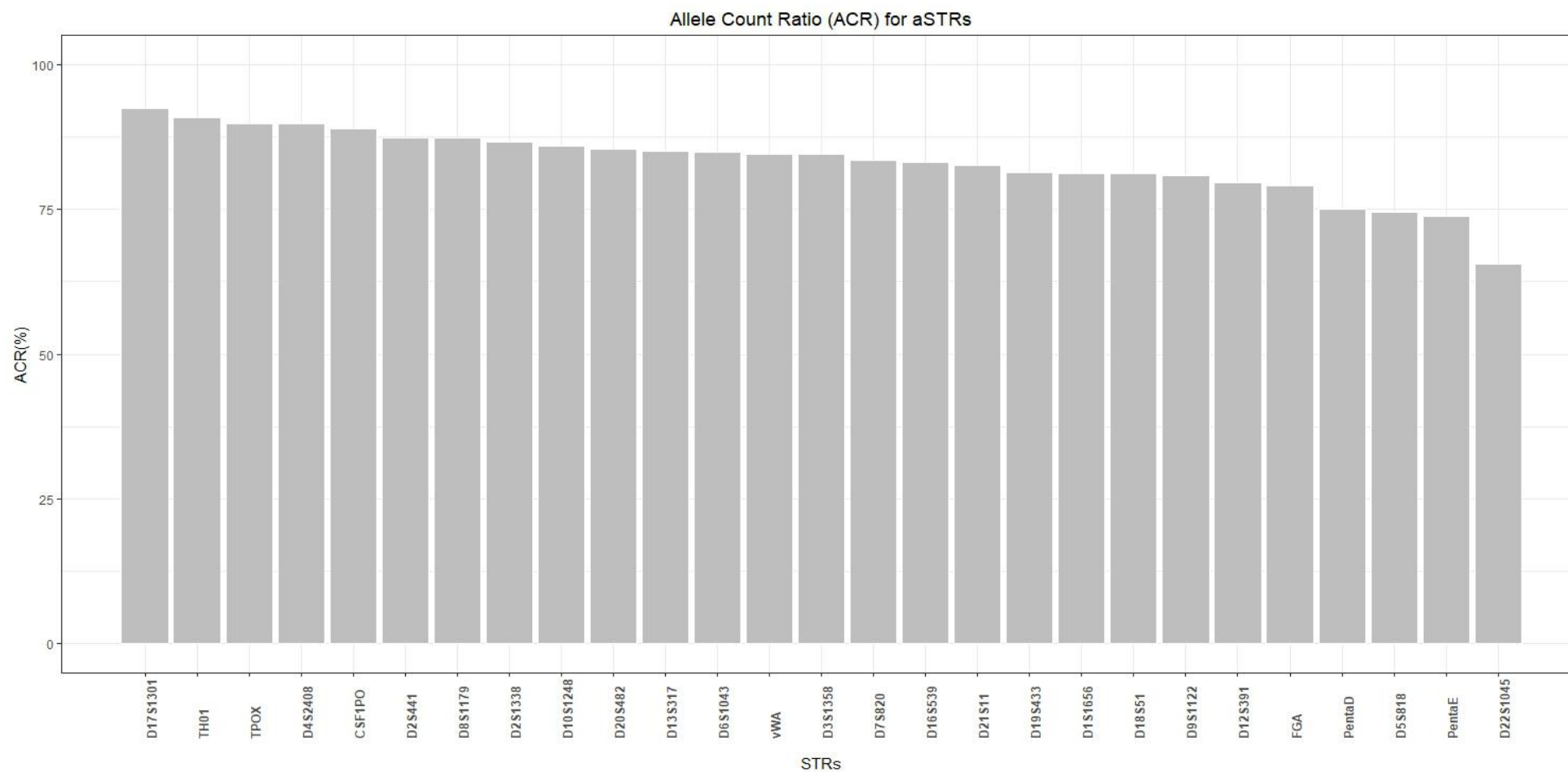


Figure 6.4. Average ACRs of 27 aSTRs. The ACRs of all aSTRs were >60% and ranged from 92.5% for D17S1301 to 65.5% for D22S1045.

Allele Count Ratios (ACR) for iSNPs

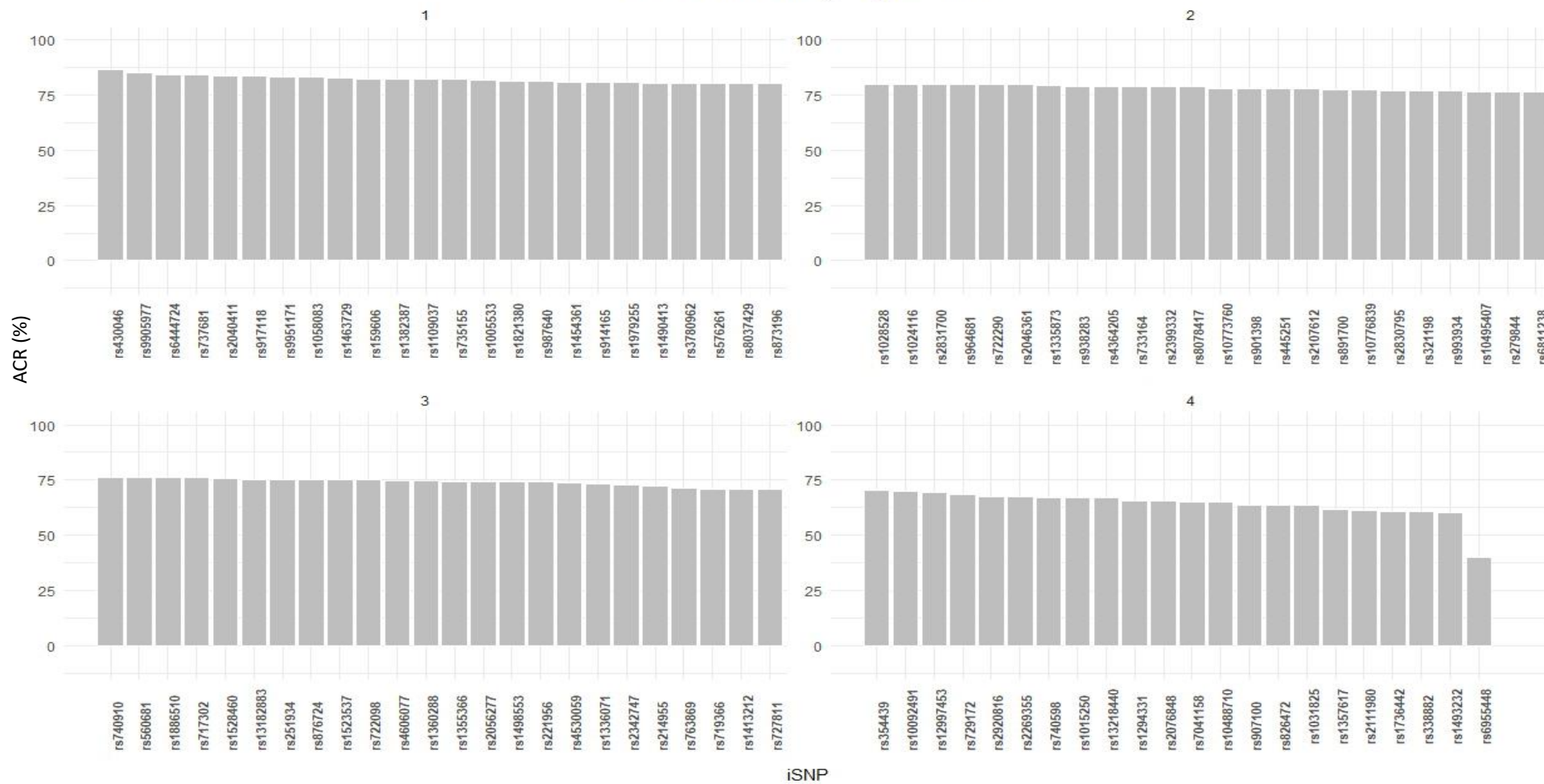


Figure 6.5. Average ACRs of 94 iiSNPs. All iiSNPs showed >60% ACRs except rs6955448 SNPs that showed an average of 40% ACR.

The stutter ratios for the 27 aSTRs were assessed that ranged from 0.6% for allele 6 in TPOX to 31.4% for allele 30 in FGA (Figure 6.6). In addition, allele variations of x.1, x.2 and x.3 showed less stutter ratios, as expected, comparing to the ratios of x-1, x-2, and x-3 alleles respectively.

6.5.2 Sequence variations

All sequence-based data are presented in Appendix 5 using the default output of the UAS software. A total of 638 sequence-based alleles (396 from the aSTRs and 242 alleles from the iiSNPs) were observed in this study (Appendix 5). This represents an average increase of 53.4% in the number of observed alleles for the aSTRs and 28.7% for iiSNPs.

Nineteen aSTRs presented greater number of observed alleles, 13 of which had more alleles based on the repeat region sequences and 8 of which had more alleles based on the flanking region sequences (two aSTRs had variants in both regions) (Figure 6.7). Examining the repeat region variations, the D2S1338 locus showed 181.8% the highest percentage of increase in the number of observed alleles (31 sequences and had 11 alleles based on the size) and the D12S391 locus showed the greatest number of observed alleles of 43 sequences (168.75% increase). Allele 23 at D12S391 reported the highest number of observed sequences (seven variants) (Appendix 5). By including the flanking regions, D21S11, D7S820, D2S441, D16S539, D20S482, D5S818, D13S317 and PentaD showed more alleles (Figure 6.7). D7S820 had 100% more alleles due to the presence of two SNP variants, rs7789995-T (GRCh38-Chr7:84160204) in 10/14 sequences and rs16887642-T (GRCh38-Chr7:84160286) in 3/14 sequences (one allele did not show any of the variants) (Appendix 5).

Stutter Ratios for the aSTRs

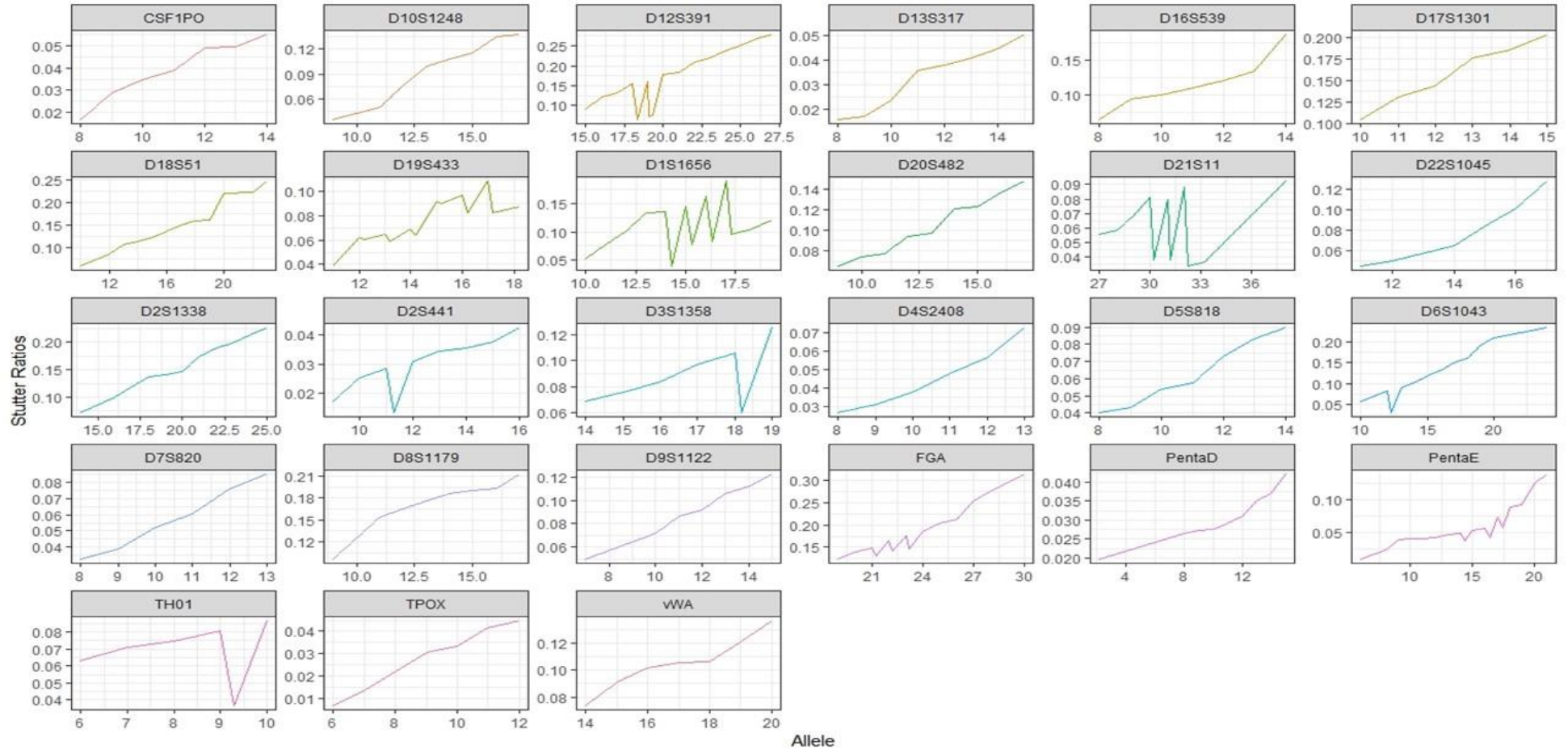


Figure 6.6. Averages of stutter ratios for the 27 aSTRs. Each STR is represented by a plot and the x-axis represents alleles and the y-axis represent stutter ratios. Stutter ratios ranged from 0.6% for allele 6 in TPOX to 31.4% for allele 30 in FGA. Allele variants of x.1, x.2 and x.3 were plotted as x.25, x.50 and x.75.

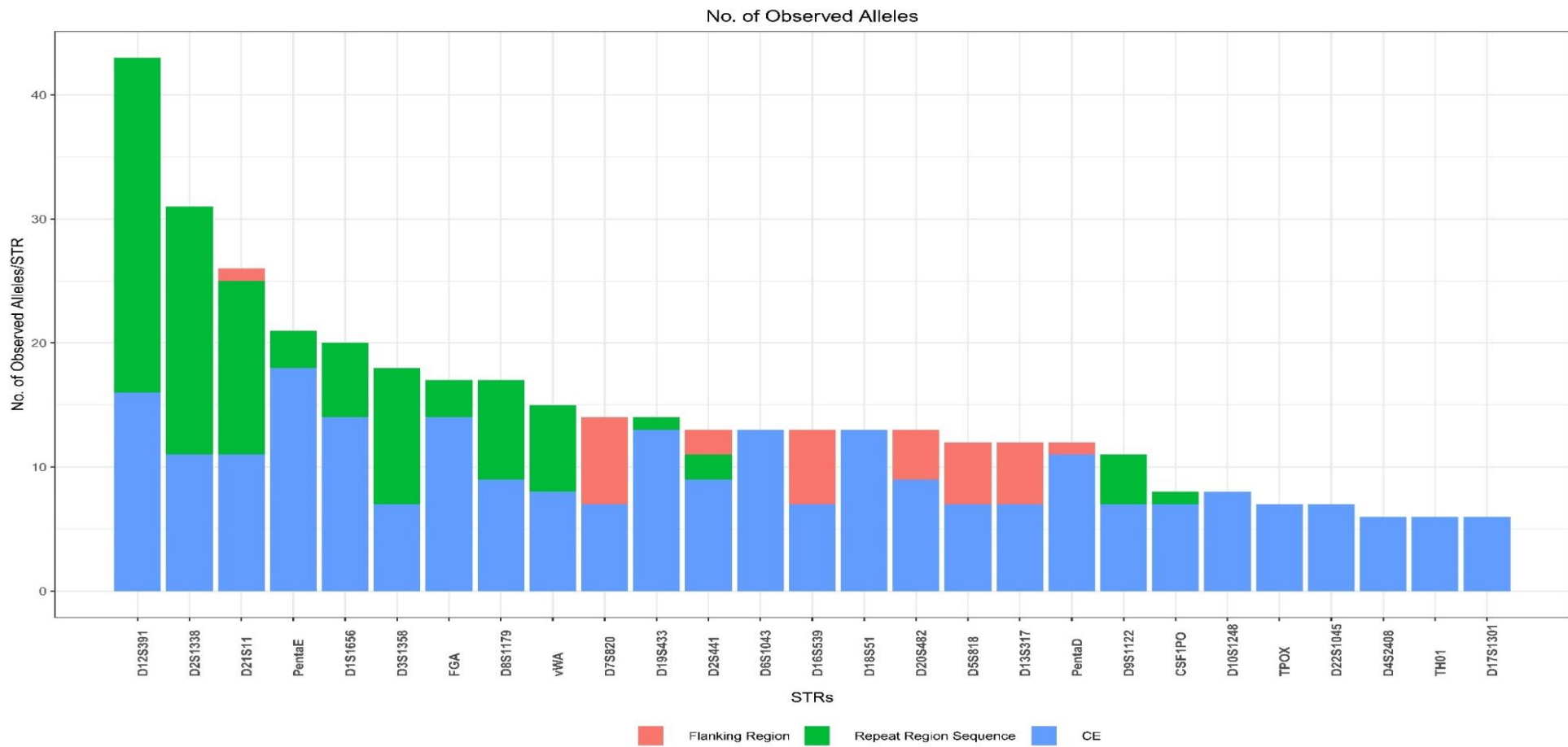


Figure 6.7. The number of observed alleles by sequencing. Nineteen aSTRs presented a greater number of observed alleles, 13 of which had more alleles based on the repeat region sequences (green), 8 aSTRs had more alleles based on the flanking region sequences (red) (two aSTRs had variants in both regions), and 8 aSTRs did not show difference in the number of observed alleles.

For the iiSNPs, 37/94 had more allele sequences per locus where rs1109037, rs8078417 and rs876724 had the greatest number of observed sequences of 7, 5 and 5 sequence variants respectively (Appendix 5). Some iiSNPs had additional variants at the flanking regions covered in the Flanking Region Report; however, they showed no additional alleles or showed a smaller number of alleles comparing to other iiSNPs with the same number of variants per amplicon. This was due to association between the target iiSNPs and variants in the flanking region (allele of one SNP perfectly predicts an allele of another SNP). Perfect associations were observed between rs6955448-T with rs6950322-A, between rs6955448-C with rs6950322-G; between rs430046-C with rs409820-C and rs430044-C, between rs430046-T with rs409820-A and rs430044-T, leading to the observation of only two alleles in both target iiSNPs (target SNP is underlined) (Table 6.4). Another associations between rs4606077-T with rs1869434-G, between rs4606077-C with rs1869434-A; between rs445251-G with rs369438-G, and between rs445251-C with rs369438-A, resulting in a smaller number of observed alleles comparing to other iiSNPs with the same number of variants per amplicons (target SNP is underlined) (Table 6.4).

Table 6.4. Perfect association between the target iiSNPs and variants in the flanking region. This table shows association that was noticed between the target iiSNPs and variants in the flanking region. Black colour indicates the target iiSNPs and the blue colour indicates variants within the flanking region. SNPs that showed perfect association are underlined.

Target iiSNP	Microhaplotype	iiSNP and Variant Reference SNP
rs6955448	<u>A</u> G A <u>I</u>	<u>rs6950322</u> <u>rs140855431</u> <u>rs143117431</u> <u>rs6955448</u>
	<u>G</u> G A <u>C</u>	<u>rs6950322</u> <u>rs140855431</u> <u>rs143117431</u> <u>rs6955448</u>
rs430046	<u>C</u> C C	<u>rs409820</u> <u>rs430044</u> <u>rs430046</u>
	<u>A</u> T <u>I</u>	<u>rs409820</u> <u>rs430044</u> <u>rs430046</u>
rs4606077	<u>I</u> T <u>G</u>	<u>rs4606077</u> <u>rs58774517</u> <u>rs1869434</u>
	<u>I</u> T <u>G</u>	<u>rs4606077</u> <u>rs58774517</u> <u>rs1869434</u>
	<u>I</u> T <u>G</u> C	<u>rs4606077</u> <u>rs58774517</u> <u>rs1869434</u> <u>rs975955864</u>
	<u>C</u> C <u>A</u>	<u>rs4606077</u> <u>rs58774517</u> <u>rs1869434</u>
rs445251	<u>C</u> T <u>G</u> <u>G</u>	<u>rs117702247</u> <u>rs535095356</u> <u>rs445251</u> <u>rs369438</u>
	<u>C</u> T <u>C</u> <u>A</u>	<u>rs117702247</u> <u>rs535095356</u> <u>rs445251</u> <u>rs369438</u>
	<u>T</u> T <u>C</u> <u>A</u>	<u>rs117702247</u> <u>rs535095356</u> <u>rs445251</u> <u>rs369438</u>

Fourteen variants at the flanking regions of two aSTRs (PentaD, D21S11) and of 12 iiSNPs rs1109037 (two variants), rs1979255, rs917118, rs4606077, rs1015250, rs735155, rs1335873, rs8078417, rs1523537, rs914165 and rs733164, were reported by the UAS but were not highlighted (Table 7.5). These variants were reported by the SR as “Novel Sequences”, eleven of which already have rs identifiers in the dbSNP database, two were observed in the Saudi population (Khubrani *et al.* 2019b), while four are novel (Table 7.5).

Table 6.5. Variants at the flanking region of two aSTRs and of 11 iiSNPs. The table shows 14 variants identified in this study which were reported by the UAS but were not highlighted in blue. The table presents the marker's name, allele call (CE), rs identifiers if exist, GRCh37 location reported by the UAS, number of observation (Obs. #), and the comprehensive nomenclature as recommended by the ISFG (Parson *et al.* 2016). It also indicates if a variant was previously observed in the Saudi population (Khubrani *et al.* 2019b) or not. Variants in black are the target iiSNPs, in blue variants that were highlighted by the UAS and in red variants that were reported by the UAS in the Flanking Region Report but were not highlighted in blue (see Table 1.6). N/A: no rs identifier were found for the correspondence variant at the dbSNP database. None of these variants was observed in the data of the Qatari population (Almohammed and Hadi 2019).

Marker	Targeted Marker		Variants in the flanking region not highlighted by the UAS				Observation in Saudi population (Khubrani <i>et al.</i> 2019b)
	Name	Allele	rs identifier	GRCh37 location	Obs. #	Comprehensive nomenclature as recommended by the ISFG	
aSTR	PentaD	8	rs927345580	Chr21:45056053	1	PentaD [CE 8]-GRCh38-Chr21-43636100-43636278 (AAAGA)8 rs927345580-C	G>C (Not observed)
	D21S11	32.2	N/A	Chr21: 20554428	1	D21S11 [CE 32.2]-GRCh38-Chr21-19181939-19182111 (TCTA)5 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)12 TA TCTA 19182110-C	A>C (Not observed)
iiSNPs	rs1109037	G	N/A	Chr.2:10085752	1	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs1109037-G; 9945623-G ; rs183533496-C; rs1109038-G	A>G (Not observed)
		G	rs999755320	Chr.2:10085721	2	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs999755320-T ; rs1109037-G; rs183533496-C; rs1109038-G	C>T (observed)
	rs1979255	G	rs190924736	Chr.4:190318065	1	rs1979255 [CE G]-GRCh38-Chr4:189396884-189396929 rs1979255-G; rs190924736-T	C>T Not observed
	rs917118	T	rs1431710768	Chr.7:4456981	1	rs917118 [CE T]-GRCh38-Chr7:4417342-4417409 rs917118-T; rs1431710768-C	A>C (Not observed)
	rs4606077	T	rs975955864	Chr.8:144656787	1	rs4606077 [CE T]-GRCh38-Chr8:143574562-143574669 rs4606077-T; rs58774517-C ; rs1869434-G; rs975955864-C	G>C (Not observed)
	rs1015250	C	rs1307278892	Chr.9:1823783	1	rs1015250 [CE C]-GRCh38-Chr9:1823726-1823793 rs6475200-G ; rs1015250-C; rs1307278892-C , rs145984676-C	G>C (observed)
	rs735155	A	rs1003612513	Chr.10:3374201	1	rs735155 [CE A]-GRCh38-Chr10:3331961-3332088 rs1003612513-A ; rs79799511-G ; rs735155-A; rs373487413-A ; rs7905965-T	G>A (Not observed)
	rs1335873	A	rs1021428287	Chr.13:20901709	1	rs1335873 [CE A]-GRCh38-Chr13:2032751-20327614 rs1335873-A; rs1021428287-T	G>T (Not observed)
	rs8078417	C	N/A	Chr.17:80461911	1	rs8078417 [CE C]-GRCh38-Chr17:82503992-82504093 rs78650971-G ; rs182919351-C ; rs567092265-C ; rs138630479-G ; 82504035-T ; rs559299986-G ; rs8078417-C	C>T (Not observed)
	rs1523537	C	N/A	Chr.20:51296113	1	rs1523537 [CE C]-GRCh38-Chr20:52679563-52679632 52679574-G ; rs538906241-G ; rs77195753-A ; rs1523537-C	T>G (Not observed)
	rs914165	A	rs192267746	Chr.21:42415913	1	rs914165 [CE A]-GRCh38-Chr21:41043962-41044005 rs192267746-C ; rs914165-A; rs755095-C	G>C (Not observed)
rs733164	A	rs1361542862	Chr.22:27816752	1	rs733164 [CE A]-GRCh38-Chr22:27420770-27420849 rs1361542862-T ; rs733164-A	A>T (Not Observed)	

6.5.3 The impact of sequence variations on discrimination power and heterozygosity

Using the size-based data, the PoD of the 19 aSTRs ranged from 84.9% for D9S1122 to 98.3% for PentaE with an average of 92.7%, while it ranged from 88.1% for CSF1PO to 99.3% for D12S391 with an average of 95.9% using the sequence-based data. The number of aSTRs that had > 90% power of discrimination (PoD) was increased from 17/27 to 21/27 with sequencing. The PoD improvement in the 19 aSTRs, with additional alleles observed by sequencing, ranged from 11.67% for D9S1122 (PoD from 84.9% to 94.9%) to 0.004% for PentaD (PoD from 95.65377% to 95.65789 %). PentaE was the most informative locus with the CE data (98.3% PoD), but the sequence-based data showed that D12S391 (99.3% PoD) and D2S1338 (98.6% PoD) have become more informative (Figure 6.8).

Using the size-based data, the PoD of the 37 iiSNPs ranged from 47.3% for rs740910 to 62.5% for rs560681 with an average of 59.7%, while it ranged from 55.2% for rs1015250 to 86.6% for rs1109037 with an average of 66.7% using the sequence-based data (Figure 6.9). The improvement ranged from 51.9% for rs876724 (PoD from 52.5% to 79.9%) to 1.26% for rs733164 (PoD from 59.5% to 60.2%).

Despite the increase in the observed alleles by sequencing in the 19 aSTRs, the heterozygosity was improved in 17 aSTRs. The heterozygosity was increased by 22.95% (from 70.1% to 86.2%) at D13S317 as the highest improvement, while no improvement was observed in D19S433 and PentaD. Twenty-seven iiSNPs showed an increased heterozygosity, three of which had >50% increase: rs876724 (75.8%), rs9905977 (67.7%) and rs740910 (56.5%).

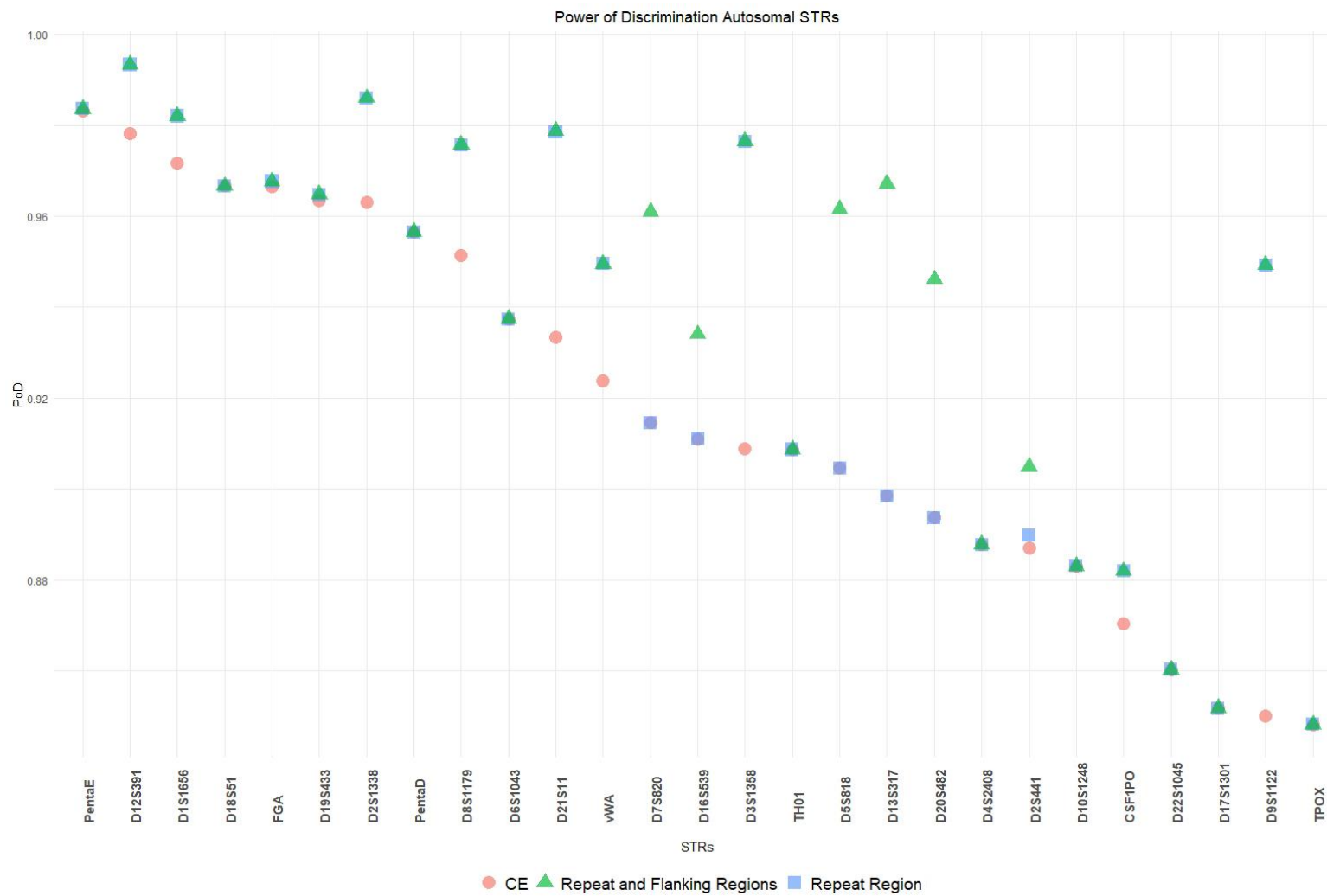


Figure 6.8. Improvements in the discrimination power of the 27 aSTRs.

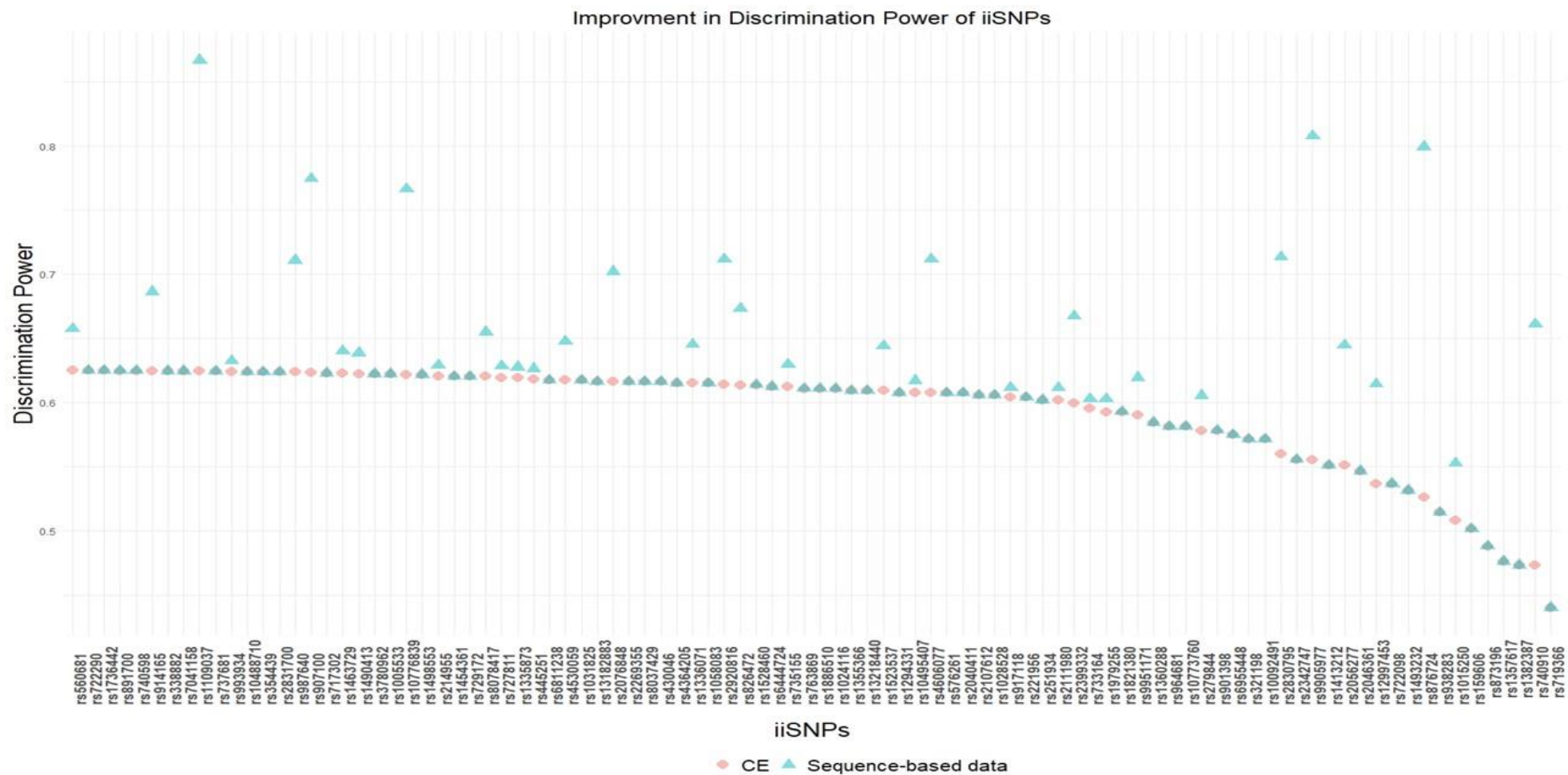


Figure 6.9. Improvements in the discrimination power of the 94 iiSNPs.

Linkage disequilibrium (LD) was tested for 292 pairs of syntenic markers (STR-STR, STR-SNP and SNP-SNP) and no linkage was detected within tested loci after Bonferroni correction (P value > 0.0001) (Appendix 5, Table 10.9).

6.5.4 Evidence of consanguinity and HWE

This study showed some level of consanguinity in our data set due the excess of homozygosity in 21/27 aSTRs (size-based data), 22/27 (repeat region sequence) and 23/27 (repeat and flanking regions); 66/94 (CE data) and 68/94 (including the flanking region) for the iSNPs. This manifestation of consanguinity was supported by the estimation of the inbreeding coefficient (F_{IS}) of 0.03924. However, none of the analysed markers showed significant deviation from HWE after Bonferroni correction (P value > 0.0004) (Appendix 5, Table 10.5 and Table 10.8).

6.5.5 Novel sequences

The novelty assessment was initiated by the SR database, by which 33 alleles were labelled as novel sequences (Table 6.6). Twelve of the alleles were previously reported in Phillips *et al.* (2018a), Almohammed and Hadi (2019) and/or Khubrani *et al.* (2019) (2019b), where the majority (11/12) were mainly observed in the Middle Eastern data set (Phillips *et al.* 2018a), the Saudi population (Khubrani *et al.* 2019b), and/or Qatari population (Almohammed and Hadi 2019). The novelty of the rest of alleles (21 alleles) were further assessed using the GenBank database, by which 8/21 alleles were reported in the GenBank database (Table 6.6). Therefore, this study reported 13 novel sequences, nine of which were due to the repeat sequence (RS) and four were due to flanking region sequence (FS).

Table 6.6. Novel alleles observed in the population of Saudi Arabia. The table show 33 novel alleles assessed based on the SR database. Shaded alleles are novel and have not been observed in (Phillips *et al.* 2018a, Khubrani *et al.* 2019b, Almohammed and Hadi 2019) or in the GenBank. The reason of the novelty types is also shown: repeat sequence (RS) and flanking region sequence (FS).

Nomenclature	Type	Obs. #	Observations			
			Qatari population	Phillips <i>et al.</i> (2018)	Saudi population	GenBank
D2S441 [CE 9]-GRCh38-Chr2-68011918-68012017 (TCTA)9	RS	1	-	ME/EUR/AFR/EA	Observed	MH167314
D2S441 [CE 9]-GRCh38-Chr2-68011918-68012017 (TCTA)9 68011922-A (rs74640515)	FS	1	-	-	-	MK570007
D2S1338 [CE 14]-GRCh38-Chr2-218014856-218014964 (GGAA)8 (GGCA)6	RS	3	-	-	-	MK569967
D2S1338 [CE 20]-GRCh38-Chr2-218014856-218014964 (GGAA)6 GAAA (GGAA)5 (GGCA)8	RS	1	-	-	-	-
D2S1338 [CE 21]-GRCh38-Chr2-218014856-218014964 (GGAA)12 (GGCA)9	RS	1	-	-	-	MH105157
D2S1338 [CE 24]-GRCh38-Chr2-218014856-218014964 (GGAA)2 GGAC (GGAA)16 (GGCA)5	RS	1	-	-	-	MK569981
D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)12	RS	1	-	ME	-	-
D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)13	RS	1	Observed	ME/SCA	Observed	MK990348
D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)2 TCTC (TCTA)14	RS	1	Observed	-	Observed	MK990350
D3S1358 [CE 18.2]-GRCh38-Chr3-45540691-45540820 TCTA (TCTG)3 TC (TCTA)14	RS	1	-	-	-	MK990351
CSF1PO [CE 12]-GRCh38-Chr5-150076318-150076389 ATCT ACCT (ATCT)10	RS	4	-	ME	Observed	-
D6S1043 [CE 12.3]-GRCh38-Chr6-91740160-91740292 (ATCT)8 ATC (ATCT)4	RS	1	-	-	-	-
D9S1122 [CE 7]-GRCh38-Chr9-77073809-77073880 (TAGA)7	RS	1	-	-	-	-
D10S1248 [CE 9]-GRCh38-Chr10 129294226-129294318 (GGAA)9	RS	2	Observed	AFR/ME	Observed	MH167056
vWA [CE 13]-GRCh38-Chr12-5983950-5984049 (TAGA)3 TGGA (TAGA)4 (TAGA)2 5983970-G (rs75219269)	FS	1	-	-	-	MK569942
vWA [CE 17]-GRCh38-Chr12-5983950-5984049 (TAGA)11 (CAGA)5 TAGA	RS	1	-	SCA/AFR	-	MH167086
D12S391 [CE 18]-GRCh38-Chr12-12296981-12297168 (AGAT)10 (AGAC)8	RS	1	-	-	-	MH167121
D12S391 [CE 19]-GRCh38-Chr12-12296981-12297168 (AGAT)9 (AGAC)9 AGAT	RS	1	-	ME	-	MK569923
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)16 (AGAC)6 AGAT	RS	1	-	-	-	-
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 (AGAT)15 (AGAC)7 AGAT	RS	1	-	ME	-	MK569936
D12S391 [CE 26]-GRCh38-Chr12-12296981-12297168 (AGAT)17 (AGAC)8 AGAT	RS	1	-	-	Observed	MH167197
D12S391 [CE 27]-GRCh38-Chr12-12296981-12297168 (AGAT)18 (AGAC)8 AGAT	RS	1	-	EA	-	MH167200
PentaE [CE 14.4]-GRCh38-Chr15-96830996-96831114 (TCTTT)14 TCTT	RS	1	-	-	-	-
PentaE [CE 16.4]-GRCh38-Chr15-96830996-96831114 (TCTTT)16 TCTT	RS	3	-	ME/SCA	Observed	-
PentaE [CE 17]-GRCh38-Chr15-96830996-96831114 (TCTTT)6 TATTT (TCTTT)10	RS	1	-	-	-	-
PentaE [CE 18]-GRCh38-Chr15-96830996-96831114 (TATTT)2 (TCTTT)16	RS	1	-	-	-	-
D16S539 [CE 8]-GRCh38-Chr16-86352664-86352781 (GATA)8 86352761-C (rs11642858)	FS	1	-	-	-	MK570017
D16S539 [CE 12]-GRCh38-Chr16-86352664-86352781 (GATA)12 86352749-C (rs906687856)	FS	1	-	-	-	-
D19S433 [CE 14]-GRCh38-Chr19-29926205-29926352 (CCTT)13 CCTA CCT TTT CCTT 29926229-29926230 DEL (rs745607776)	FS	1	-	-	-	-
D21S11 [CE 32]-GRCh38-Chr21-19181939-19182111 (TCTA)6 (TCTG)7 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)11	RS	1	-	-	-	-
* D21S11 [CE 32.2]-GRCh38-Chr21-19181939-19182111 (TCTA)5 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)12 TA TCTA 19182110-C (no rs identifier)	FS	1	-	-	-	-
D21S11 [CE 38]-GRCh38-Chr21-19181939-19182111 (TCTA)10 (TCTG)8 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA (TCTA)12	RS	1	-	-	-	-
* PentaD [CE 8]-GRCh38-Chr21-43636100-43636278 (AAAGA)8 43636172-C (rs927345580)	FS	1	-	-	-	-

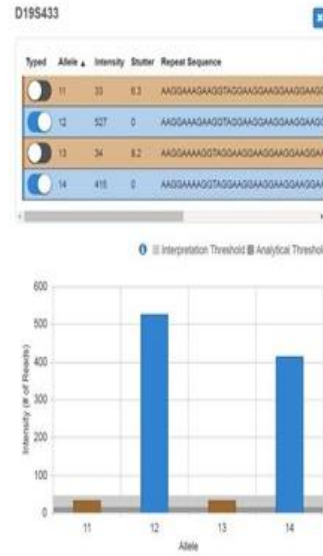
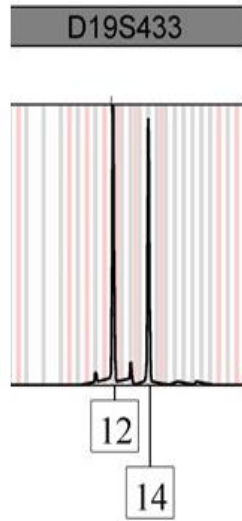
ME: Middle East, SCA: South Central Asia, EA: East Asian, EUR: European, AFR: African

* The variants C at 19182110 (D21S11) and at 43636172 (PentaD) were not highlighted by the UAS.

6.5.6 Concordance study

Apart of the drop out events presented in Table 6.2; all samples showed 100% chemistry concordance at the common 23 aSTRs. In addition, all samples showed 100% bioinformatical concordance between the UAS and the SR. Interestingly, one sample had size-based alleles of 12,14 at the D19S433 locus using the GlobalFiler™ kit and showed, in this study, the same alleles call when analysed by the UAS and the SR, but was labelled as a novel sequence by the SR. Examining the sequences, the allele 14 had [AAGG] AAA AGG [TAGG] [AAGG]₁₃ in the repeat region, which is allele 14.2, suggesting a deletion of 2 bp in the flanking region (Figure 6.10 A). The sequence revealed an uncommon deletion (*rs745607776* AG DEL, reverse strand) at 3' end of the locus and the SR was not able to call the allele as 14.2 as the deletion is located within the 5' and the 3' anchors (Figure 6.10 B).

A



Allele Call			Sequence
CE	UAS	SR	
14	14	14	AAGG AAAG AAGG TAGG [AAGG]12 AGAGAGGAAGAAAGAGAGAAGATTTTATT
14	14	14	AAGG AAA AGG TAGG [AAGG]13 AGAG ^{rs745607776} GAAGAAAGAGAGAAGATTTTATT
14.2	14.2	14.2	AAGG AAA AGG TAGG [AAGG]13 AGAGAGGAAGAAAGAGAGAAGATTTTATT

B

GAGGCTGC^{5'}AAAAAGCTATAATTGTACCACTGCACTCCAGCCTGGGCAACAGAATAAGATTCTG
 TTGAAGGAAAAGGTAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
 GGAAGGAAGGAGAG^{**}GAAGAAAGAGAGAAGATTTTATT^{3'}CGGGTAATGGGTGCACCAAAA

Figure 6.10. Allele of interest at D19S433. A) shows the genotype of the sample using the GlobalFiler™ kit, the sequencing results using the ForenSeq™ kit, and typical sequences of the alleles 14 and 14.2 comparing to the allele of interest. B) shows the repeat region (blue) and the location of AG deletion (green) and the 5' and the 3' anchors used by the SR (yellow).

6.5.7 Sequence-Based Saudi Population Data for The SE33 Locus

The SE33 sequences of the 87 samples were recovered using the FASTAQ files and SR, 83 of which were within the designated limits (≥ 10 reads and $\geq 20\%$ ACR), and the remaining four samples were recovered manually due lower ACR ($< 20\%$). The ACR of heterozygous sequences ranged from 6.5% to 99.4% and showed an average of 58.6%, the four manually typed samples had ACR of 6.5% for alleles 6.3, 31.2, 8.14% for alleles 14,35.2, 12.17% for alleles 13.3,31.2, and 12.8% for alleles 17,34 (Alsafiah *et al.* 2019b).

The total coverage of the SE33 locus in all samples was 53,956 reads and the average reads number of recovered sequences was 742 reads that ranged from 32 to 2196 reads for alleles 31.2 and 6.3 respectively (Alsafiah *et al.* 2019b).

The number of observed sequence-based alleles was 130% more (69 alleles) comparing to 30 size-based alleles (Figure 6.11). Most sequence variants were observed in x.2 alleles where alleles 27.2 and 30.2 had the highest number of observed sequences (7 sequence variants/allele) (Figure 6.12) (Alsafiah *et al.* 2019b).

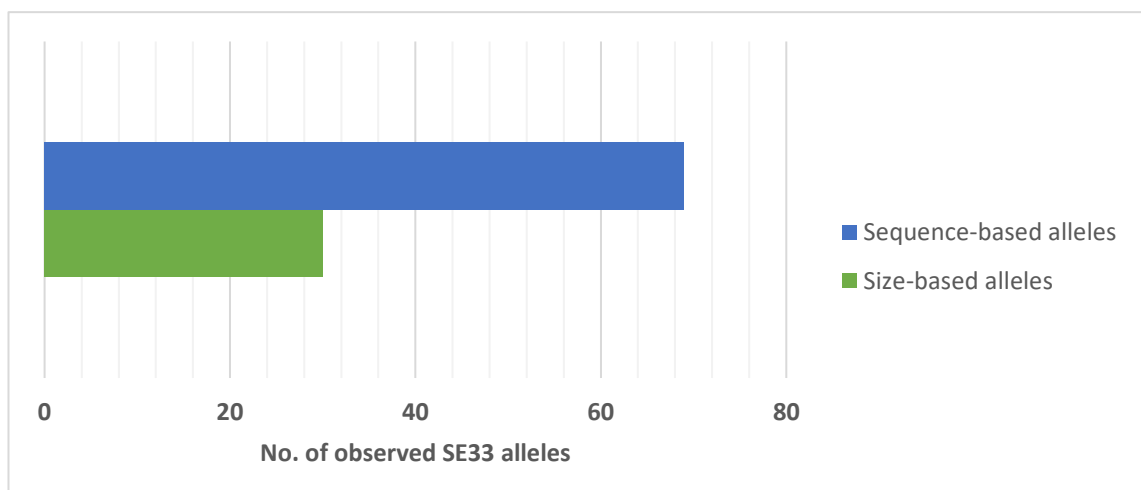


Figure 6.11. The number of observed size and sequence-based SE33 alleles.

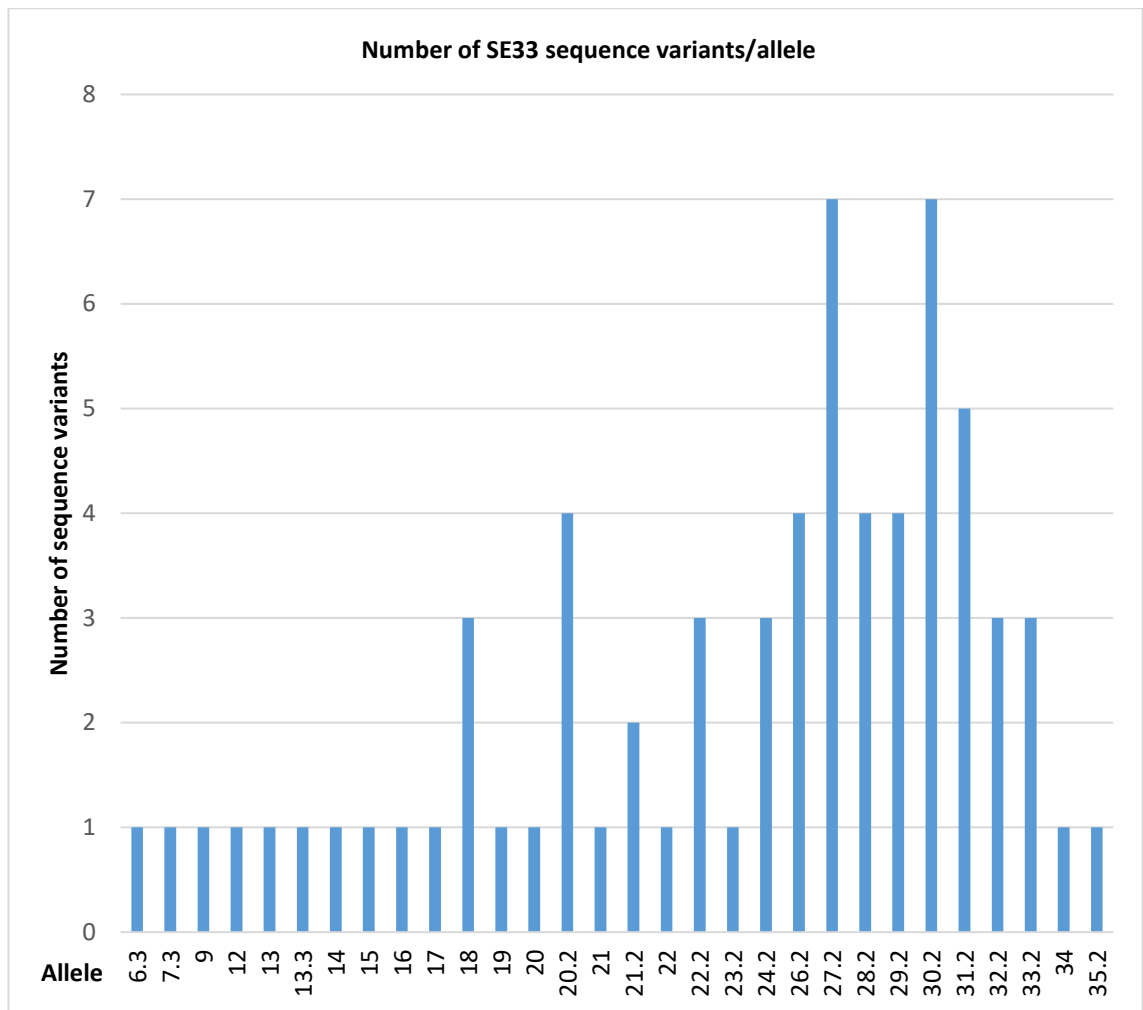


Figure 6.12. The number of SE33 sequence variants observed per allele.

The SE33 motif patterns of the 69 sequences showed that 66 alleles were within the classification of Borsuk *et al.* (2018) and most of these alleles (53 alleles), as expected, had an A0 or A1 motif. Two new motif patterns were observed in three alleles that are shown in Table 6.7. Following on from the earlier study we suggested two new motif IDs (D4 & D5) (Alsafiah *et al.* 2019b). In addition, seven sequences, which fall within the motif classification, but have not been reported in the GenBank database before, were observed (Table 6.7) (Alsafiah *et al.* 2019b).

Table 6.7. Motif patterns of the SE33 locus observed in the samples from the population of Saudi Arabia. A total of 66 allele sequences were within motif patterns classified by Borsuk *et al.* (2018), 53 of which, as expected, had the A0 and A1 motif patterns. Two unreported motif patterns were observed in three alleles and were classified as D4 and D5 motif IDs. Rows in red indicates novel motifs observed in the Saudi population and shaded rows indicates novel alleles that were not reported before in the GenBank database (Alsafiah *et al.* 2019b).

Alleles	Motif	Obs.	ID	Novelty
9-22	CT [CTTT] ₃ C [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	13	A0	Novel sequence (Allele 9)
20.2-33.2	CT [CTTT] ₂ CCTT C [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	40	A1	Reported in (Borsuk <i>et al.</i> 2018)
30.2	CT [CTTT] ₂ CCTT C [CTTT] _n CT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	A3	Reported in (Borsuk <i>et al.</i> 2018)
34	CT [CTTT] ₂ CCTT C [CTTT] _n TT [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	A7	Novel sequence
35.2	CT [CTTT] ₂ CCTT C [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT CTTT	1	A8	Novel sequence
6.3 & 7.3	CT [CTTT] ₃ [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	2	C2	Novel sequence (Allele 7.3)
13.3	CT [CTTT] ₃ C [CTTT] _n C [CTTT] _n [CTTT] ₃ CT [CTTT] ₂	1	B2	Novel sequence
18	CT [CTTT] ₂ C [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	B1	Reported in (Borsuk <i>et al.</i> 2018)
20.2 & 22.2	CT [CTTT] ₃ C [CTTT] _n CT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	2	B3	Novel sequence
26.2	CT [CTTT] ₂ [CCTT] ₃ C [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	B9	Reported in (Borsuk <i>et al.</i> 2018)
28.2	CT [CTTT] ₂ [CCTT] ₂ C [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	A4	Reported in (Borsuk <i>et al.</i> 2018)
27.2	CT [CTTT] ₂ CCTT C [CTTT] _n CTGT [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	C4	Novel sequence
27.2	CT [CTTT] ₂ CCTT C [CTTT] _n TT [CTTT] _n CT TTTT [CTTT] ₂ CT [CTTT] ₂	1	D4*	Novel motif
29.2 & 30.2	CT [CTTT] ₂ CCTT C [CTTT] _n CCTT [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	2	D5*	Novel motif
30.2	CT [CTTT] ₂ C [CTTT] _n TT [CTTT] _n CT [CTTT] ₃ CT [CTTT] ₂	1	B5	Reported in (Borsuk <i>et al.</i> 2018)

*The D4 and D5 IDs were suggested to continue the work done by Borsuk *et al.* (2018).

A single discordance was observed, where the sample had 19,31.2 in the sequence-based data while it had 18,31.2 in the size-based data. The allele 19 had CT [CTTT]₃ C [CTTT]₁₉ CT [CTTT]₃ CT [CTTT]₂ (counted part of the repeat region is in bold) suggesting a deletion of four bp within the flanking region. Examination of the FASTQ file of the sample revealed the presence of rs369314007-DEL, a [TTTT] deletion at 88277355_88277358 (GRCh38), when compared to the reference sequence of the locus. This was further investigated by Sanger sequencing and the deletion was confirmed (Figure 6.13) (Alsafiah *et al.* 2019b).

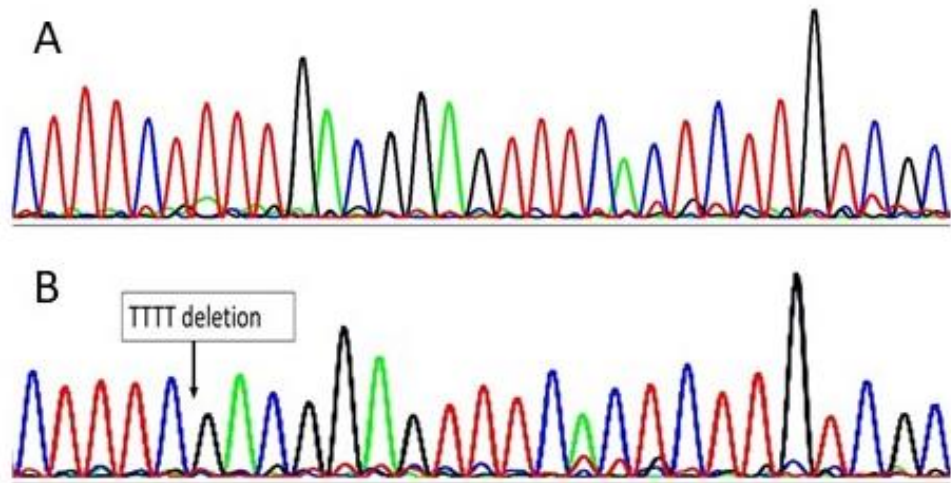


Figure 6.13. Sanger sequencing results for the discordance event. (A) the reference sequence of nucleotides 88277350 – 88277381 (GRCh38) at the 3' flanking region of the SE33 locus. (B) the sequence of the sample showed the discordance event. It shows a TTTT deletion at 88277355_88277358 (GRCh38) that explains the discordance between sequence and CE data.

The data showed that the heterozygosity was increased from 90.8% (79 heterozygous samples) to 91.9% (80 heterozygous samples), and both data were within the expectations of HWE (P value > 0.05) and the power of discrimination increased from 99.3% to 99.7% (Alsafiah *et al.* 2019b).

6.6 Discussion

In this study, 122 autosomal markers included in the ForenSeq™ DNA Signature Prep Kit were analysed for the successfully sequenced samples (87 samples). All run indicators ensured that the average quality of the generated reads is within the optimal ranges (Verogen 2018b). The reads elevation of the n-4 stutter of allele 1 compared to the second allele's read (the true allele), at D22S1045, have been observed previously (e.g. (Gettings *et al.* 2018, Jäger *et al.* 2017)) and was also mentioned in the manufacturer's reference guide (Verogen 2018a).

The rs6955448 SNP showed an ACR of 40%, which was very low compared to other iiSNPs. The SNP has been known as a poorly balanced SNP in many reports (e.g.

(Churchill *et al.* 2016, Guo *et al.* 2017, King *et al.* 2018)), and even in the developmental validation of the MiSeq FGx™ carried out by Illumina (Jäger *et al.* 2017). Guo *et al.* (2017) suggested that the rs6955464 (C: 68%/ T: 32%, Chr.7:4310397, GRCh37) located in the primer binding site (amplicon of rs6955448 starts at 4310285 and ends at 4310404, GRCh37 (Verogen 2018a)) could be the reason of this imbalance. King *et al.* (2018) used an in-house capture panel and found that the presence of the rs6955464-T variant showed reduced ACR (42%) in comparison to the rs6955464-C variant demonstrating a preferential amplification with the rs6955464-C variant. The stutter ratios for aSTRs estimated in this study were as expected showing the correlation between the allele sizes and the complexity of the repeat regions.

The sequencing results showed that 46/121 loci had more observed alleles by sequencing allowing more differentiation between alleles with the same sizes (STRs) or between genotypes with the same iiSNPs. As expected from aSTRs with complex repeat structures like D12S391, D2S1338, D21S11, and D3S1358, they showed more than a two-fold increase in the observed alleles, which improved the random match probabilities greatly from 0.0218 to 0.0066, 0.0370 to 0.0139, 0.0666 to 0.0211 and from 0.0910 to 0.0235 respectively. Despite the significant increase on the number of observed alleles in those loci, this was not reflected in the heterozygosity, with a maximum of 10.1% gain for D21S11 that is obviously due the massive heterozygosity level of those loci even with the size-based data (>77%). D9S1122 had the greatest heterozygosity gain of 15.4% (from 75.8% to 89.6%) due to variants rs4281164-C (A: 58.3% C: 41.7%, Chr. 9: 77073831, GRCh38) within the repeat region.

Variants at the flanking region of aSTRs significantly increased the number of observed alleles in five aSTRs D7S820 (100%), D16S539 (85.7%), D5S818 (71.4%), and

D13S317 (71.4%) reducing the match probabilities by two-fold or more. Notably, the flanking region variants at D13S317 and D5S818 allowed the highest percentage of heterozygosity gain with 18.6% (from 70.1% to 86.2%) and 17.5% (from 70.1 to 85%) respectively. Variants at the flanking region of the target iiSNPs are also reported. This led to that rs1109037 becoming more informative (0.133 MP) than the lowest discriminating aSTRs TPOX (0.151 MP) D17S1301 (0.148 MP) and displayed a higher level of heterozygosity (67.8%) than five aSTRs TPOX (66.66%), D4S2408 (66.66%), TH01 (65.51%), D17S1301 (60.92%) and D22S1045 (60.97%). It is clear, at least in this study, that variants at the flanking region had significant impact on the heterozygosity, especially in iiSNPs.

However, the UAS does not highlight all variants at the flanking region as it looks at specific positions for pre-defined variants that mostly were taken from (King *et al.* 2018) and from the dbSNP database. As the variants still reported in the Flanking Region Report, this study was able to identify 14 additional variants that were not highlighted in blue by the UAS (Table 6.5). King, J *et al.* (2018) studied four major populations African American, East Asian, US Caucasian, and Southwest US Hispanic which may explain why the 14 variants have not been pre-defined in the UAS as they may be restricted to geographical region. Ten/fourteen variants already have been assigned to rs identifiers in the dbSNP database, two of which (rs999755320-T at rs1109037 and rs1307278892-C at rs1015250) were observed in the previous study of population of Saudi Arabia (Khubrani *et al.* 2019b). The rest of reported variants 19182110-C at D21S11, 9945623-G at rs1109037, 82504035-T at rs8078417 and 52679574-G at rs1523537 are novel (Table 6.5).

Perfect associations between SNP variants at four pairs rs6955448- rs6950322 (separated by 48 bp), rs430046-rs409820-rs430044 (separated by 17 and 6 bp respectively), rs4606077-rs1869434 (separated by 11 bp), and rs445251-rs369438 (separated by 40 bp) (target SNP is underlined), were observed in this data set (Table 6.4). All these associations were also observed in the previous study of the population of Saudi Arabia (Khubrani *et al.* 2019b), and is due to physical linkage. Although the variants are known to be closely linked (6 bp – 48 bp apart), other variants are also closely linked but do not show perfect associations. This can be explained by the presence of recombinational hot spot between any two linked but not associated variants or by several mutational events through generations (Carothers and Wright 1992). However, Khubrani *et al.*(2019b) reported another perfect association between rs279844-A with rs279845-T and between rs279844-T with rs279845-A (separated by 68 bp) that was not observed in this study due the observation of the allele rs279844-T, rs279845-T (0.02299 frequency) (Appendix 5) (Table 6.8).

Table 6.8. Perfect association between the target iiSNP and variant in the flanking region observed in Khubrani *et al.* (2019b) but not in this study. In Khubrani *et al.* (2019b), perfect association was observed between rs279844 and rs279845 but was not observed in this study due to the presence of the allele TT at rs279844_rs279845 (shaded). Black colour indicates the target iiSNPs and the blue colour indicates variants within the flanking region. SNPs that showed perfect association are underlined.

Study	Target iiSNP	Microhaplotype	iiSNP and Variant Reference SNP
This study	rs279844	TA	rs279844_rs279845
		TT	rs279844_rs279845
		AT	rs279844_rs279845
Khubrani <i>et al.</i> (2019b)	rs279844	<u>TA</u>	<u>rs279844</u> rs279845
		<u>AT</u>	rs279844 <u>rs279845</u>

This has raised a question whether these associations are also present in other populations. Therefore, all associations were further investigated in the data of five major populations (African, ad Mixed American, East Asian, European and South Asian) generated by the 1000 Genomes Project (Phase 3) using LDlink v3.7.2. (<https://ldlink.nci.nih.gov>) (Machiela and Chanock 2015). The four associations observed in this study and in (Khubrani *et al.* 2019b) (Table 6.4) were observed in the five populations too, with an R^2 (a measure of association between alleles for two SNPs, where 0: variants are completely independent, 1: an allele of one SNP perfectly predicts an allele of another SNP) value of 0.9982 for rs6955448-rs6950322 pair, 1 for rs430046-rs409820-rs430044 pair, 0.9796 for rs4606077-rs1869434 and of 1 for rs445251-rs369438 pair (Machiela and Chanock 2015) (Table 6.9 A-F). Interestingly, the pair rs279844-rs279845 (previously reported in the population of Saudi Arabia (Khubrani *et al.* 2019b), but not observed in this study) (Table 6.8), showed lower R^2 value of 0.7826 due to the observation of the rs279844-T/rs279845-T microhaplotype in 300/5008 samples, 284 of which are from the African population (Table 6.9 F and G). This suggests that the donor of the four samples that showed the TT microhaplotype (in this study) at the rs279844-rs279845 pair may have descended from an African descent.

Table 6.9. Associations between five SNP pairs observed in the Saudi population and in five major populations (African, Ad Mixed American, East Asian, European and South Asian) generated by the 1000 Genomes Project (Phase 3) using LDlink v3.7.2. (Machiela and Chanock 2015). (A) is for pairs rs6955448-rs6950322, (B) rs430046-rs409820, (C) rs409820-rs430044, (D) rs430046-rs430044, (E) rs4606077-rs1869434, (F) rs445251-rs369438, (G) rs279844-rs279845 (all populations) and (H) is for rs279844-rs279845 (Africans). Each table shows the haplotypes frequencies across 5008 samples per all population (A-G)/African population (H), D' (an indicator of allelic segregation for two genetic variants. A D' value of 0 presents no linkage of alleles and a D' value of 1 indicates at least one expected haplotype combination is not observed), R2 value, Chi-sq. and p-value (High chi-square statistics and low p-values are evidence that haplotype counts deviate from expected values and suggest linkage disequilibrium may be present). Each table shows a statement for the correlation between the variants of interest and if ($R^2 > 0.1$), the variants are correlated.

A	All populations	rs6950322		Total	Frequency	Haplotypes	Statistics						
			A					G					
		rs6955448	C					1	3408	3409	0.681	C_G: 3408 (0.681)	D': 0.9991
			T					1598	1	1599	0.319	T_A: 1598 (0.319)	R2: 0.9982
			Total					1599	3409	5008		C_A: 1 (0.0)	Chi-sq: 4998.802
			Frequency					0.319	0.68			T_G: 1 (0.0)	p-value: <0.0001
rs6955448(C) allele is correlated with rs6950322(G) allele rs6955448(T) allele is correlated with rs6950322(A) allele													
B	All populations	rs430046		Total	Frequency	Haplotypes	Statistics						
			A					C					
		rs409820	C					4	3246	3250	0.649	C_C: 3246 (0.648)	D': 1
			T					1758	0	1758	0.351	T_A: 1758 (0.351)	R2: 0.9965
			Total					1762	3246	5008		C_A: 4 (0.001)	Chi-sq: 4990.481
			Frequency					0.352	0.64			T_C: 0 (0.0)	p-value: <0.0001
rs430046(C) allele is correlated with rs409820(C) allele rs430046(T) allele is correlated with rs409820(A) allele													
C	All populations	rs430044		Total	Frequency	Haplotypes	Statistics						
			C					T					
		rs409820	A					0	1762	1762	0.352	C_C: 3246 (0.648)	D': 1
			C					3246	0	3246	0.648	A_T: 1762 (0.352)	R2: 1
			Total					3246	1762	5008		A_C: 0 (0.0)	Chi-sq: 5008
			Frequency					0.648	0.35			C_T: 0 (0.0)	p-value: <0.0001
rs409820(A) allele is correlated with rs430044(T) allele rs409820(C) allele is correlated with rs430044(C) allele													
D	All populations	rs430044		Total	Frequency	Haplotypes	Statistics						
			C					T					
		rs430046	C					3246	4	3250	0.649	C_C: 3246 (0.648)	D': 1
			T					0	1758	1758	0.351	T_T: 1758 (0.351)	R2: 0.9965
			Total					3246	1762	5008		C_T: 4 (0.001)	Chi-sq: 4990.481
			Frequency					0.648	0.35			T_C: 0 (0.0)	p-value: <0.0001
rs430046(C) allele is correlated with rs430044(C) allele rs430046(T) allele is correlated with rs430044(T) allele													
E	All populations	rs1869434		Total	Frequency	Haplotypes	Statistics						
			A					G					
		rs4606077	C					3315	22	3337	0.666	C_A: 3315 (0.662)	D': 0.9991
			T					1	1670	1671	0.334	T_G: 1670 (0.333)	R2: 0.9796
			Total					3316	1692	5008		C_G: 22 (0.004)	Chi-sq: 4905.84
			Frequency					0.662	0.33			T_A: 1 (0.0)	p-value: <0.0001
rs4606077(C) allele is correlated with rs1869434(A) allele rs4606077(T) allele is correlated with rs1869434(G) allele													

Table 6.9. continued.

F	All populations	rs369438		Total	Frequency	Haplotypes	Statistics
		C	T				
rs445251	C	2927	0	2927	0.584	C_C: 2927 (0.584)	D': 1
	G	0	2081	2081	0.416	G_T: 2081 (0.416)	R2: 1
	Total	2927	2081	5008		C_T: 0 (0.0)	Chi-sq: 5008
	Frequency	0.584	0.41			G_C: 0 (0.0)	p-value: <0.0001
rs445251(C) allele is correlated with rs369438(C) allele rs445251(G) allele is correlated with rs369438(T) allele							
G	All populations	rs279845		Total	Frequency	Haplotypes	Statistics
		A	T				
rs279844	A	1	2680	2681	0.535	A_T: 2680 (0.535)	D': 0.9991
	T	2027	300	2327	0.465	T_A: 2027 (0.405)	R2: 0.7826
	Total	2028	2980	5008		T_T: 300 (0.06)	Chi-sq: 3919.368
	Frequency	0.405	0.59			A_A:1 (0.0)	p-value: <0.0001
rs279844(A) allele is correlated with rs279845(T) allele rs279844(T) allele is correlated with rs279845(A) allele							
H	African	rs279845		Total	Frequency	Haplotypes	Statistics
		A	T				
rs279844	A	0	690	690	0.522	A_T: 690 (0.522)	D': 1
	T	348	284	632	0.478	T_A: 348 (0.263)	R2: 0.3901
	Total	348	974	1322		T_T: 284 (0.215)	Chi-sq: 515.6841
	Frequency	0.263	0.73			A_A: 0 (0.0)	p-value: <0.0001
rs279844(A) allele is correlated with rs279845(T) allele rs279844(T) allele is correlated with rs279845(A) allele							

The previous studies in the Saudi population using GlobalFiler™ kit (Alsafiah *et al.* 2017) (Chapter 3) and SureID® 23comp kit (Alsafiah *et al.* 2019a) (Chapter 5) showed excess of homozygosity in 20/21 ($F_{IS} = 0.03560$) and in 14/17 ($F_{IS} = 0.02977$) respectively, which revealed some level of consanguinity in the population of Saudi Arabia. D20S482 was the only STR among 38 loci investigated, in those studies, that showed a significant deviation from HWE (P value= 0), which was not clear if this was due null alleles or because of the consanguinity. Although this study cannot eliminate null alleles theory as both kits (SureID® 23comp and ForeSeq) may use the same primer pairs (100% concordance), sequencing results showed the presence of rs77560248-T variant (Chr.20:4525680, GRCh38) at flanking region with 16.67% frequency that increased the heterozygosity by 44.4% improving the P value from 0.03 (size-based) to 0.1 (sequence-based). Despite this improvement in the P value, both theories (null alleles and consanguinity) are still possible.

In this study, the excess of homozygosity was also detected in the sequence-based data of 23/27 for aSTRs and 68/94 of iiSNPs, which was supported by 0.03924 value for inbreeding coefficient (F_{IS}). Khubrani *et al.* (2019b) reported similar figures where 23/27 aSTRs and 63/91 had excess of homozygosity and the F_{IS} was 0.04131.

A total of 33 potentially novel alleles assessed based on the SR database were further investigated. Twelve alleles have been reported before in (Phillips *et al.* 2018a), (Almohammed and Hadi 2019) and/or (Khubrani *et al.* 2019b), eleven of which were observed in the Middle Eastern population, Qatari population and/or the Saudi population. Eight out the 33 alleles were only reported in the National Centre for Biotechnology Information (NCBI) by the STRSeq project (Gettings *et al.* 2017). This study reported 13 novel alleles where the novelty of 9 alleles was due to the repeat sequence (RS) and 4 alleles was due to the variant at the flanking sequence (FS) (Table 6.6). It is believed that when more samples from the Saudi population or from neighbouring countries are sequenced, the novel alleles will be observed and reflect ascertainment bias in the database.

Three additional non-CODIS loci D9S1122, D17S1301 and D20S482 were included in the concordance study, which were previously genotyped using the SureID® 23comp kit (Alsafiah *et al.* 2019a) (Chapter 5). In addition, the sequence data of three iiSNPs rs1736442, rs2920816 and rs719366, which were not covered in the previous publication (Khubrani *et al.* 2019b), are also included.

The 27 aSTRs showed CMP of 6.26E-32 for the size-based data and 6.52E-37 for the sequence-based data that are comparable to combined CMPs estimated for Caucasians (6.28E-32 and 3.63E-36) and for Asians (6.37E-32 and 8.66E-36) (Novroski *et al.* 2016). As expected, the 94 iiSNPs alone showed 1.24E-37 for the CE data and 5.6E-41 by

sequencing providing higher CMP than what aSTRs provided. The 121 autosomal loci combined allowed $1.97E-68$ and $3.65E-77$ using the size-based data and by sequencing respectively. In Khubrani *et al.* (2019b), the aSTRs showed CMP of $2.62E-30$ to $3.49E-34$ and the iiSNPs showed $9.97E-37$ to $6.83E-40$ CMP for size and sequence-based data respectively, which are relatively higher than figures generated from this study. Although this can be, in part, due to exclusion of three iiSNPs in that study, clearly the major impact came from aSTRs. The population of Saudi Arabia is highly structured (Khubrani, *et al.* 2018), and some parts can be distinguished from others even by using aSTRs (Khubrani *et al.* 2019a), and thus variations in the CMP within different data set can be expected in the population of Saudi Arabia.

The SE33 locus showed the lowest ACR of all aSTRs analysed. Among the 87 samples, the four manually recovered samples had the largest size difference between the long and short allele that ranged from 99 bp to 68 bp demonstrating the ACR correlation with the size difference of the heterozygous allele pair. This correlation was observed when 1036 U.S. population samples were sequenced that was attributed to a decline in the reads number of the second allele (Borsuk *et al.* 2018).

A single discordance was observed between the GlobalFiler™ kit data (Chapter 3) and data generated from this study at SE33. Sanger sequencing confirmed the presence of the rs369314007 deletion, which was found to be associated with the A0 motif (Borsuk *et al.* 2018), which is the motif pattern of allele 19 (Alsafiah *et al.* 2019b).

The data showed that the SE33 heterozygosity was increased from 90.8% (79 heterozygous samples) to 91.9% (80 heterozygous samples), and both data were within the expectations of HWE (P value > 0.05). As expected, SE33 still the most informative loci for the population of Saudi Arabia even when using the size-based data (PoD 99.4%)

that was improved to 99.7% by sequencing. This can further improve the CMP obtained from the ForenSeq™ DNA Signature Prep Kit to $5.02E-71$ and $1.01E-79$. Although, the figures emphasize the value of using SE33 in forensic applications especially with mixture analysis and in paternity testing, it was difficult to be confident in sequence-based data for SE33 without CE data support due to variation in reads depth and ACR values (Borsuk *et al.* 2018).

6.7 Conclusion

In this chapter, 87 samples from the population of Saudi Arabia were successfully sequenced using ForenSeq™ DNA Signature Prep Kit with a MiSeq FGx™ instrument and data of 122 autosomal markers (28 aSTRs and 94 iiSNPs) were analysed. It was shown that the 122 autosomal markers presented more discrimination power and heterozygosity by sequencing. Using the kit allowed obtaining the CE data of four aSTRs (PentaE, PentaD, D6S1043, and D13S317) and the sequence-based data of 28 autosomal markers including the most polymorphic well-characterised STR (SE33). In addition, the data of 94 iiSNPs for 87 samples were obtained in one sequencing run that could not be achieved using other SNP genotyping methods like TaqMan® Real-Time PCR assay or SNaPshot assay.

The sequence-based data allowed CMP of $6.52E-37$ for the 27 aSTRs, $5.6E-41$ for the 94 iiSNPs, and of $3.65E-77$ for the 121 autosomal loci combined. The CMP obtained from the ForenSeq™ DNA Signature Prep Kit increased to $1.01E-79$ when including the SE33 taking in account possible allele drop out. Associations between the target iiSNPs and variants at the flanking region were observed in the Saudi population, which were also observed in different populations. These associations have limited or reduced the number of observed sequence-based alleles in five iiSNPs. However, no LD was detected

within the target autosomal markers allowing the maximum benefit of combining the 122 markers.

This study reported 13 novel sequences and 14 variants at the flanking region that were not highlighted in blue in the Flanking Region Reports, which can be added to the pre-defined list of variants at the target iiSNPs. The novel sequences and variants may be observed when samples from the Middle East or more samples from the Saudi population are sequenced.

As MPS systems are being established in Saudi Arabia, sequencing more samples is needed to establish a representative sequence-based database for both aSTRs and iiSNPs. Although the total number of sequenced samples from the Saudi population is 176 samples, the highly structured nature of the Saudi population necessitates the analysis of more samples.

7 Chapter Seven: Evaluation Study of 136 DNA Markers for Kinship Applications in Saudi Arabia.

7.1 Overview of experiment

The data of DNA markers (42 aSTRs and 94 iiSNPs) obtained from Chapters 3,5 and 6 were further assessed for kinship testing in Saudi Arabia. The assessment was carried out to measure the impact on the extra markers could have on the resolution of kinship testing in Saudi Arabia. Different combinations of autosomal markers included in Identifiler Plus (15 aSTRs), GlobalFiler (21 aSTRs), GlobalFiler and SureID23 (38 aSTRs), Fusion 6C and SureID23 (40 aSTRs), ForenSeq DNA Signature Prep kit (27 aSTRs and 94 iiSNPs (121 loci)), all markers (42 aSTRs and 94 iiSNPs (136 loci)) and 94 iiSNPs alone, were used in this study to test five types of relationships (eight different scenarios in total) (Table 2.5).

Testing additional markers increases the number of loci situated on the same chromosome (syntenic loci) and raises concerns regarding their independence (Tillmar and Phillips 2017, O'Connor and Tillmar 2012). Syntenic loci are regarded as independent (unlinked) if they are 50 centimorgans (cM) or more apart (at which point the probability of recombination between them is 0.5) (Lobo and Shaw 2008). As recombination rates vary along chromosomes, using the physical distance (bp) may underestimate or overestimate the genetic distance between syntenic loci (Westen *et al.* 2012). Therefore, family studies have been undertaken to estimate the recombination fraction (RF) between syntenic loci (Westen *et al.* 2012, Liu *et al.* 2014, Budowle *et al.* 2011, Wu *et al.* 2014). However, family studies are expensive and, sometimes, may not be informative enough due to the need of a large number of generations (meioses) and high percentage of heterozygote genotypes (Liu *et al.* 2014,

Phillips *et al.* 2012). An alternative approach employed by Phillips *et al.* (2012) used the high-density multi-point SNP data of HapMap to approximate the genetic distance between syntenic loci to estimate the RFs, which showed RF values similar to those generated using the family studies (Alsafiah *et al.* 2019a).

Three potential types of syntenic pairs resulted from using the 42 aSTRs and 94 iiSNPs: STR-STR, STR-SNP and SNP-SNP pairs, that can impact the LR estimation in kinship testing.

7.2 Aims of the study

This chapter aimed to evaluate that to what extent DNA markers, characterised in previous chapters, can improve the confidence in kinship testing in Saudi Arabia. The evaluation included seven different combinations of markers combined in Identifiler Plus (currently used kit in Saudi Arabia), GlobalFiler, GlobalFiler and SureID 23, Fusion 6C and SureID 23, ForenSeq DNA Signature Prep kit, 94 iiSNPs alone and all markers. Based on the simulation study, this part will feed into guidelines for the Supreme Council of Magistracy in Saudi Arabia in defining the LR threshold that can be used for kinship testing, and for the genetic laboratories in Saudi Arabia regarding the number/type of markers that would allow sufficient differentiation between tested hypotheses.

It also aimed to estimate the genetic distance between syntenic markers located on the same arm (p-p and q-q) using the high-density multi-point SNP data of HapMap as described by Phillips *et al.* (2012) and to calculate the RF using the Kosambi function. Based on the estimated RFs, the study highlighted those syntenic pairs that would have significant impact on the LRs estimations to be considered.

7.3 Objectives

- 1- Creating a hypothetical pedigree using an in-house Excel software that contains all relationships to be tested.
- 2- Prepare the input files for the Familias3 software v3.2.7 for each simulation test. As each simulation contained a certain number of markers, a total of 14 files were prepared, 7 of which contained the allele frequencies and 7 contained the profiles of the hypothetical members of the pedigree.
- 3- Setting up the mutation rate for each marker in the Familias3 software.
- 4- Confirm the parent-child relationship within the members of the hypothetical pedigree using the blind search in Familias3 software.
- 5- Carry out the simulation tests for the five relationships 1000 times using the seven sets of markers.
- 6- Define the LR limits for each simulation tests.
- 7- Study the impact of the number of markers and the number of relatives included in the simulation study on the LR estimates.
- 8- Using the cumulative genetic map distances (cM) of 41 aSTR published in (Phillips 2017) and the approximated cumulative genetic map distances (cM) for D16S539 and 94 iSNPs to calculate RFs for syntenic pairs as described by Phillips *et al.* (2012).
- 9- Highlight syntenic pairs that can potentially impact the LR estimation.

7.4 Materials and Methods

Materials and methods used in this part are described in Section 2.9.

7.5 Results and discussion

7.5.1 Confirmation of the parent-child relationship of the pedigree's members.

A hypothetical pedigree consisted of three generations with 13 members including all tested relationships was created using an in-house Excel software (Figure 7.1). To confirm that all members had appropriate genotypes for the 136 loci, the parent-child relationships between the pedigree's members (A&B with D&F; D&E with N, O, J and I; D&C with H; F&G with L) were tested using the blind search in Familias3 software. All parent-child relationships were confirmed demonstrating the correct genotypes generated by the in-house Excel tool (Figure 7.2).

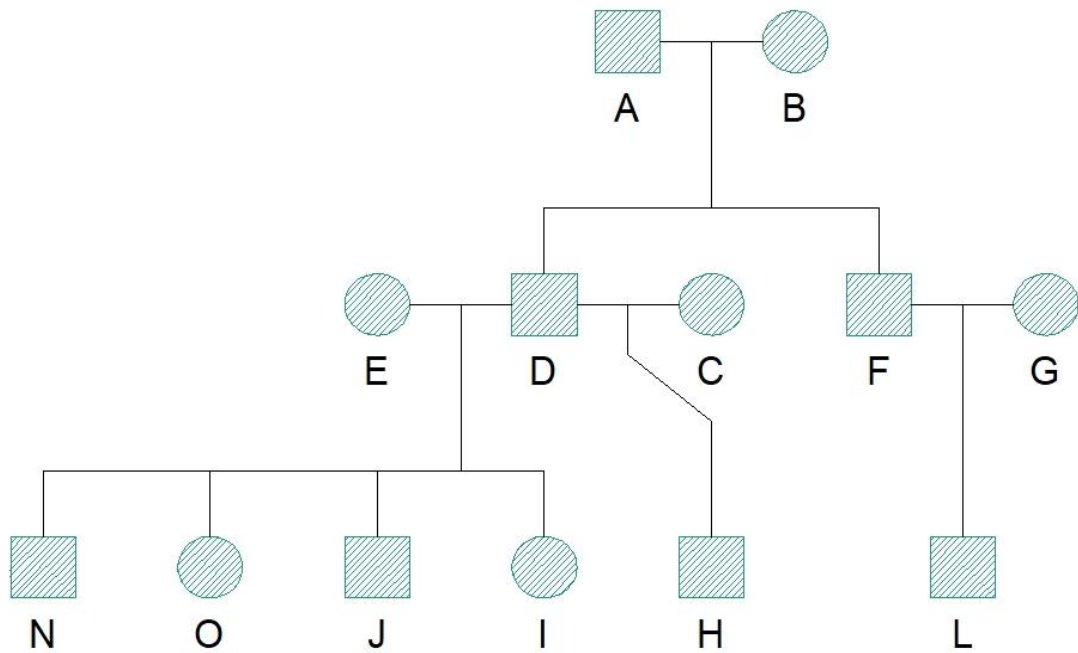


Figure 7.1. A hypothetical pedigree created by an in-house Excel sheet. The hypothetical pedigree comprised of three generations and 13 members. Circles represent female members and squares represent male members (This figure is a copy of Figure 2.2).

Blind search

This module performs a blind search on the imported data set. #Persons: 15, #Matches: 16

Person 1	Person 2	Relationship	LR	Inconsistencies	Overlapping markers	Cluster	Sharec
C	H	Parent-Child	1.5627314e+030	0	136	1	74.
B	F	Parent-Child	5.8639527e+025	0	136	2	72.
B	D	Parent-Child	6.8696892e+024	0	136	1	74.
D	O	Parent-Child	6.9462534e+023	0	136	1	73.
F	L	Parent-Child	1.0690008e+023	0	136	2	72.
A	F	Parent-Child	6.5556681e+022	0	136	2	71.
E	I	Parent-Child	9.2601018e+021	0	136	2	73.
D	I	Parent-Child	9.01668e+021	0	136	1	75.
E	N	Parent-Child	1.2063557e+021	0	136	2	72.
E	J	Parent-Child	1.1260743e+021	0	136	2	71.
G	L	Parent-Child	8.1825753e+020	0	136	2	68.
E	O	Parent-Child	4.1837129e+020	0	136	1	71.
D	J	Parent-Child	1.8203493e+020	0	136	1	73.
A	D	Parent-Child	5.3362158e+019	0	136	1	74.
D	N	Parent-Child	3.4481498e+019	0	136	1	74.
D	H	Parent-Child	3.1943873e+019	0	136	1	68.

Buttons: New search, View match, Merge samples, Remove, Remove all, Sort, Color clusters, Export matrix, Export list, Report match, Create summary, Close

Figure 7.2. A confirmation of the parent-child relationship assumed between the pedigree's members. The figure shows a screen shot from the Familias3 software for the results of the blind search (parent-child). Each parent-child relationship was validated for the 136 loci before starting the simulation tests.

7.5.2 Simulation results

To evaluate the differentiation power between related and unrelated individuals that can be achieved when using more DNA markers, five types of relationships: parent-child/unrelated (mother not available (duo pedigree)), full-siblings/unrelated (3 scenarios), half-siblings/unrelated, first-cousins/unrelated (2 scenarios) and grand-parent or grand-child/unrelated, were simulated using allele frequency data generated from Chapters 3, 5, and 6 and the hypothetical pedigree. In addition, seven different combinations of DNA markers included in different commercially available kits: Identifiler Plus (15 aSTRs), GlobalFiler (21 aSTRs), GlobalFiler and SureID23 (38 aSTRs), Fusion 6C and SureID23 (40 aSTRs), ForenSeq DNA Signature Prep (27 aSTRs and 94 iiSNPs (121 loci)), all loci (42 aSTRs and 94 iiSNPs (136 loci)) and 94 iiSNPs, were used. For the rest of this study, the number of the markers are used rather than the name of the kits.

For each relationship tested, Familias3 software (Kling *et al.* 2014) calculated the Likelihood ratio (LR) for the two hypotheses, which are shown in (Table 2.5), by dividing the probability of hypothesis 1 by the probability of hypothesis 2. The LR will have a value of >1 (\log_{10} of 1 = 0) if the probability of hypothesis 1 higher than the probability of hypothesis 2 (related as claimed in hypothesis 1) and will have a value of <1 if the probability of hypothesis 2 higher than the probability of hypothesis 1 (unrelated). In all relationships tested in this study, hypothesis 1 is the correct relationship between tested members.

Each relationship was simulated 1000 times for each marker set and six LR limits (thresholds) (from 1, 10 ... to 100,000) were applied to measure the true positive (TP) and the false positive (FP) of each scenario at each threshold. Here, the TP, for example

when using LR threshold of 1, represents the percentage of related simulations that appeared as related (they are related and $LR > 1$), while the FP represents the percentage of unrelated simulations that appeared as related (they are unrelated, but $LR > 1$). False negative (FN) represents the percentage of related simulations that appeared as unrelated (they are related, but $LR < 1$), which can be calculated by $100\% - TP\%$ (Figure 7.3). The software generates a data file for each simulation run that can be visualised by plotting in the RStudio platform (RStudio Team 2016).

Typically, when more loci are added, the LRs for related individuals are increased (more shifting to the right) and the LRs for unrelated individuals are decreased (more shifting to the left) in comparison to fewer markers. This shifting reduces the overlapping area (uncertainty area) between the LRs of related and unrelated and thus reduces the chance of false inclusion (FP) or exclusion (FN) (Figure 7.3).

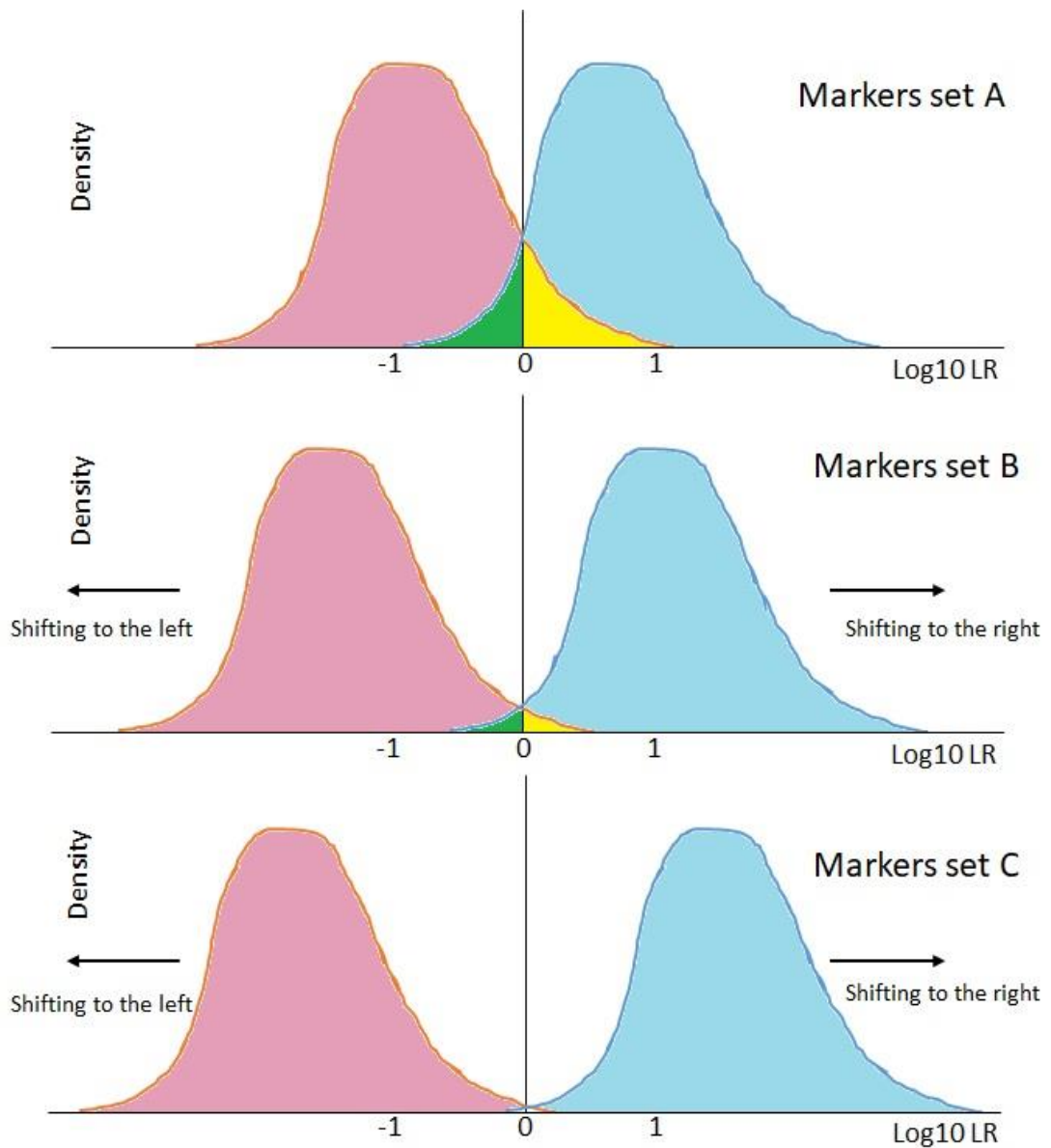


Figure 7.3. The impact of adding more DNA markers to the simulation tests on the LR. The figure shows how testing more DNA markers improves the LR and thus reduces uncertainty. The blue line represents LR distribution for related simulations, the red line represents LR distribution for unrelated simulations, the light blue area represents the true positive (TP), the light red area represents the true negative (TN), the yellow area represents the false positive (FP), and the green area represents false negative (FN). The marker sets A, B and C are examples of different marker sets where the number of markers in set A lower than in set B, which is lower than in set C. The green and yellow areas are the uncertainty areas. When more markers are used (e.g. set B) LR distribution of related moves to the right (LR increased) and LR distribution of unrelated moves left (LR decreased). The uncertainty areas are decreased when more markers are added (e.g. set C) (\log_{10} of LR 1 = 0) (an original figure).

Parents (mother and father)-child/unrelated relationship (trio pedigree) was not included in the simulation study as 15 loci were found to be enough to differentiate between parents-child and unrelated with 100% TP and 0% FP up to a LR threshold of 100,000 (Figure 7.4). In addition, previous work demonstrated that even the 13 CODIS STRs were able to differentiate between parent-child and unrelated with 0% FP and FN at a LR threshold of 1 (O'Connor *et al.* 2010).

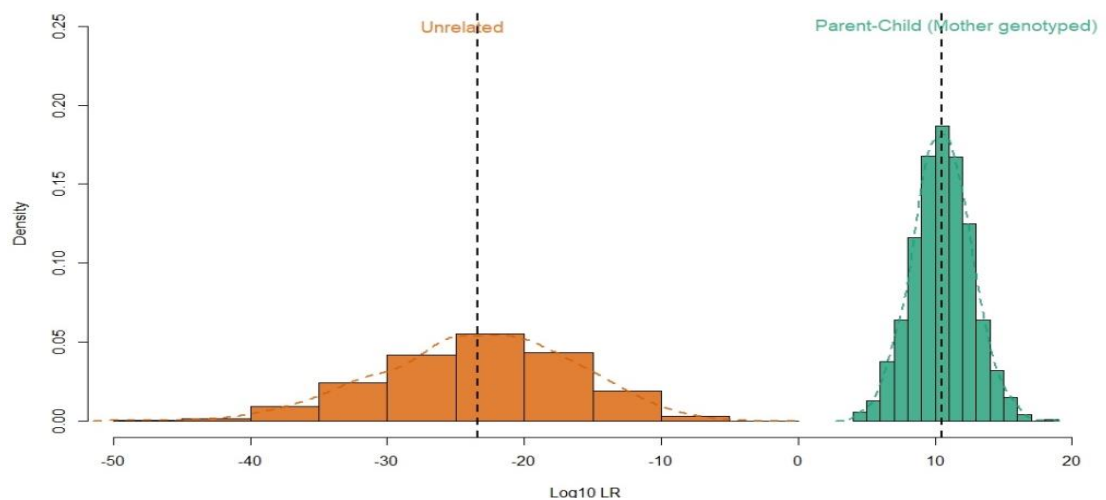


Figure 7.4. LR distributions of the simulation study for parents-child relationship (trio pedigree) using 15 aSTRs included in the Identifiler kit, which was plotted based on data generated by Familias3 software. The green histogram represents LR distributions for the true positive simulations (parents-child relationship), the orange histogram represents the LR distributions of true negative simulations (unrelated). The 15 aSTRs showed 100% TP and 0% FP up to the 100,000 LR threshold.

However, more loci may be needed when a meiotic mutation is observed (Jia *et al.* 2015), a mother less pedigree (duo pedigree) (Poetsch *et al.* 2013, González-Andrade *et al.* 2009), or when the alleged fathers/mothers are close relatives (Dogan *et al.* 2015, Canturk *et al.* 2016).

For the rest of relationships, the assessments started with studying the impact of additional markers on the LR. The simulation results showed that the LR, as expected, improved (increased in related simulations and decreased in unrelated simulations) when more loci were used, and the improvements were correlated to the number of

loci added. However, the level of improvement varied and was impacted by the type of relationship tested and by the number of relatives included in the simulation (Figure 7.5 and Figure 7.6) (Table 7.1). For example, the LR medians for parent-child relationship (duo pedigree) ranged from 24564.25 (15 aSTRs) to 1.05665E+20 (136 loci) while they ranged from 1.086125 (15 aSTRs) to 2.31342 (136 loci) for first-cousins/unrelated (Scenario 1). When more relatives were included in the simulation, the LRs medians of full-siblings, for example, were significantly improved from 7.373645 (Scenario 1, two siblings were tested) to 344.3885 (Scenario 2, three siblings were tested) and to 43126.8 (Scenario 3, four siblings were tested) using the 15 aSTRs (Figure 7.5 and Figure 7.6) (Table 7.1). Table 7.1 summarises the improvements in LR medians of related/unrelated when more loci are added for each relationship.

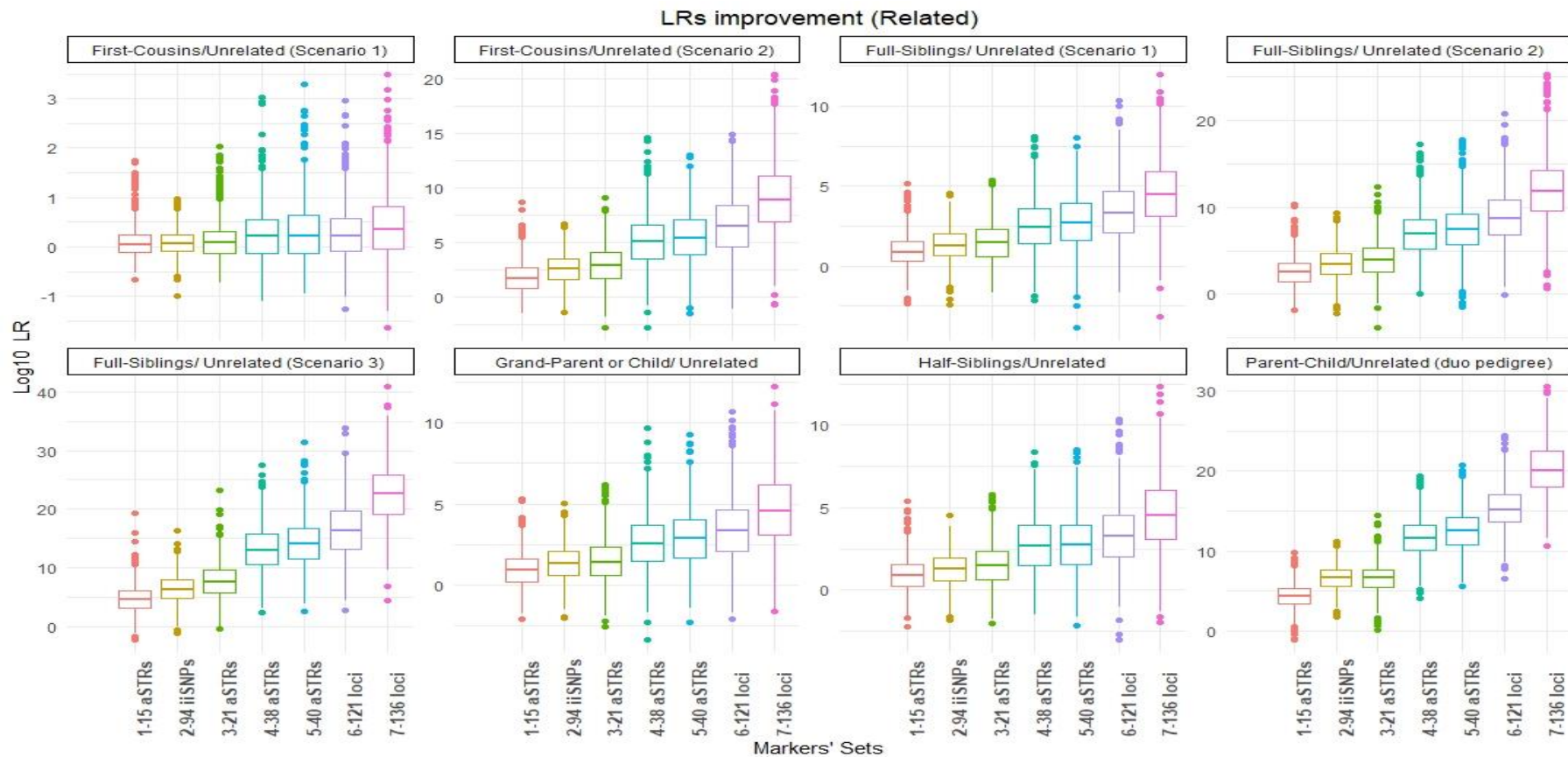


Figure 7.5. LR improvements (increment) for different relationships using the seven marker sets for related simulations. The figure shows LR improvements when more loci used and shows the impact of type of the relationship simulated and impact of including relatives in the simulation tests, on the LRs. In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar).

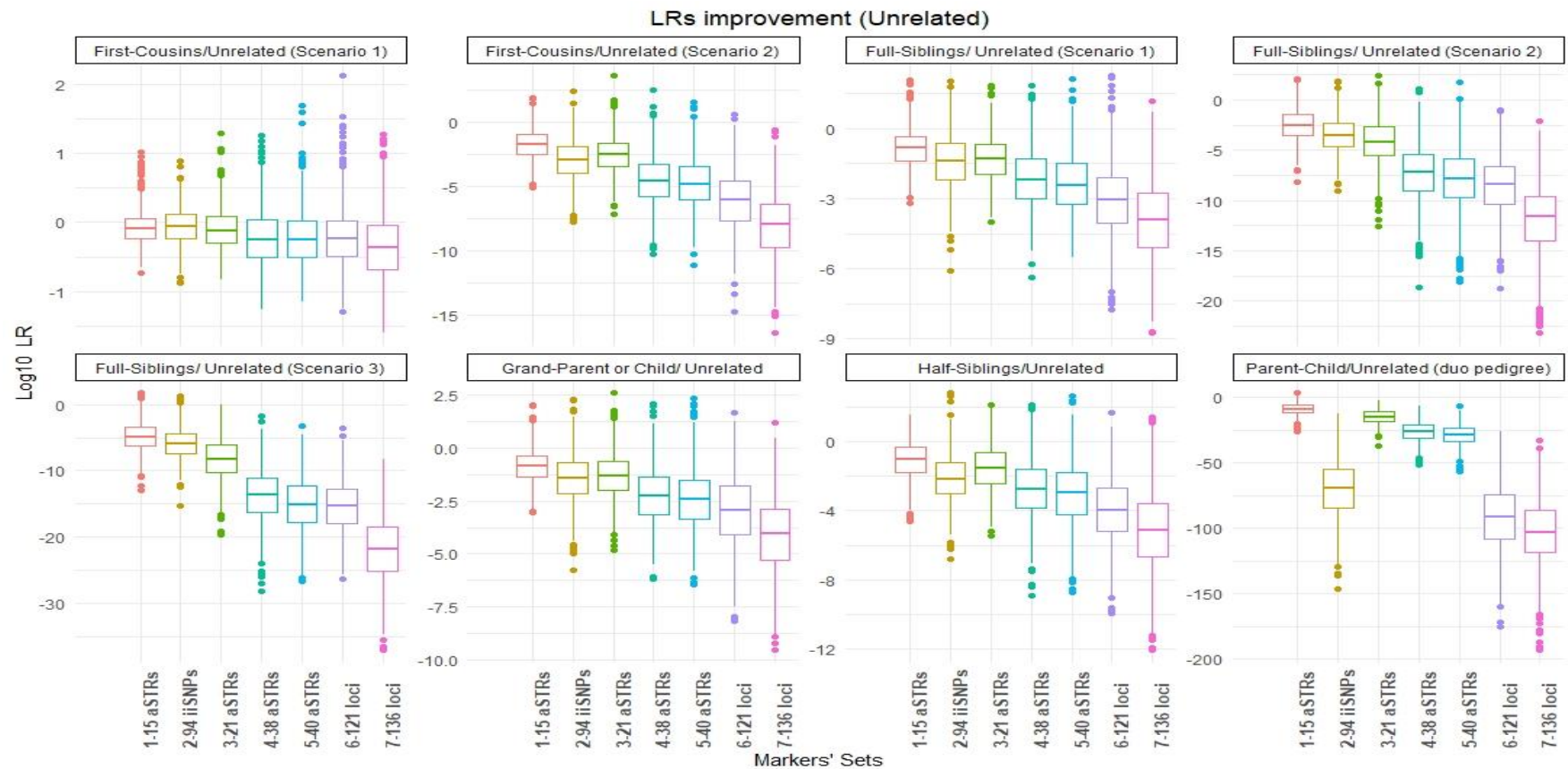


Figure 7.6. LR improvements (decrease) for different relationships using the seven marker sets for unrelated simulations. The figure shows LR improvements when more loci were used and shows the impact of type of the relationship simulated and impact of including relatives in the simulation tests, on the LRs. Higher impact of the 94 iiSNPs on parent-child relationship (when used alone or when they were included in the 121 or the 136 loci) can be seen in the bottom right (will be discussed at the end of this study). In the box plots, the lower whisker represents 25% of the lowest data, the upper whisker represents 25% of the highest data. The rectangle shows that 75% of the data are below the upper line, 25% of the data are below the lower line, and the centre bar represents the median of the data (50% of the data above this bar and 50% of the data below the bar)..

Table 7.1. LR medians for eight scenarios simulated using seven different markers sets for related and unrelated simulations. The table shows the improvement on LRs when more markers were used for the tested relationships. It also shows the case pedigrees (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members (i.e. genotyped members) and crossed member were assumed as not available for testing. As expected, LRs improved when more loci were added. The improvement varied and was impacted by the type relationship tested and by the number of relatives included in the simulation.

Relationship	Simulated pedigree (hypothesis 1)	Markers set	LR Medians (Related)	LR Medians (Unrelated)
Parent-Child/Unrelated (duo pedigree)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	2.46E+04 4.70E+06 3.86E+11 3.39E+12 1.54E+15 1.06E+20 4.36E+06	1.23E-09 1.62E-15 1.12E-26 1.27E-29 7.31E-92 5.94E-104 4.37E-70
Full-Siblings/ Unrelated (Scenario 1)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	7.37 3.01E+01 2.91E+02 5.30E+02 2.02E+03 3.15E+04 2.02E+01	1.49E-01 5.29E-02 6.65E-03 3.85E-03 8.69E-04 1.27E-04 4.14E-02
Full-Siblings/ Unrelated (Scenario 2)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	3.44E+02 7.96E+03 9.68E+06 2.67E+07 5.44E+08 7.73E+11 3.12E+03	3.11E-03 6.85E-05 5.96E-08 1.30E-08 4.09E-09 2.55E-12 3.08E-04
Full-Siblings/ Unrelated (Scenario 3)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	4.31E+04 4.03E+07 8.66E+12 1.00E+14 2.32E+16 4.02E+22 2.23E+06	1.38E-05 6.02E-09 2.50E-14 7.90E-16 4.23E-16 1.64E-22 1.18E-06
First-Cousins/Unrelated (Scenario 1)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	1.09 1.22 1.63 1.67 1.67 2.31 1.17	8.16E-01 7.52E-01 5.64E-01 5.67E-01 5.87E-01 4.27E-01 8.64E-01
First-Cousins/Unrelated (Scenario 2)		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	5.72E+01 7.78E+02 1.31E+05 2.37E+05 3.58E+06 7.47E+08 3.93E+02	2.14E-02 3.04E-03 2.66E-05 1.58E-05 9.73E-07 1.22E-08 1.33E-03
Half-Siblings/Unrelated		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	7.02 2.82E+01 4.31E+02 5.55E+02 1.73E+03 3.22E+04 1.78E+01	1.42E-01 4.70E-02 5.12E-03 4.42E-03 9.67E-04 1.14E-04 3.77E-02
Grand-Parent or Child/ Unrelated		15 aSTRs 21 aSTRs 38 aSTRs 40 aSTRs 121 loci 136 loci 94 iiSNPs	8.06 2.57E+01 3.56E+02 7.72E+02 2.36E+03 3.83E+04 2.18E+01	1.36E-01 4.88E-02 5.29E-03 4.04E-03 1.15E-03 9.67E-05 3.61E-02

The data files generated by Familias3 software for each relationship using each marker set were used to generate two plots: LR distribution and exceedance probability (a figure that shows the improvement in probabilities at different LR thresholds). To compare between the marker sets, the LR distribution plots for each relationship were integrated in one plot and are presented (Figures 8.7, 8.9, 8.11, 8.13, 8.15, 8.17, 8.19 and 8.21). In addition, another type of figures that shows the percentages of TP and FP estimated for each relationship using each marker set at different LR thresholds (Figures 8.8, 8.10, 8.12, 8.14, 8.16, 8.18, 8.20 and 8.22) are also presented. The combined exceedance probability figures (8 Figures) are presented in Section 10.6.1 (Appendix 6). All figures were plotted using RStudio platform (RStudio Team 2016).

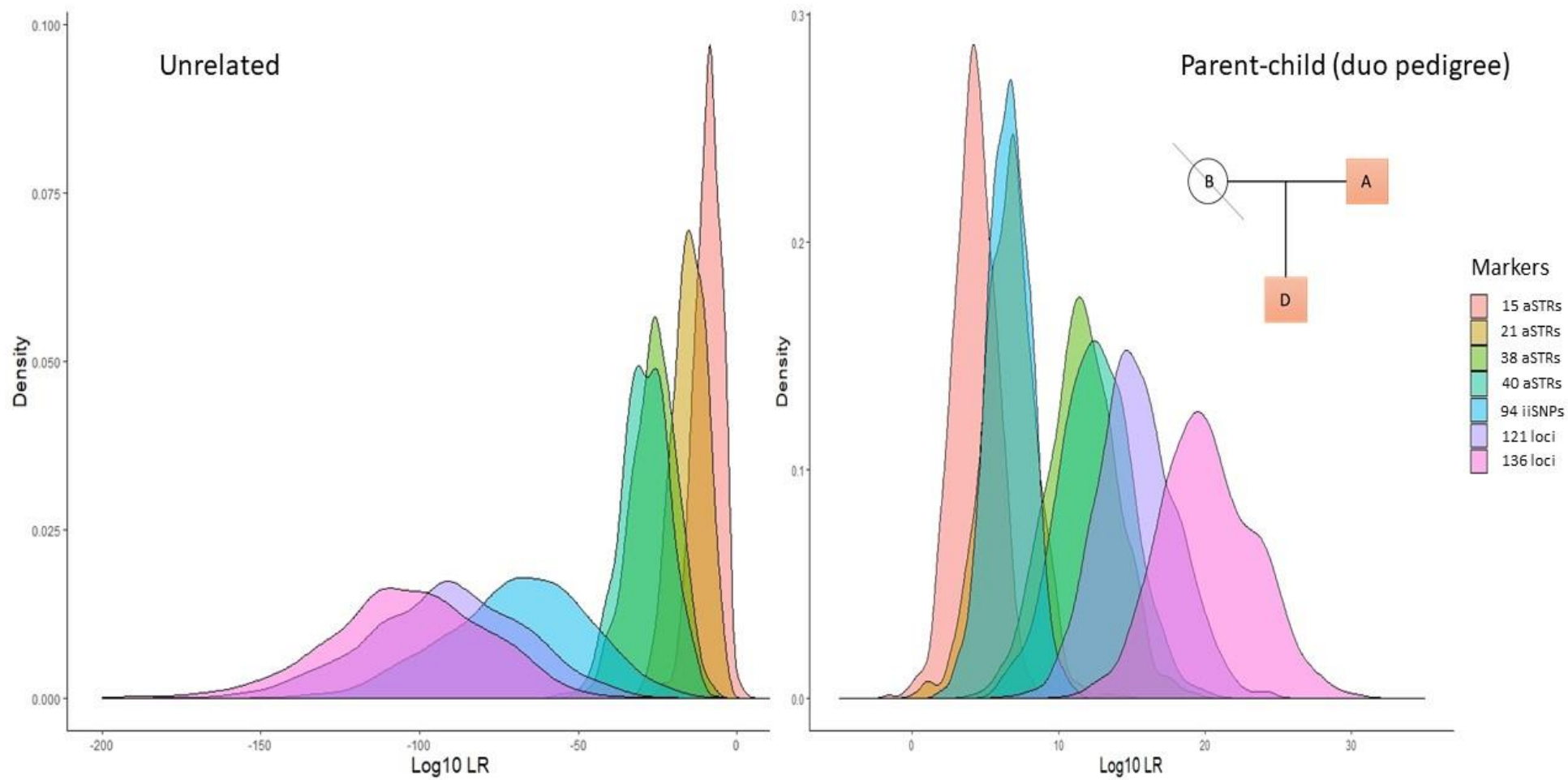


Figure 7.7. LR distributions of the simulation study for parent-child relationship (duo pedigree) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed member was not available for testing.

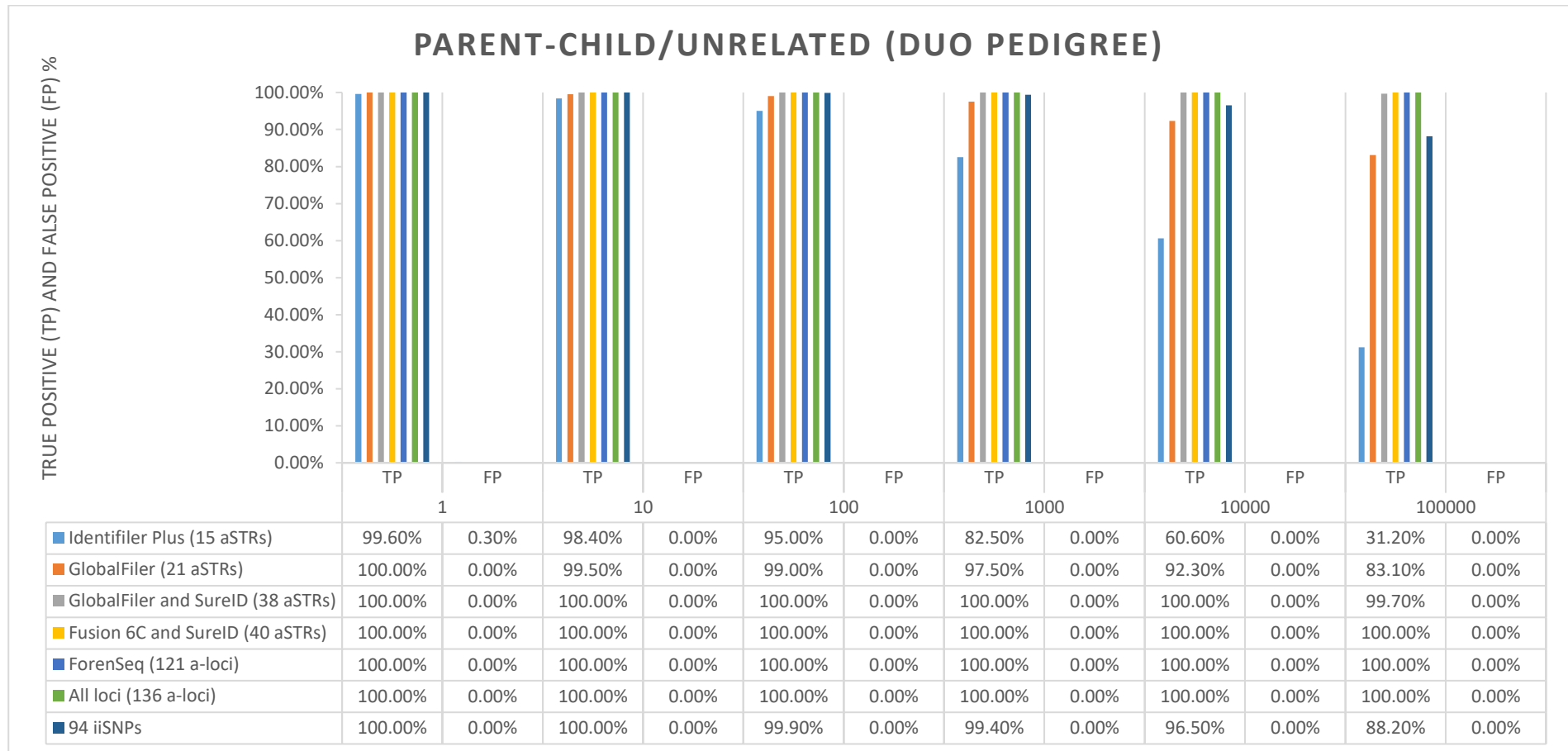


Figure 7.8. The TP and FP at different LR thresholds generated from the simulation study for parent-child relationship (duo pedigree) using different marker combinations. Each marker set is represented by a unique colour. True positive (TP) and false positive (FP).

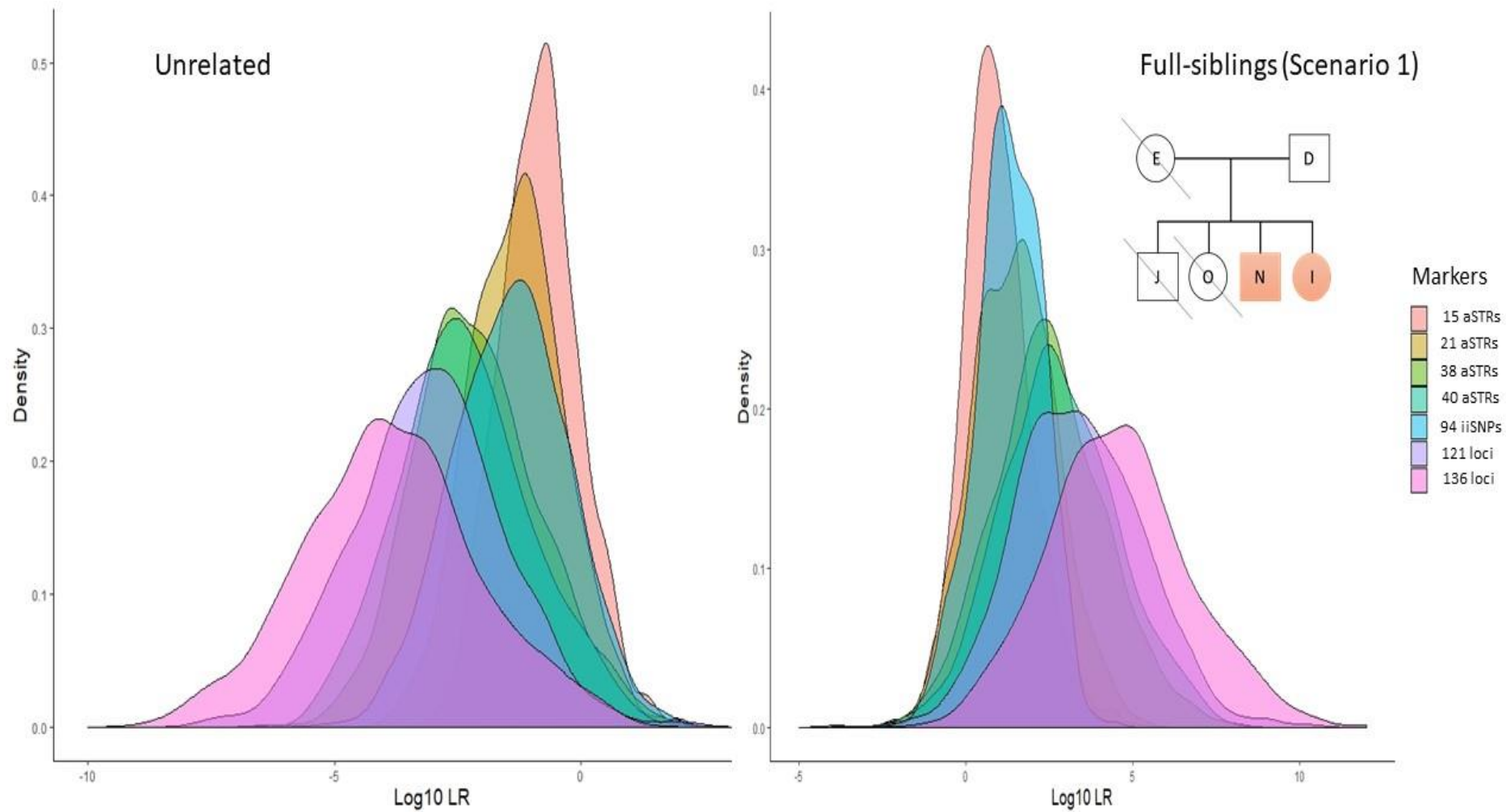


Figure 7.9. LR distributions of the simulation study for full-siblings/unrelated (Scenario 1) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.

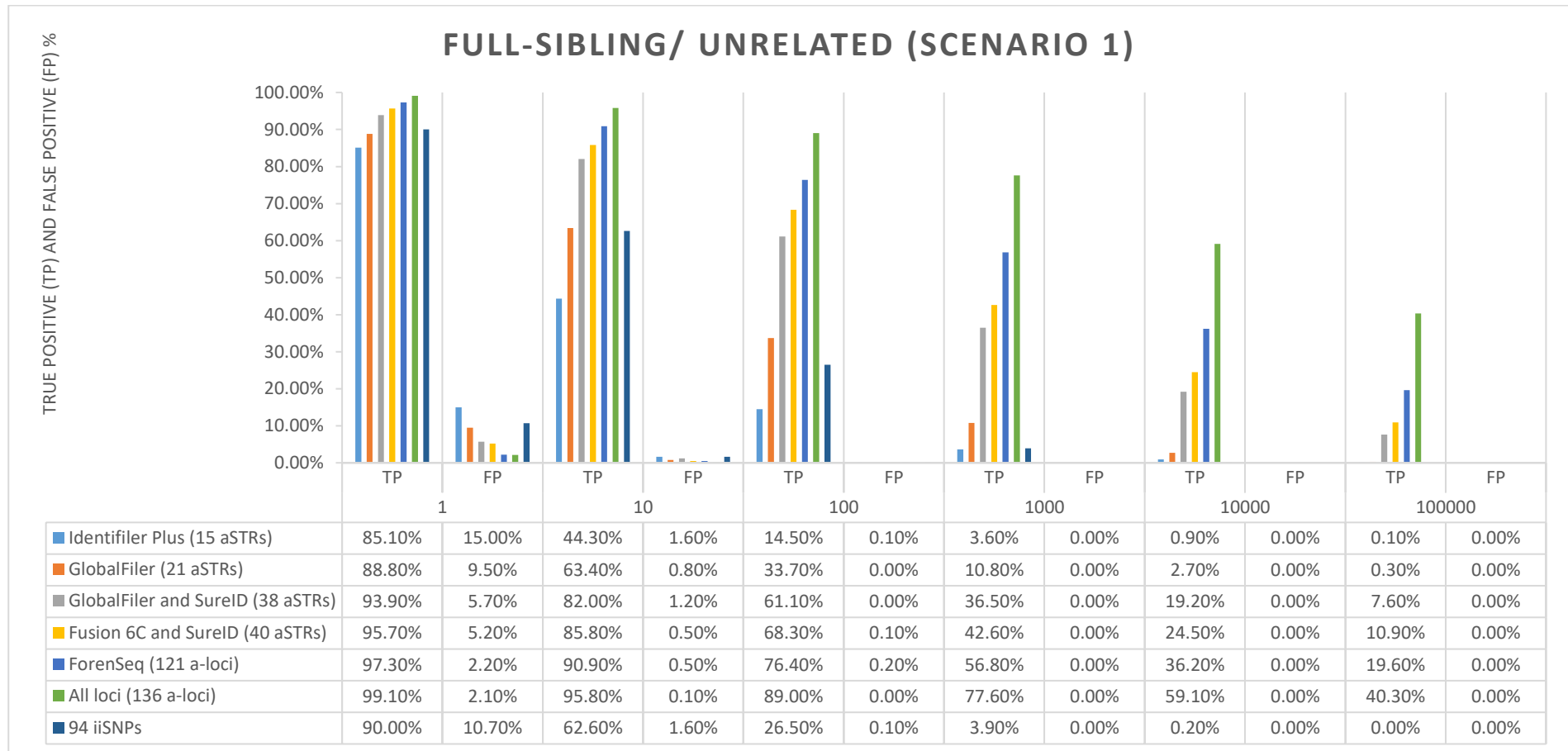


Figure 7.10. The TP and FP at different LR limits generated from the simulation study for full-siblings/unrelated (Scenario 1) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

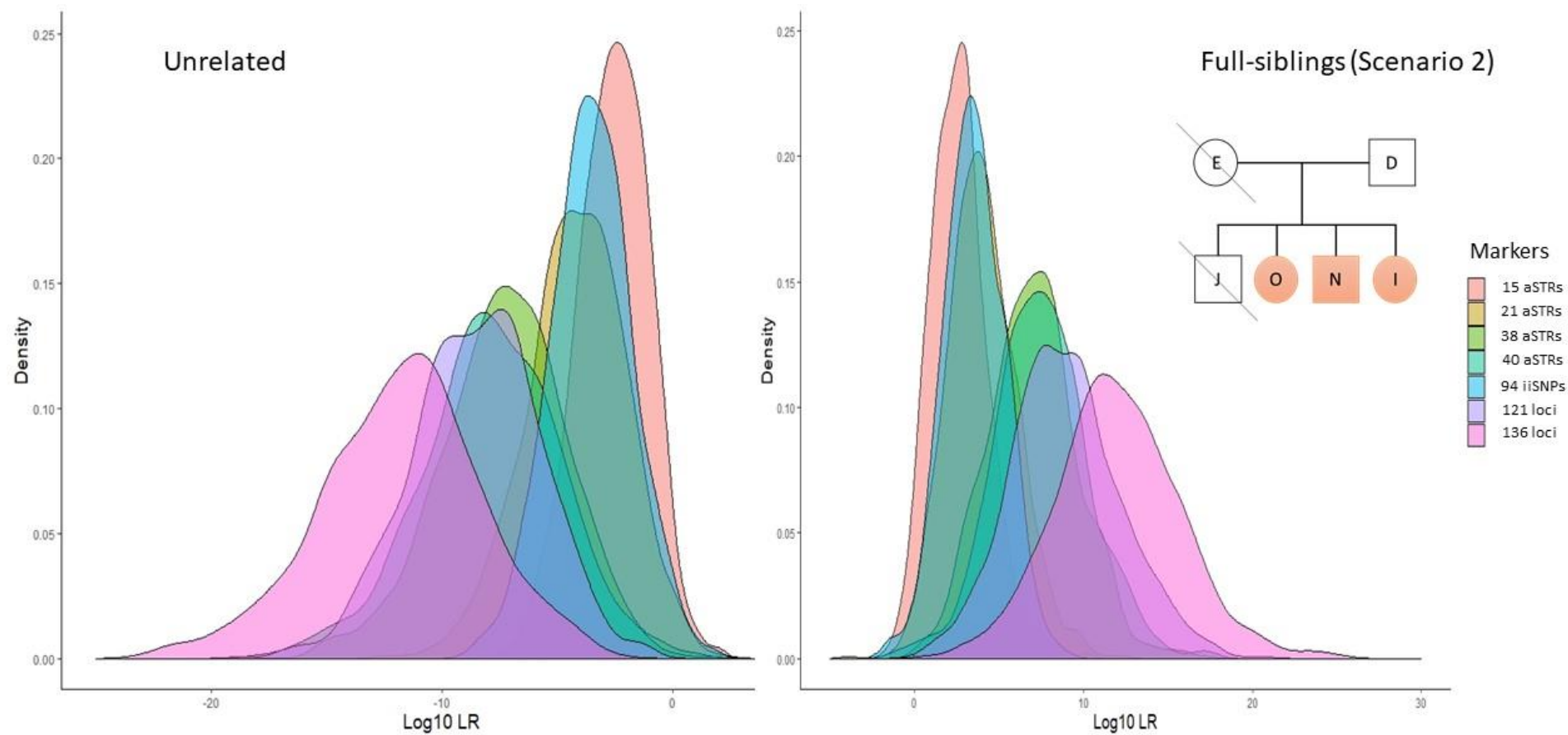


Figure 7.11. LR distributions of the simulation study for full-siblings/unrelated (Scenario 2) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.

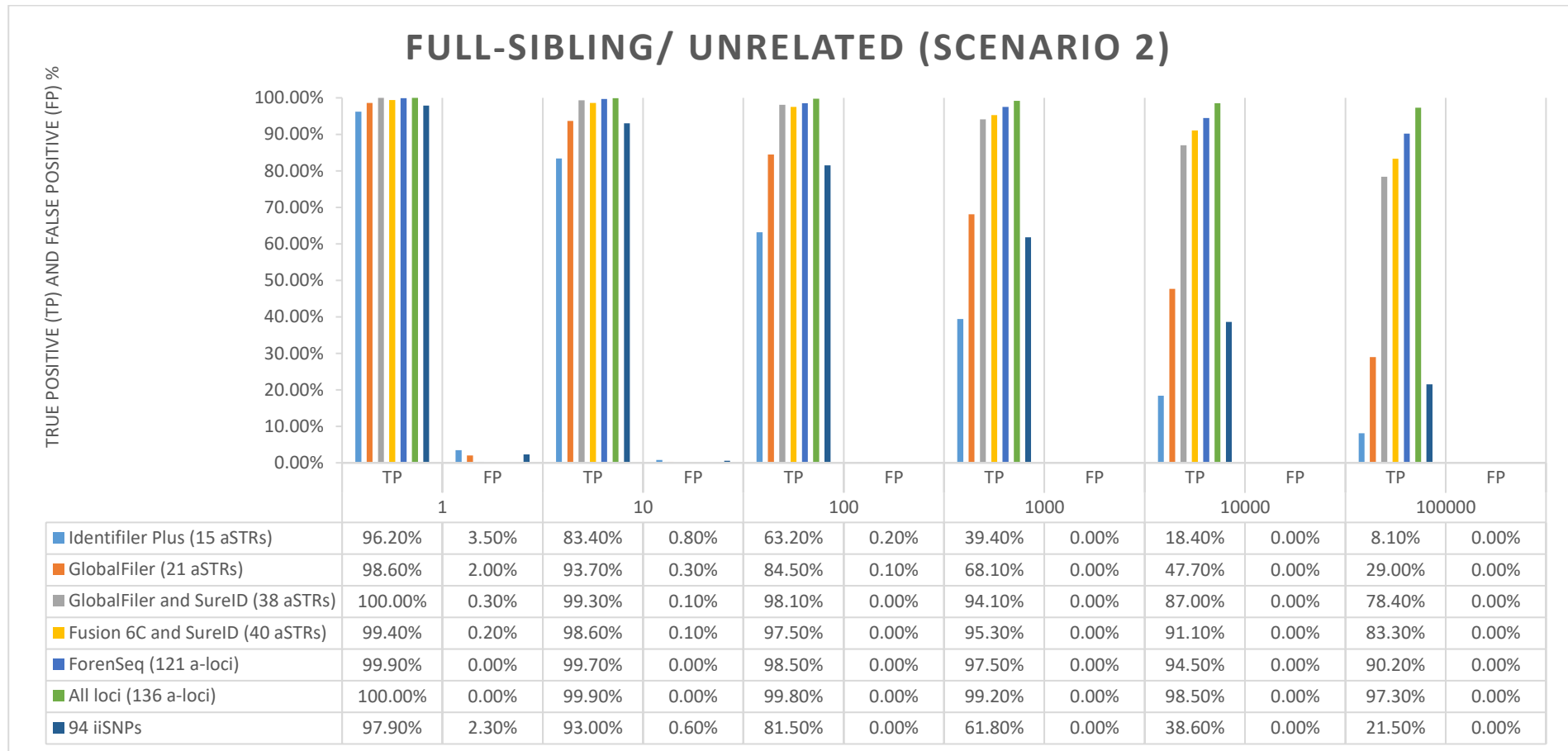


Figure 7.12. The TP and FP at different LR limits generated from the simulation study for full-siblings/unrelated (Scenario 2) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

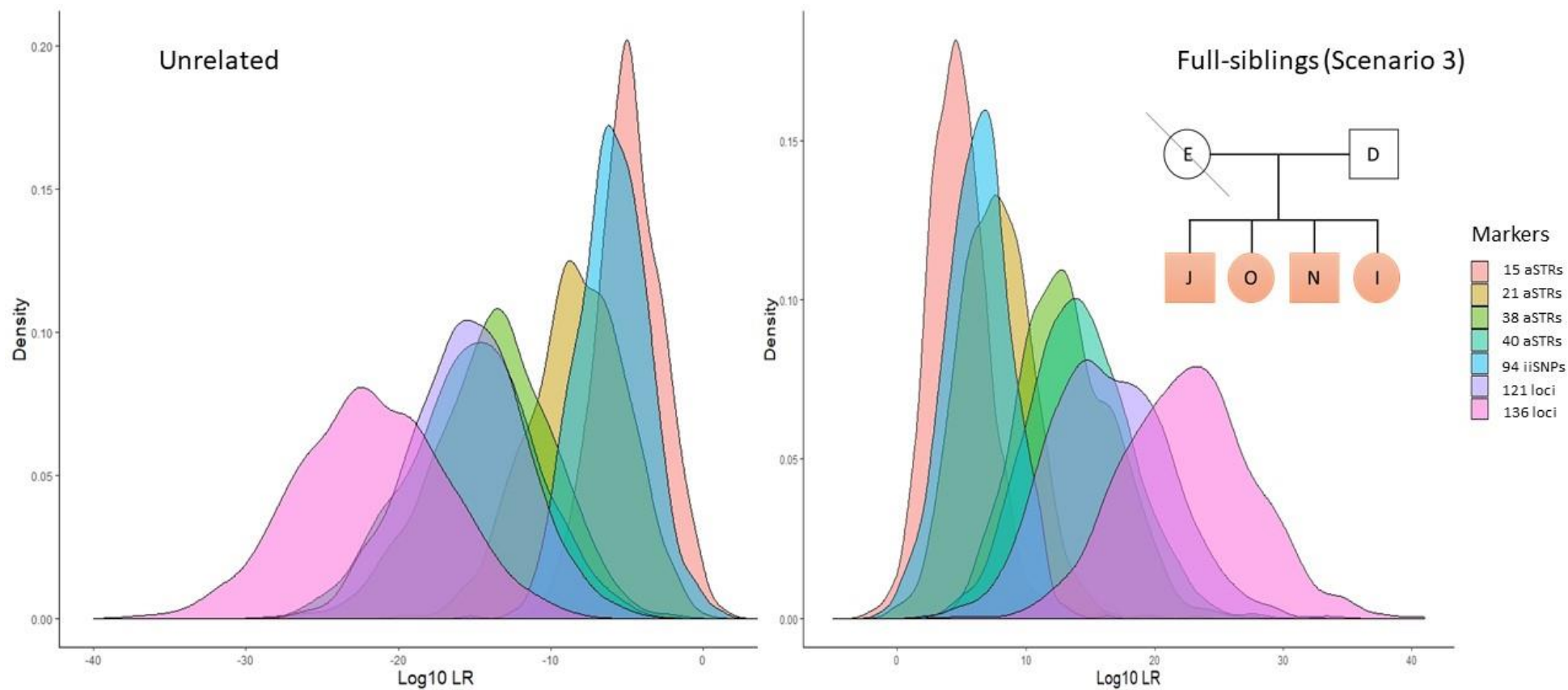


Figure 7.13. LR distributions of the simulation study for full-siblings/unrelated (Scenario 3) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed member was not available for testing.

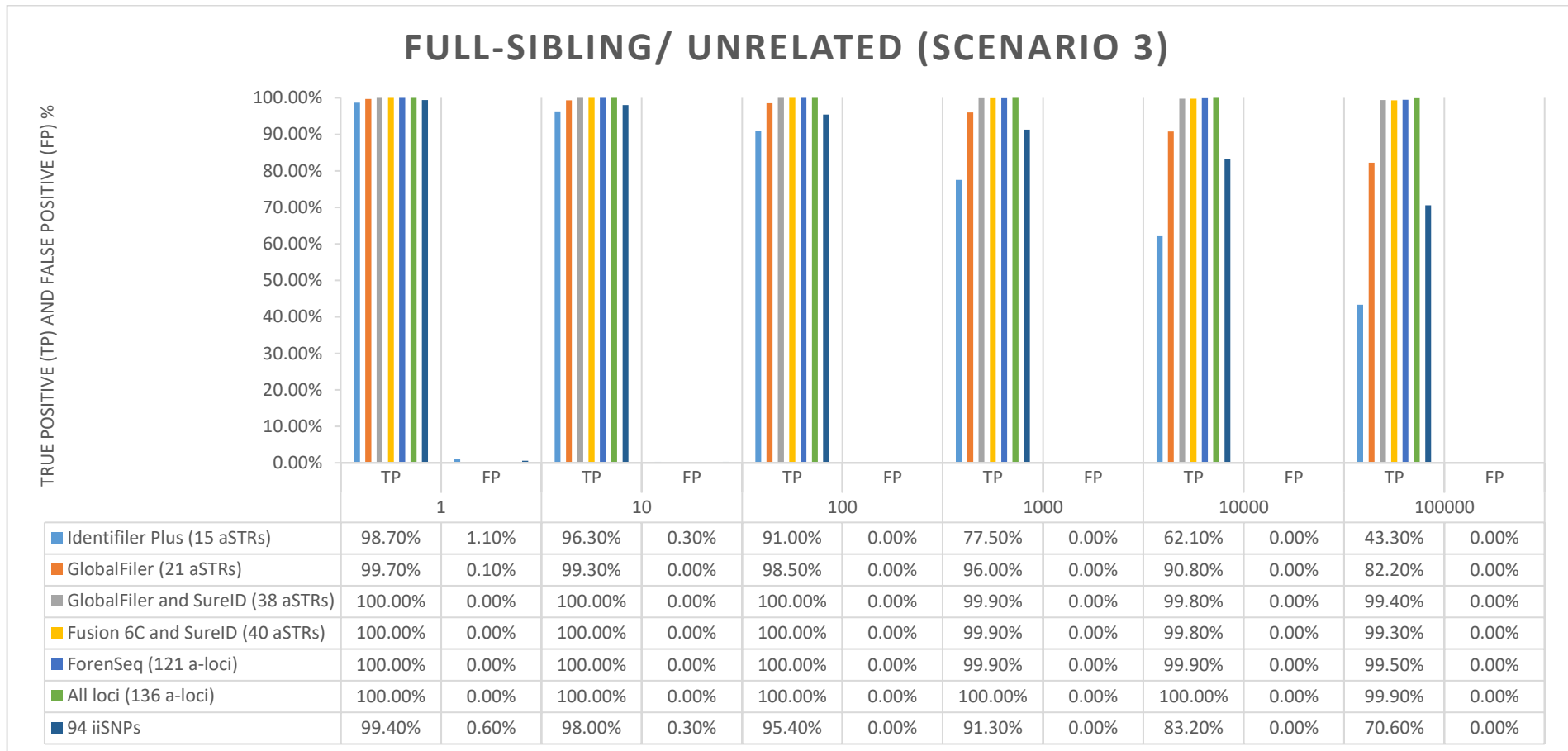


Figure 7.14. The TP and FP at different LR limits generated from the simulation study for full-siblings/unrelated (Scenario 3) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

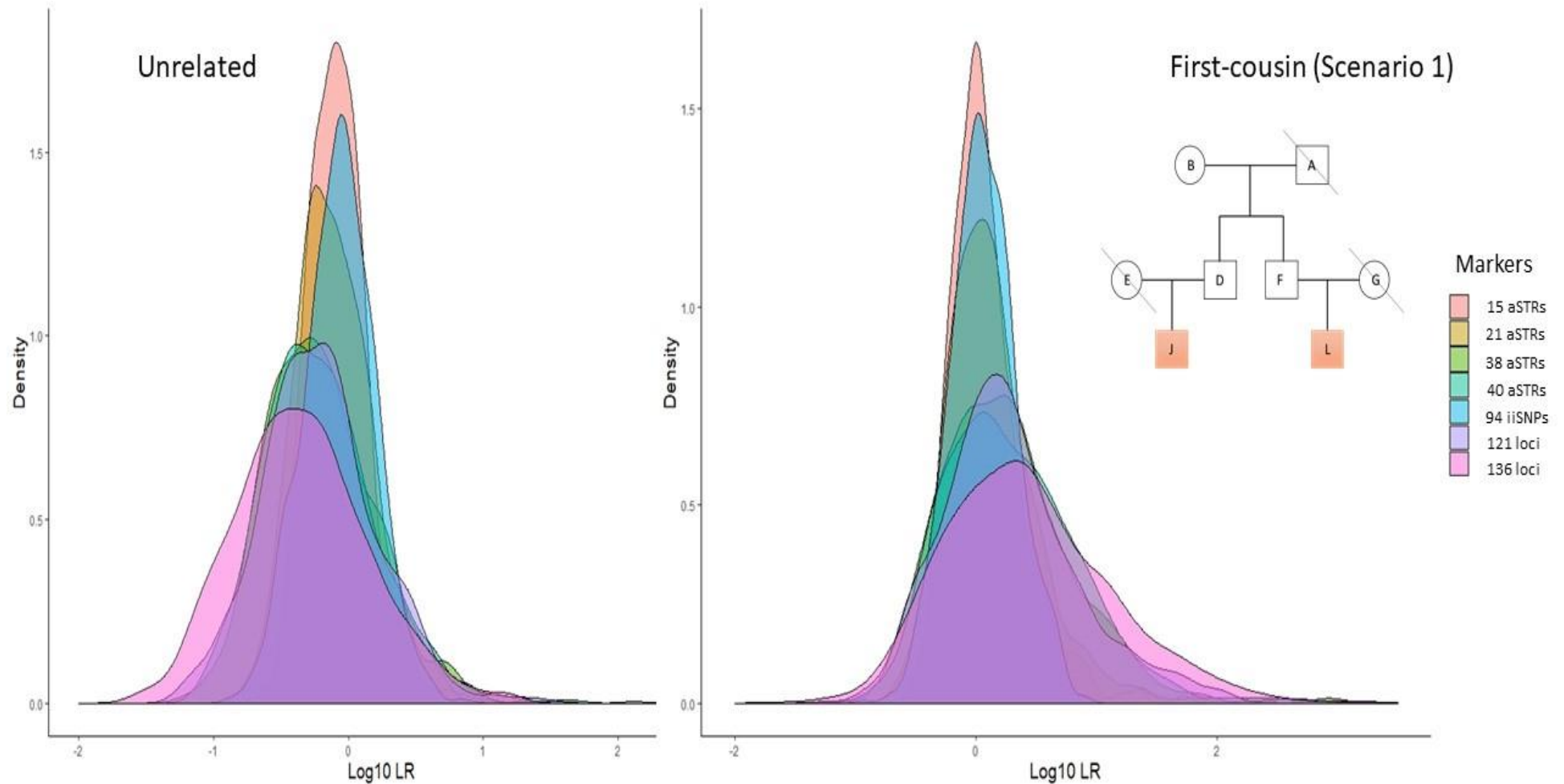


Figure 7.15. LR distributions of the simulation study for first-cousin/unrelated (Scenario 1) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.

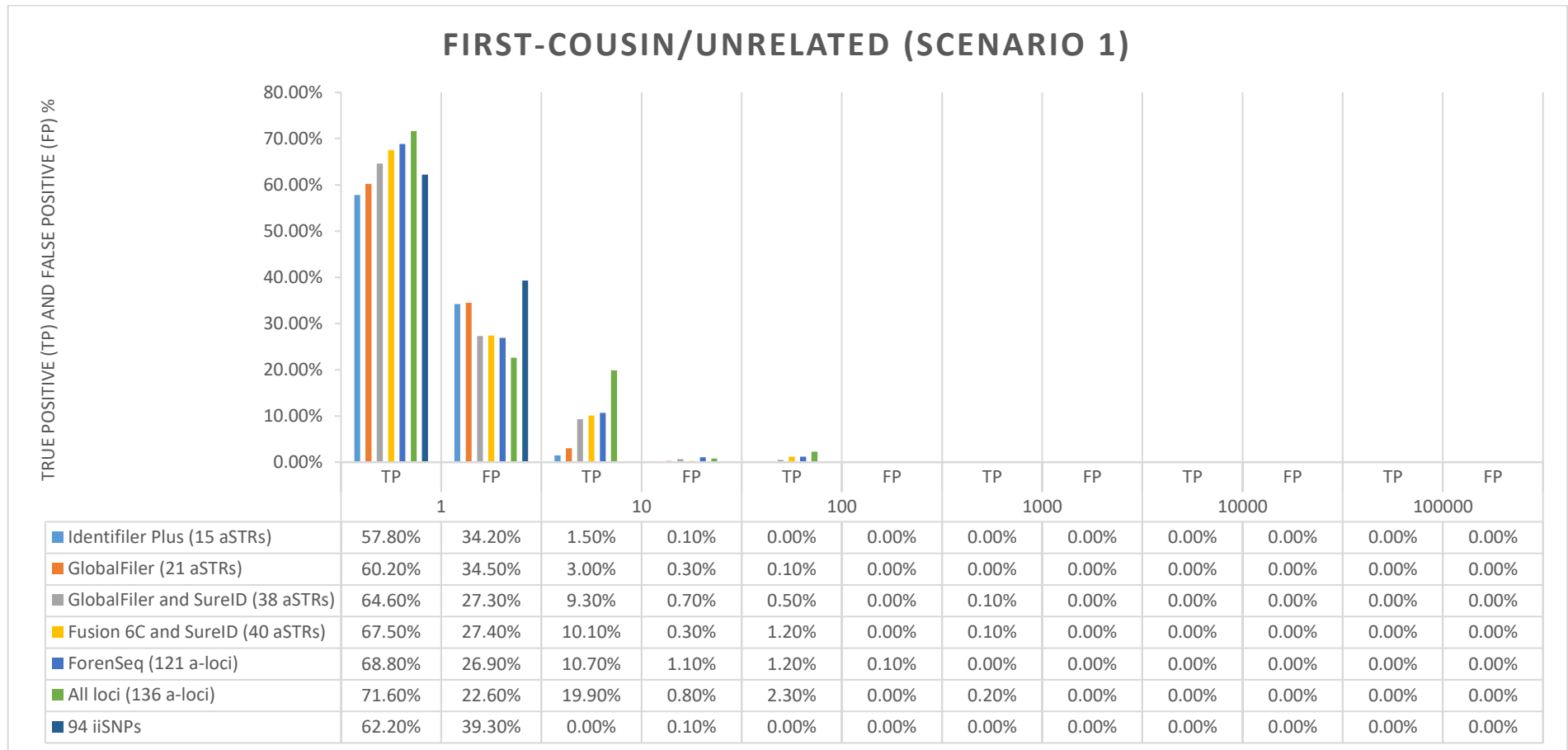


Figure 7.16. The TP and FP at different LR limits generated from the simulation study for first-cousin/unrelated (Scenario 1) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

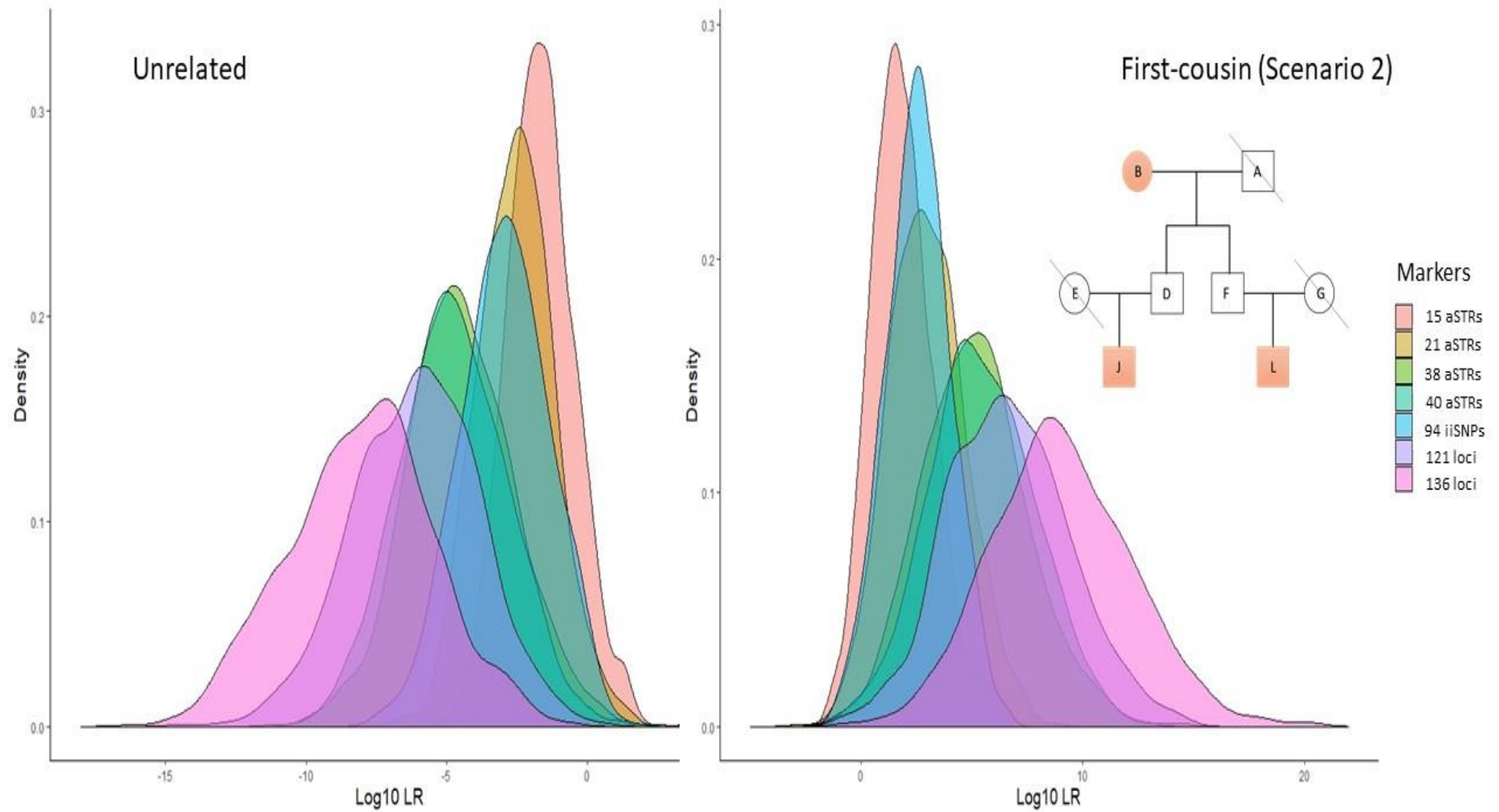


Figure 7.17. LR distributions of the simulation study for first-cousin/unrelated (Scenario 2) using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing.

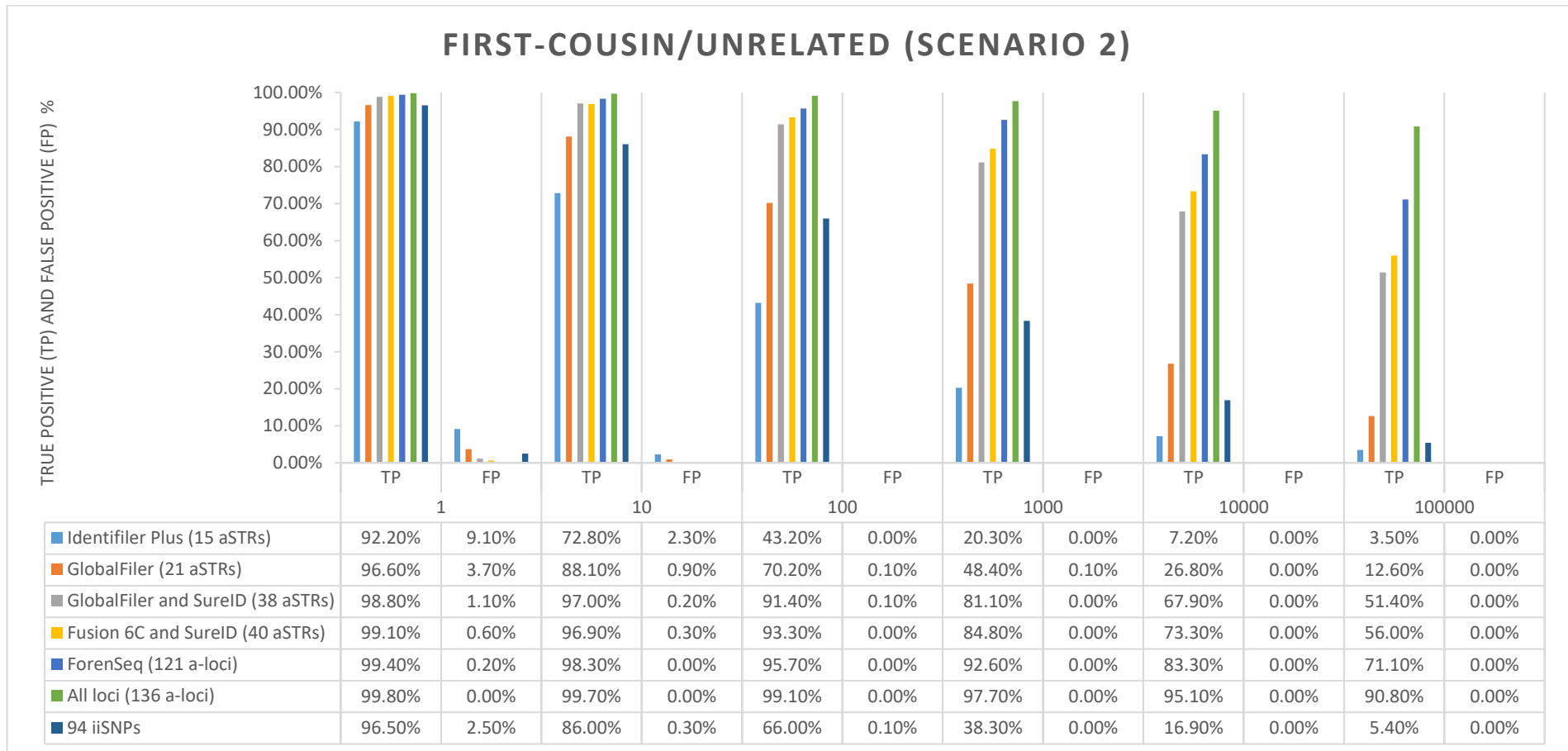


Figure 7.18. The TP and FP at different LR limits generated from the simulation study for first-cousin/unrelated (Scenario 2) using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

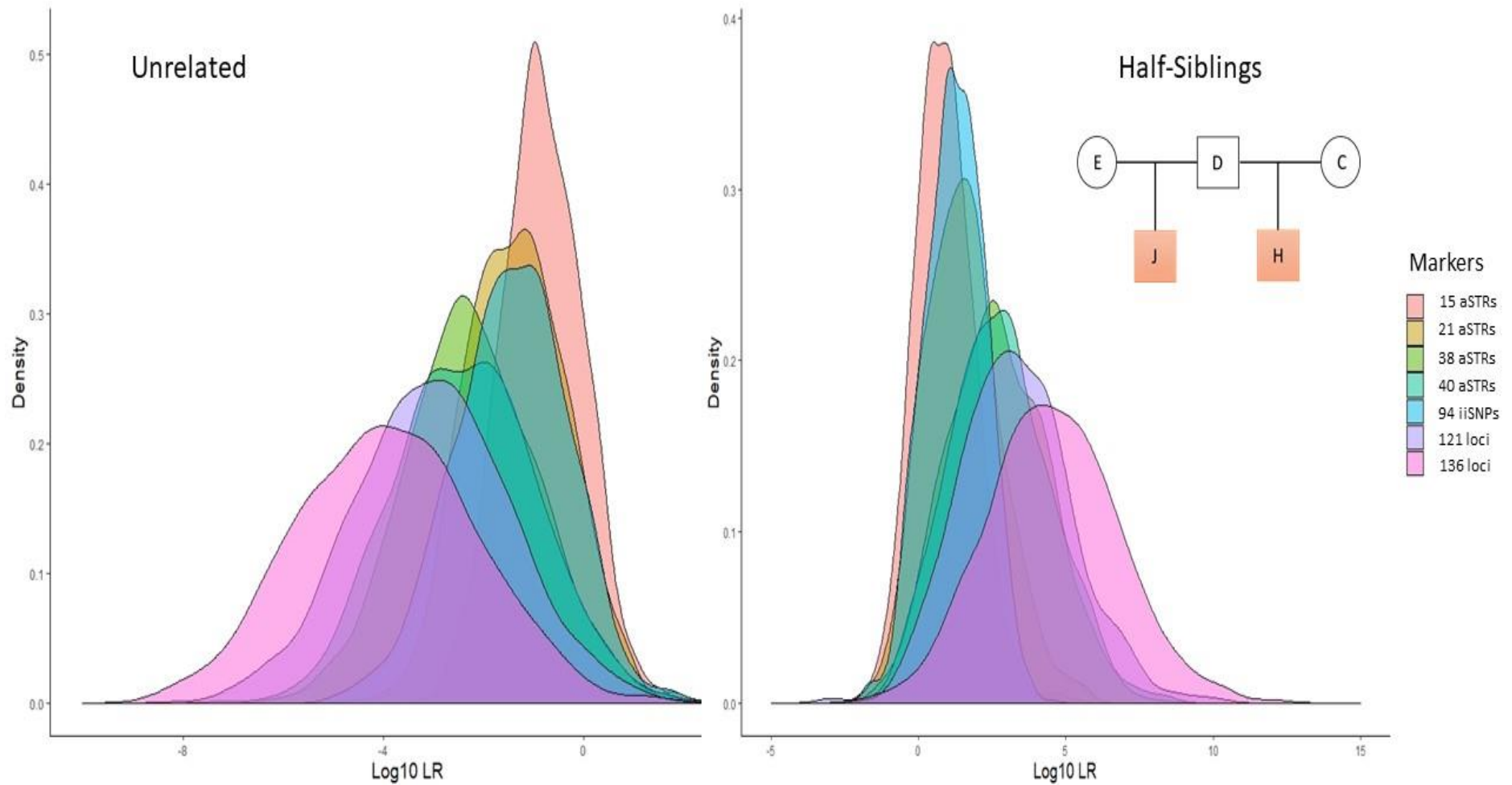


Figure 7.19. LR distributions of the simulation study for half-siblings/unrelated using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members.

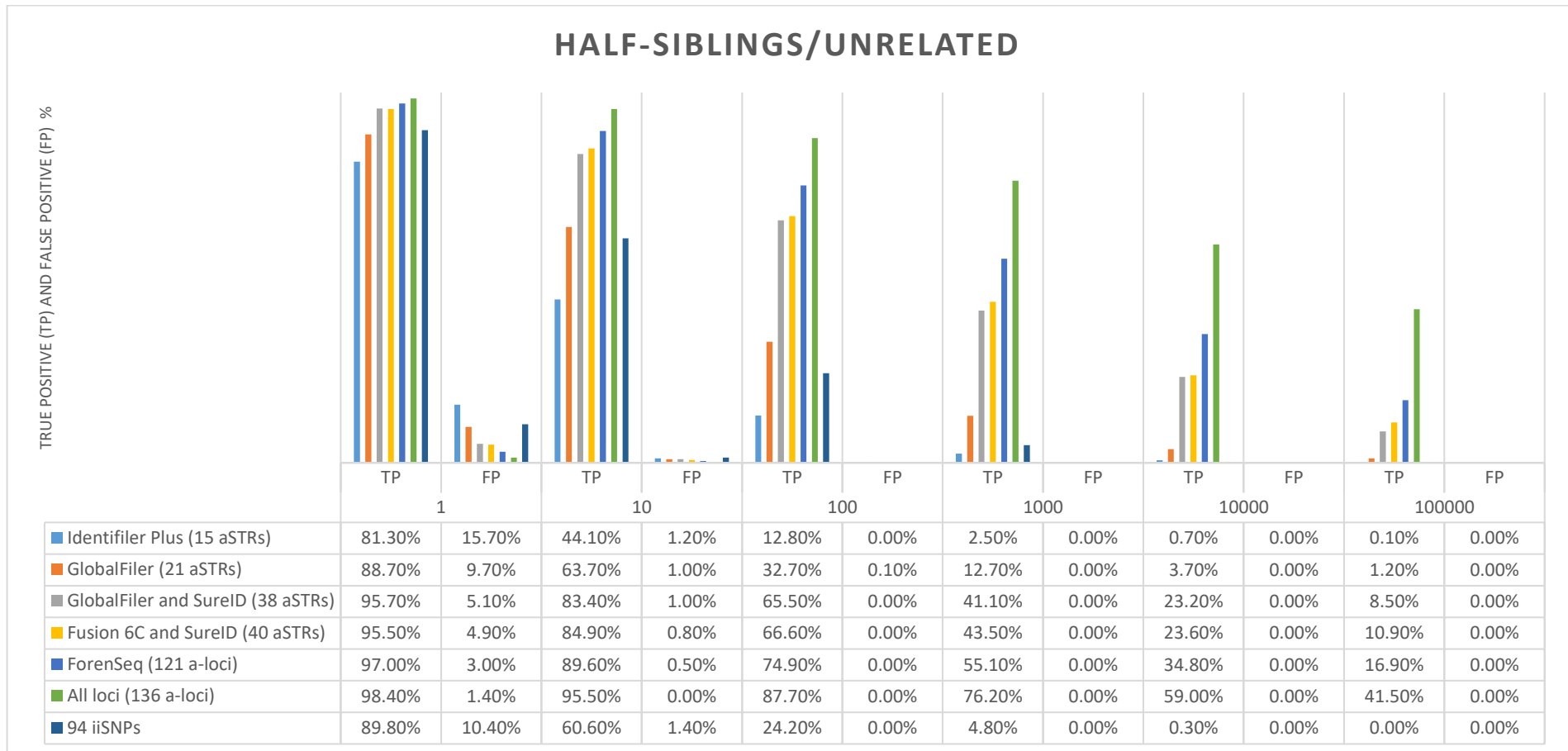


Figure 7.20. The TP and FP at different LR limits generated from the simulation study for half-siblings/unrelated different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

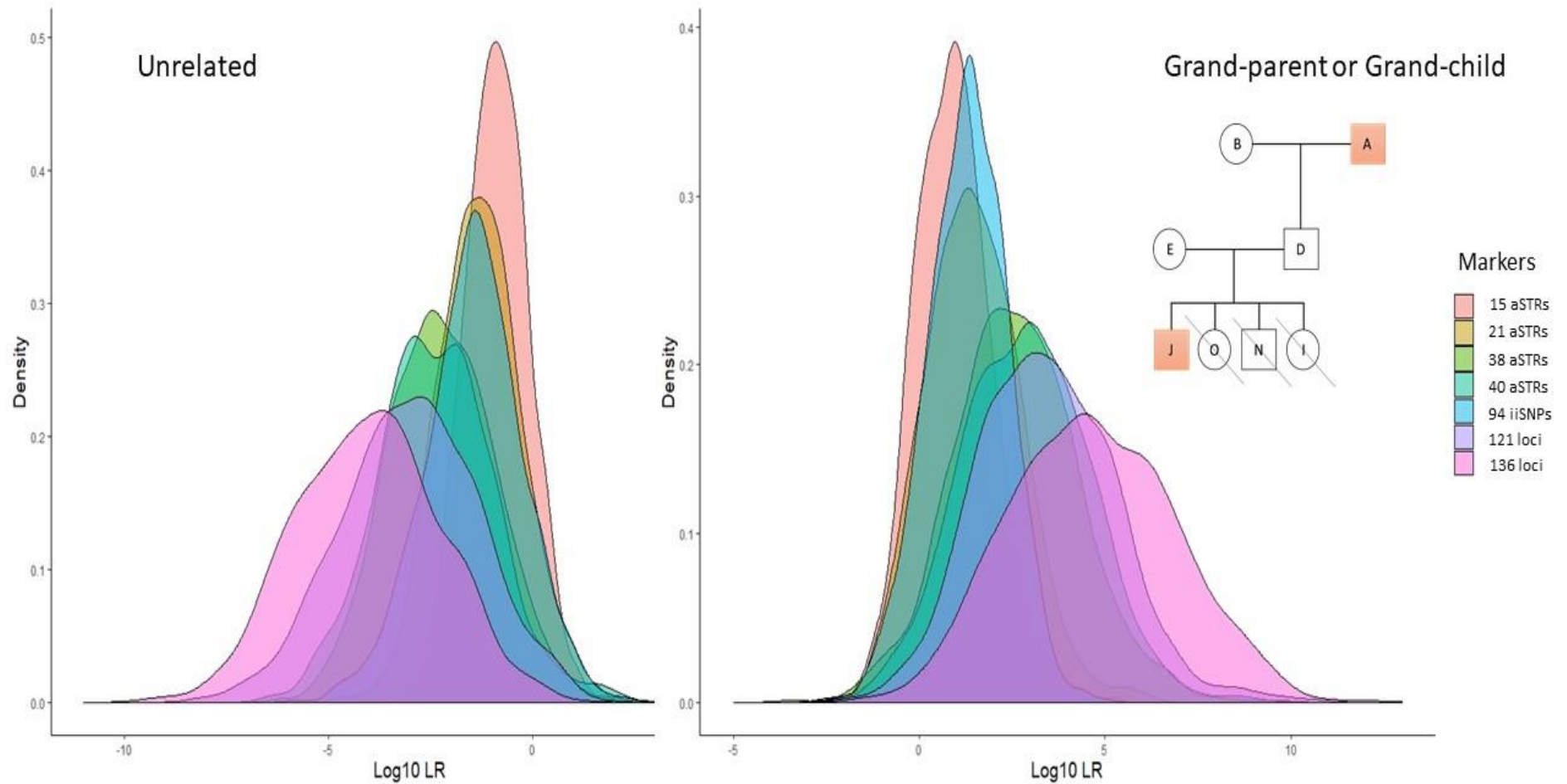


Figure 7.21. LR distributions of the simulation study for grand-parent or child/unrelated using different marker combinations. The figure also shows the case pedigree (hypothesis 1) as in (Table 2.5) where orange colour represents simulated members and crossed members were not available for testing..

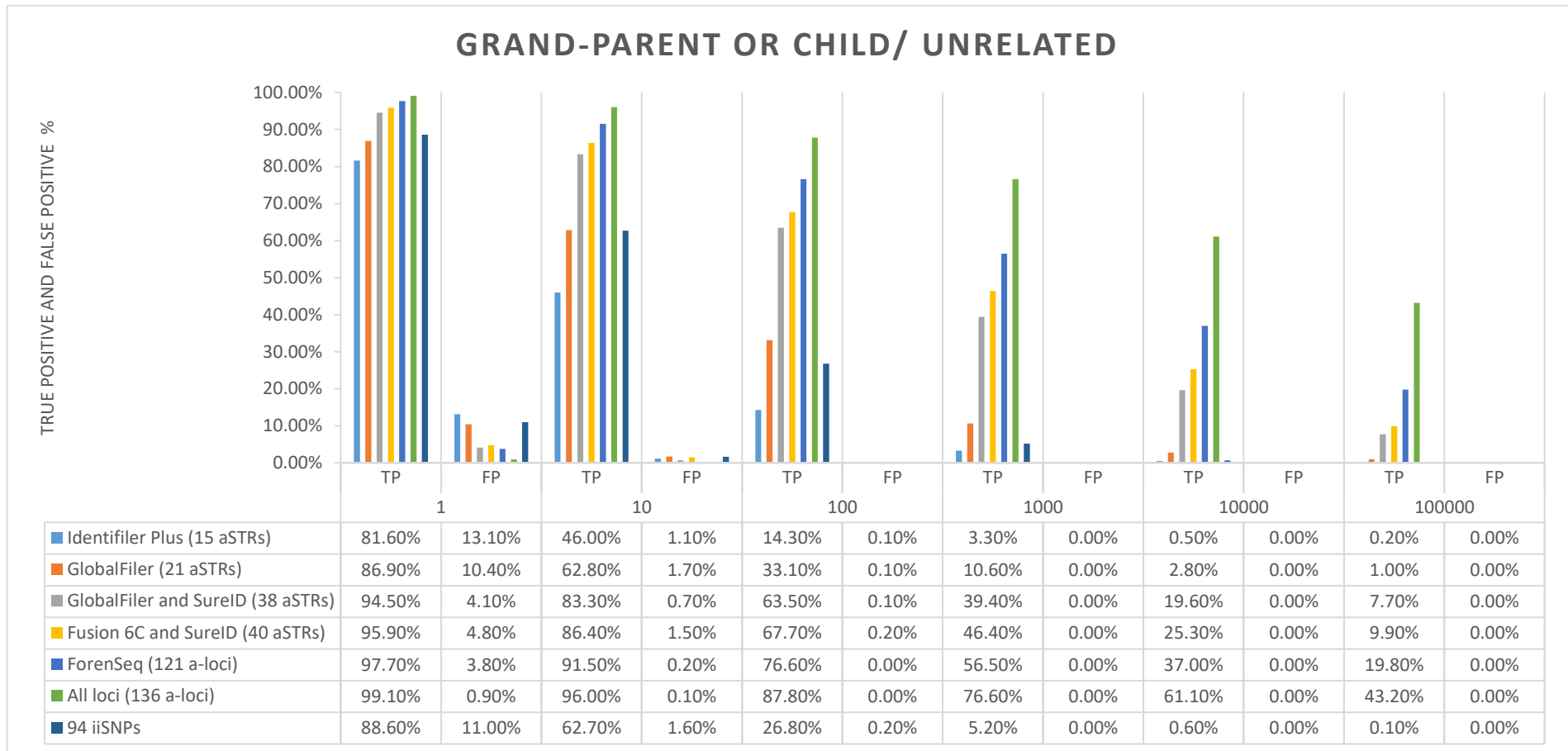


Figure 7.22. The TP and FP at different LR limits generated from the simulation study for grand-parent or child/unrelated using different marker combinations. Each marker set represented by a unique colour. True positive (TP) and false positive (FP).

In the parent-child relationship, where it was assumed that the genotype of the mother was not available, all sets had, at LR of 1, 100% TP and 0% FP except the 15 aSTRs that showed 99.6% TP with 0.3% FP (Figure 7.7 and Figure 7.8). This supports that using 15 aSTRs is not enough for duo pedigrees due to the probability of false inclusion or false exclusion (Poetsch *et al.* 2013). Although, the 21 aSTRs were able to reduce the false inclusion and exclusion to 0%, some cases where putative fathers are relatives e.g. (Goodwin *et al.* 2004), may need more loci to be included as the LR distributions for related (parent-child) and unrelated are nearly overlapping (Figure 7.7). In such cases, 38 or 40 aSTRs will be robust to differentiate between parent-child and unrelated even at LR of 10,000 with 100% TP and 0% FP (Figure 7.8).

The relationship of full-siblings/unrelated was also tested using three different scenarios that included simulation of only two siblings (Scenario 1) (Figure 7.9 and Figure 7.10), simulation of three siblings (Scenario 2) (Figure 7.11 and Figure 7.12), and simulation of four siblings (Scenario 3) (Figure 7.13 and Figure 7.14) (Table 7.1). Significant improvements in the TP and FP at all LR thresholds were observed when the third (Scenario 2) and the fourth (Scenario 3) siblings were included. For example, the 136 loci had TP percentage of 40.3% (Scenario 1), 97.30% (Scenario 2) and of 99.9% (Scenario 3) at LR threshold of 100,000. At the same LR level, the TP was also improved, when using 94 iiSNPs, from 0% (Scenario 1) to 21.5% (Scenario 2) and to 70.6% (Scenario 3). In addition, at LR of 1, the FP was reduced from 15% (15 loci) and from 10.7% (94 iiSNPs) (Scenario 1) to 3.5% and 2.3% (Scenario 2) and to 1.1% and 0.6% (Scenario 3), respectively.

First-cousin/unrelated relationship (Scenario 1) recorded the lowest percentage of TP and the highest percentage of FP in comparison with other relationships. Even with

the 136 loci the TP was 71.6%, which means 28.4% of related simulations appeared as unrelated (FN), and the FP was 22.6% (LR 1). Moreover, the 15 aSTRs and 94 iiSNPs had 57.8% and 62.2% TP (34.2% and 39.3% FP) at LR of 1 respectively, but both had 0% TP when using $LR \geq 100$ (Figure 7.15 and Figure 7.16). However, the differentiation was significantly improved when a grand-parent's genotypes were available and included in the simulation (Scenario 2). For example, at LR 1, the 136 loci had 99.8% TP with 0% FP and the performance of the 15 aSTRs and the 94 iiSNPs was improved to 92.2% (9.1% FP) and 96.5% (2.5% FP) respectively (Figure 7.17 and Figure 7.18).

The relatively poor discrimination when testing full-siblings and first-cousins compared to parent-child relationship came from the fact both relationships have lower probability of sharing alleles than in parent-child relationship. In parent-child relationship, there are 100% probability that a child shares half of the father's alleles and half of the mother's alleles. The shared alleles are termed as identical by descent (IBD) as they have come from the parents' ancestors. In full-siblings relationship; however, there is 25% probability of not having an IBD allele, 50% probability of having one IBD allele and 25% probability of having two IBD alleles (Figure 1.8). This is more difficult in first-cousin relationship as there is 75% probability of not having an IBD allele and 25% probability of having one IBD allele (Figure 1.8).

Despite the relatively poor discrimination when testing full-siblings and first-cousin, significant improvement can be obtained when more relatives were included. The probability of having IBD alleles is increased when more relatives are included and thus the TP and FP will be improved. It has been reported previously that adding more relatives to the tested pedigree is more powerful than studying more loci (Wenk and Shao 2012).

Half siblings could be differentiated from unrelated using 136 loci that showed 99.4% TP and only 1.1% FP at LR of 1. Interestingly, both 121 and 136 loci recorded 0% of FP at LR of 100 with 82.4% and 91.4% TP of the simulations respectively, while the 15 aSTRs had only 5.1% TP (Figure 7.19 and Figure 7.20). The grand-parent or child (Figure 7.21 and Figure 7.22) had similar discrimination power to that observed in half siblings. This may be due to that both relationships have the same probabilities of having IBD alleles (50% probability of not having IBD allele and 50% probability of having one shared allele), that has led to similar improvements in the LRs (Table 7.1).

7.5.3 Performance of the marker sets

Although the 94 iiSNPs have shown higher CMP of $1.24E-37$ (Chapter 6) compared to that can be obtained by the 21 aSTRs ($1.42E-26$, Chapter 3), their performance in kinship testing was similar over all relationships (Table 7.1) and (Figure 7.5 and Figure 7.6). This was due the fact binary markers have much lower performance than the multi-allelic markers (STRs). In parent-child/unrelated; however, more significant impact was noticed when using the 94 iiSNPs set (alone or when included to the 121 and 136 loci) in the LRs of unrelated compared to related simulations. This can be explained by the lower mutation rate of SNPs comparing to STRs and the impact was less when related was simulated since mutations were not present in related simulations (Daniel Kling, personal communication). In addition, this was only observed with parent-child relationship due the expected shared component of DNA between parents and offspring (100% probability of sharing one IBD allele at any locus).

As expected, the performance of the 40 aSTRs (Fusion 6C and SureID 23) was slightly higher than the 38 aSTRs (GlobalFiler and SureID 23) due to the additional two STRs (PentaD and PentaE). Using the Fusion 6C kit (Promega Corporation) would add value to

other forensic application (human identification) as two rapidly mutating (RM) Y-STRs (DYS576 and DYS570) are also included in the kit.

The 121 autosomal loci included in the ForenSeq™ DNA Signature Prep Kit (Primer Mix A) showed the highest discrimination power could be obtained from one kit, which was due to the number of typed markers and the inclusion of 94 iiSNPs. In addition, using the kit would improve the resolution of kinship testing by including sequence variants that can be observed in STRs (Ma *et al.* 2016) and in the flanking region of iiSNPs (King *et al.* 2018). Moreover, the kit includes 24 Y-STRs and 7 X-STRs that have been useful in many cases (e.g. (Junge *et al.* 2006) and (Li *et al.* 2012) respectively).

7.5.4 Potential linkage effect of closely located markers

It is worth knowing that all simulation tests carried out in this study were under the assumption of no linkage between all markers used. However, using more markers has raised concerns regarding including them in the product rule as independent markers, e.g. (Tillmar and Phillips 2017, O'Connor and Tillmar 2012). To address the impact of linkage in the LR estimation, family studies have been carried out to estimate RF between four syntenic pairs residing on the same chromosome arm: vWA-D12S391, D5S818-CSF1PO, TPOX-D2S441 and D21S11-PentaD typed when using any of the commonly used STR kits (Liu *et al.* 2014, Westen *et al.* 2012, Budowle *et al.* 2011, O'Connor and Tillmar 2012). RF values were 0.17 (Westen *et al.* 2012), 0.13 (Liu *et al.* 2014), 0.089 (O'Connor and Tillmar 2012) and 0.11 (Budowle *et al.* 2011) for vWA-D12S391; 0.197 (Buckleton and Triggs 2006) for D5S818-CSF1PO; 0.53 (Westen *et al.* 2012) for TPOX-D2S441 and 0.316 (Buckleton and Triggs 2006) D21S11-PentaD. The high-density multi-point SNP data of HapMap was also used to approximate the genetic distance between these syntenic loci and gave RF values of 0.12 for vWA-D12S391, 0.25

for D5S818-CSF1PO, 0.36 for D21S11-PentaD, and of 0.47 for TPOX-D2S441 (Phillips *et al.* 2012); these values are similar to those generated using family studies (Alsafiah *et al.* 2019a).

When using additional markers available in supplementary kits (e.g. SureID 23) in conjunction with commonly used kits, or when using MPS kits that include a large number of markers, the number of syntenic loci will be increased. Although excluding the less informative locus from the probability estimation (Budowle *et al.* 2011) is an option, this may lead to an overestimation or to an underestimation of the strength of an evidence (Gill *et al.* 2012).

Gill *et al.* (2012) has addressed the impact of using the closely located vWA-D12S391 and concluded that, for most pedigrees, syntenic loci with RF value of ~ 0.12 has almost zero effect in any population as long as no linkage disequilibrium is detected. For some pedigrees, where at least one individual has a heterozygote genotype in both syntenic loci and is involved in at least two transmissions of genetic components, linkage can have a significant effect in the product rule calculation in kinship testing (i. e. incest cases) (Alsafiah *et al.* 2019a).

FamLink software v.1.16 (Kling *et al.* 2012) allows the calculation of case specific LRs taking in account linkage between syntenic pairs. In addition, it can perform simulations to study the impact of ignoring the linkage on the LRs calculation for each linked pair. To use the software, the RFs between potentially linked pairs should be estimated and used in the FamLink setting.

Therefore, LD test was carried out for all syntenic loci included in this study using the data of the 500 samples (38 aSTRs) from Chapters 3 and 5 and the data of 87 samples (121 loci) from Chapter 6. In addition, the RFs were estimated for each syntenic loci using

high-density multi-point SNP data of HapMap as described by Phillips *et al.* (2012) after estimating the cumulative genetic map distances for each locus.

The cumulative genetic map distances (cM) of 41/42 aSTR were already published in (Phillips 2017), while the cumulative genetic map distances (cM) of D16S539 and 94 iiSNPs were estimated using the HapMap data as described by Phillips *et al.* (2012). The cumulative genetic map distance (cM) of D16S539 was not estimated in (Phillips 2017) as there was not any other STR located in chromosome 16 in that study. However, in this study, there are four iiSNPs located in chromosome 16 (rs729172, rs2342747, rs430046 and rs1382387) that necessitates estimating the cumulative genetic map distances of D16S539.

The cumulative genetic map distances (cM) of 95 SNPs (94 iiSNPs and rs925658351 SNP for D16S539) were estimated and are shown in Section 10.6.2 (Appendix 6). This was followed by calculating the RFs of all syntenic loci using the Excel tool provided by Phillips *et al.* (2012) as shown in Figure 2.4.

Three tables are presented for the results of LD test and for RFs estimations. Table 7.2 shows the results of syntenic pairs at the same arm resulted from using SureID 23 kit in conjunction with GlobalFiler or with Fusion 6C kits (12 or 14 STR-STR pairs respectively). Table 7.3 shows the results of syntenic pairs at the same arm (166 STR-STR, STR-SNP and SNP-SNP pairs) resulted when using the ForenSeq DNA Signature Prep kit alone (with an assumption that SE33 was typed as in Chapter 6). Table 7.4 shows the results of additional 50 syntenic pairs at the same arm (STR-STR and STR-SNP pairs) resulted when using the 136 loci included in all kits (GlobalFiler, SureID 23 and ForenSeq DNA Signature Prep kit).

Table 7.2. The results of the LD test for 14 syntenic STR-STR pairs (at the same arm) resulted from using SureID 23 kit in conjunction with GlobalFiler (12 syntenic pairs, P value = 0.004) or with Fusion 6C (14 syntenic pairs, P value = 0.0035) and their RF values. The RFs were calculated using Kosambi mapping function using genetic map distance in cM estimated using cumulative genetic map distance in cM which were reviewed from (Phillips 2017). None of the syntenic pairs showed LD after Bonferroni correction. The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/12$ STRs = 0.004 and $0.05/14$ = 0.0035. Shaded rows show all syntenic pairs with RFs < 0.12. Cautions should be considered when including D18S51-D18S1364 and PentaD-D21S2055 pairs in the calculation of LR_s due to low RFs. The pair vWA-D12S391 will not have significant impact for most pedigrees as RF is ~ 0.12 (Gill *et al.* 2012).

No.	Location		Syntenic pair		LD test p value	Cumulative genetic map distance in cM		Genetic map distance in cM	RFs from Kosambi mapping function
	Chr.	Arm	Locus 1	Locus 2		Locus 1	Locus 2		
1	Chr.2	p-p	TPOX	D2S441	0.97764	1.6661	90.47903	88.81293	0.472145669
2	Chr.5	q-q	D5S2800	D5S818	0.15198	70.3208	126.67284	56.35204	0.405002025
3	Chr.5	q-q	D5S2800	CSF1PO	0.85646	70.3208	154.43395	84.11315	0.466577301
4	Chr.5	q-q	D5S818	CSF1PO	0.69008	126.67284	154.43395	27.76111	0.252211952
5	Chr.6	q-q	SE33	D6S474	0.99963	95.44921	118.66248	23.21327	0.216777122
6	Chr.8	q-q	D8S1132	D8S1179	0.23577	119.96228	136.44313	16.48085	0.159088296
7	Chr.11	p-p	TH01	D11S2368	0.20421	4.48933	32.88891	28.39958	0.256941387
8	Chr.12	p-p	vWA	D12S391	0.89307	15.63031	27.57129	11.94098	0.117190251
9	Chr.13	q-q	D13S325	D13S317	0.97422	44.90825	79.83074	34.92249	0.301691441
10	Chr.18	q-q	D18S51	D18S1364	0.04312	84.639759	91.21746	6.577701	0.065400163
11	Chr.21	q-q	D21S11	D21S2055	1	14.64555	49.46478	34.81923	0.301033962
12	Chr.22	q-q	D22S1045	D22GATA198B05	0.98893	46.21362	7.39585	38.81777	0.325304911
13	Chr.15	q-q	PentaE	D15S659	0.99899*	124.05054	49.51748	74.53306	0.451723167
14	Chr.21	q-q	PentaD	D21S2055	0.96176*	59.37591	49.46478	9.91113	0.097833282

* the LD test was carried out using the data of 500 samples from Chapter 3 and 5, except pairs No. 13 and 14 where LD test was carried out using the data of PentaE and PentaD from the 87 samples in Chapter 6 as they are not included in the SureID23 or in the GlobalFiler kits.

Table 7.3. The results of the LD test for 166 syntenic (STR-STR, STR-SNP and SNP-SNP) pairs (at the same arm) resulted from using ForenSeq DNA Signature Prep kit alone and the RF values. The RFs were calculated using Kosambi mapping function using genetic map distance in cM estimated using high-density multi-point SNP data of HapMap as described by Phillips *et al.* (2012). The cumulative genetic map distance in cM of 27 aSTRs were reviewed from (Phillips 2017) and of the 94 iSNPs were estimated as described by Phillips *et al.* (2012) (Appendix 6, Section 10.6.2). None of the syntenic pairs showed LD after Bonferroni correction (P value = 0.0003). The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/166$ pairs = 0.0003. Shaded rows present pairs with RFs < 0.12 (43 pairs). This table assumed that SE33 was typed as shown in Chapter 6. The data of the 87 samples were used in the test of LD.

No.	Location		Pair type	Syntenic pair		LD P value	Cumulative genetic map distance in cM		Genetic map distance in cM	RFs from Kosambi mapping function
	Chr.	Arm		Locus 1	Locus 2		Locus 1	Locus 2		
1	Chr.1	q-q	STR-SNP	D1S1656	rs560681	0.92158	244.2349	173.5187	70.71623	0.444204471
2	Chr.1	q-q	STR-SNP	D1S1656	rs1294331	0.0004	244.2349	252.6893	8.454454	0.083747908
3	Chr.1	q-q	SNP-SNP	rs560681	rs1294331	0.20475	173.5187	252.6893	79.17068	0.459566663
4	Chr.1	q-q	STR-SNP	D1S1656	rs10495407	0.98907	244.2349	264.5096	20.27471	0.192320103
5	Chr.1	q-q	SNP-SNP	rs560681	rs10495407	0.95563	173.5187	264.5096	90.99094	0.474410174
6	Chr.1	q-q	SNP-SNP	rs1294331	rs10495407	0.60182	252.6893	264.5096	11.82026	0.116048705
7	Chr.1	q-q	STR-SNP	D1S1656	rs891700	0.84749	244.2349	266.7565	22.52162	0.211127162
8	Chr.1	q-q	SNP-SNP	rs560681	rs891700	0.34016	173.5187	266.7565	93.23784	0.476558202
9	Chr.1	q-q	SNP-SNP	rs1294331	rs891700	0.11079	252.6893	266.7565	14.06716	0.137073921
10	Chr.1	q-q	SNP-SNP	rs10495407	rs891700	0.37552	264.5096	266.7565	2.246905	0.022453937
11	Chr.1	q-q	STR-SNP	D1S1656	rs1413212	0.26974	244.2349	275.1116	30.87668	0.274704236
12	Chr.1	q-q	SNP-SNP	rs560681	rs1413212	0.66458	173.5187	275.1116	101.5929	0.4831053
13	Chr.1	q-q	SNP-SNP	rs1294331	rs1413212	0.93969	252.6893	275.1116	22.42223	0.210309797
14	Chr.1	q-q	SNP-SNP	rs10495407	rs1413212	0.4255	264.5096	275.1116	10.60197	0.104458858
15	Chr.1	q-q	SNP-SNP	rs891700	rs1413212	0.92589	266.7565	275.1116	8.355065	0.082781581
16	Chr.2	p-p	STR-STR	TPOX	D2S441	0.92419	1.6661	90.47903	88.81293	0.472145669
17	Chr.2	p-p	STR-SNP	TPOX	rs876724	0.54976	1.6661	0.054278	1.611822	0.016112639
18	Chr.2	p-p	STR-SNP	D2S441	rs876724	0.94563	90.47903	0.054278	90.42475	0.473839353
19	Chr.2	p-p	STR-SNP	TPOX	rs1109037	0.66548	1.6661	25.84589	24.17979	0.224559402
20	Chr.2	p-p	STR-SNP	D2S441	rs1109037	0.97426	90.47903	25.84589	64.63314	0.429911158
21	Chr.2	p-p	SNP-SNP	rs876724	rs1109037	0.4241	0.054278	25.84589	25.79162	0.237238494
22	Chr.2	q-q	STR-SNP	D2S1338	rs993934	0.99215	223.4832	143.1388	80.34436	0.461349352
23	Chr.2	q-q	STR-SNP	D2S1338	rs12997453	0.96577	223.4832	196.6693	26.81395	0.245083072
24	Chr.2	q-q	SNP-SNP	rs993934	rs12997453	0.53139	143.1388	196.6693	53.53041	0.394845118
25	Chr.2	q-q	STR-SNP	D2S1338	rs907100	0.17878	223.4832	261.3676	37.88436	0.319856327
26	Chr.2	q-q	SNP-SNP	rs993934	rs907100	0.29721	143.1388	261.3676	118.2287	0.491243368
27	Chr.2	q-q	SNP-SNP	rs12997453	rs907100	0.2164	196.6693	261.3676	64.69831	0.43008088
28	Chr.3	p-p	STR-SNP	D3S1358	rs1357617	0.69048	67.1789	1.267142	65.91176	0.433172183

Table 7.3. continued.

29	Chr.3	p-p	STR-SNP	D3S1358	rs4364205	0.01625	67.1789	56.4601	10.7188	0.10557559
30	Chr.3	p-p	SNP-SNP	rs1357617	rs4364205	0.21329	1.267142	56.4601	55.19296	0.400940491
31	Chr.3	q-q	SNP-SNP	rs2399332	rs1355366	0.0964	120.1666	209.7995	89.63285	0.473020138
32	Chr.3	q-q	SNP-SNP	rs2399332	rs6444724	0.14662	120.1666	214.0278	93.86121	0.477122283
33	Chr.3	q-q	SNP-SNP	rs1355366	rs6444724	0.35106	209.7995	214.0278	4.22836	0.042183089
34	Chr.4	p-p	STR-SNP	D4S2408	rs2046361	0.0473	49.54939	26.4958	23.05359	0.215478653
35	Chr.4	p-p	STR-SNP	D4S2408	rs279844	0.88737	49.54939	68.75248	19.20309	0.183114834
36	Chr.4	p-p	SNP-SNP	rs2046361	rs279844	0.99515	26.4958	68.75248	42.25668	0.34425927
37	Chr.4	q-q	STR-SNP	FGA	rs6811238	0.82269	156.8129	174.3913	17.57833	0.168882112
38	Chr.4	q-q	STR-SNP	FGA	rs1979255	0.98527	156.8129	213.0553	56.24236	0.404624171
39	Chr.4	q-q	SNP-SNP	rs6811238	rs1979255	0.55961	174.3913	213.0553	38.66403	0.324416484
40	Chr.5	q-q	STR-STR	D5S818	CSF1PO	0.61363	126.6728	154.434	27.76111	0.252211952
41	Chr.5	p-p	SNP-SNP	rs717302	rs159606	0.70771	6.711702	33.52614	26.81443	0.245086742
42	Chr.5	q-q	STR-SNP	D5S818	rs1318288	0.29308	126.6728	139.7681	13.09522	0.128037968
43	Chr.5	q-q	STR-SNP	CSF1PO	rs1318288	0.95781	154.434	139.7681	14.66589	0.142592815
44	Chr.5	q-q	STR-SNP	D5S818	rs251934	0.63361	126.6728	191.9862	65.3134	0.431664031
45	Chr.5	q-q	STR-SNP	CSF1PO	rs251934	0.51731	154.434	191.9862	37.55229	0.317886224
46	Chr.5	q-q	SNP-SNP	rs1318288	rs251934	0.21562	139.7681	191.9862	52.21818	0.389802681
47	Chr.5	q-q	STR-SNP	D5S818	rs338882	0.56016	126.6728	199.6403	72.96742	0.448762987
48	Chr.5	q-q	STR-SNP	CSF1PO	rs338882	0.36063	154.434	199.6403	45.20631	0.359150533
49	Chr.5	q-q	SNP-SNP	rs1318288	rs338882	0.40301	139.7681	199.6403	59.8722	0.416436656
50	Chr.5	q-q	SNP-SNP	rs251934	rs338882	0.68537	191.9862	199.6403	7.654021	0.07594789
51	Chr.6	q-q	STR-SNP	D6S1043	rs1336071	0.03743	99.86628	100.6511	0.784821	0.007847566
52	Chr.6	q-q	STR-SNP	D6S1043	rs214955	0.51	99.86628	159.8483	59.98204	0.416772515
53	Chr.6	q-q	SNP-SNP	rs1336071	rs214955	0.2689	100.6511	159.8483	59.19722	0.414345667
54	Chr.6	q-q	STR-SNP	D6S1043	rs727811	0.94169	99.86628	180.0571	80.19079	0.461120464
55	Chr.6	q-q	SNP-SNP	rs1336071	rs727811	0.76128	100.6511	180.0571	79.40597	0.459930248
56	Chr.6	q-q	SNP-SNP	rs214955	rs727811	0.376	159.8483	180.0571	20.20875	0.191757788
57	Chr.6	q-q	STR-STR	D6S1043	SE33	1	99.86628	95.44921	4.41707	0.044056152
58	Chr.6	q-q	SNP-STR	rs1336071	SE33	0.87309	100.6511	95.44921	5.201891	0.051832037
59	Chr.6	q-q	SNP-STR	rs214955	SE33	0.99894	159.8483	95.44921	64.39911	0.429298586
60	Chr.6	q-q	SNP-STR	rs727811	SE33	0.98595	180.0571	95.44921	84.60786	0.467210713
61	Chr.7	p-p	SNP-SNP	rs6955448	rs917118	0.5883	6.912354	7.494464	0.58211	0.005820837
62	Chr.7	q-q	STR-SNP	D7S820	rs321198	0.61806	100.2012	145.3779	45.17667	0.359007011
63	Chr.7	q-q	STR-SNP	D7S820	rs737681	0.37979	100.2012	181.9196	81.71839	0.463340543
64	Chr.7	q-q	SNP-SNP	rs321198	rs737681	0.98621	145.3779	181.9196	36.54172	0.311787756
65	Chr.8	p-p	SNP-SNP	rs763869	rs1009249	0.94969	1.957165	56.01666	54.0595	0.396819976
66	Chr.8	q-q	STR-SNP	D8S1179	rs2056277	0.06427	136.4431	156.441	19.9979	0.189956504
67	Chr.8	q-q	STR-SNP	D8S1179	rs4606077	0.73678	136.4431	166.5671	30.12392	0.269405425
68	Chr.8	q-q	SNP-SNP	rs2056277	rs4606077	0.37239	156.441	166.5671	10.12603	0.09989821
69	Chr.9	p-p	SNP-SNP	rs1015250	rs7041158	0.58814	4.30155	53.00553	48.70398	0.3752458

Table 7.3. continued.

70	Chr.9	q-q	STR-SNP	D9S1122	rs1463729	0.78652	81.15767	136.0526	54.89488	0.39987129
71	Chr.9	q-q	STR-SNP	D9S1122	rs1360288	0.58519	81.15767	137.9145	56.75681	0.406384914
72	Chr.9	q-q	SNP-SNP	rs1463729	rs1360288	0.13896	136.0526	137.9145	1.861931	0.018610708
73	Chr.9	q-q	STR-SNP	D9S1122	rs10776839	0.35676	81.15767	155.8453	74.68764	0.452006465
74	Chr.9	q-q	SNP-SNP	rs1463729	rs10776839	0.7046	136.0526	155.8453	19.79276	0.188198434
75	Chr.9	q-q	SNP-SNP	rs1360288	rs10776839	0.95171	137.9145	155.8453	17.93083	0.171997414
76	Chr.10	p-p	SNP-SNP	rs826472	rs735155	0.9813	3.568912	6.796451	3.227539	0.032230636
77	Chr.10	p-p	SNP-SNP	rs826472	rs3780962	0.82519	3.568912	38.18664	34.61773	0.299746241
78	Chr.10	p-p	SNP-SNP	rs735155	rs3780962	0.008	6.796451	38.18664	31.39019	0.278269047
79	Chr.10	q-q	STR-SNP	D10S1248	rs740598	0.61305	169.8992	143.7301	26.16903	0.240152437
80	Chr.10	q-q	STR-SNP	D10S1248	rs964681	0.24918	169.8992	175.6694	5.770234	0.057447533
81	Chr.10	q-q	SNP-SNP	rs740598	rs964681	0.85486	143.7301	175.6694	31.93926	0.282035914
82	Chr.11	p-p	STR-SNP	TH01	rs1498553	0.44699	4.48933	11.57216	7.082833	0.070358341
83	Chr.11	p-p	STR-SNP	TH01	rs901398	0.40305	4.48933	20.23465	15.74532	0.152446978
84	Chr.11	p-p	SNP-SNP	rs1498553	rs901398	0.43619	11.57216	20.23465	8.662483	0.085768417
85	Chr.11	q-q	SNP-SNP	rs10488710	rs2076848	0.19635	119.9957	157.8437	37.84798	0.319641288
86	Chr.12	p-p	STR-STR	vWA	D12S391	1	15.63031	27.57129	11.94098	0.117190251
87	Chr.12	p-p	STR-SNP	vWA	rs2107612	0.092	15.63031	2.139891	13.49042	0.131723262
88	Chr.12	p-p	STR-SNP	D12S391	rs2107612	0.11141	27.57129	2.139891	25.4314	0.234437749
89	Chr.12	p-p	STR-SNP	vWA	rs2269355	0.34054	15.63031	17.7073	2.076994	0.020758002
90	Chr.12	p-p	STR-SNP	D12S391	rs2269355	0.01118	27.57129	17.7073	9.863986	0.097379808
91	Chr.12	p-p	SNP-SNP	rs2107612	rs2269355	0.09493	2.139891	17.7073	15.56741	0.150831581
92	Chr.12	q-q	SNP-SNP	rs2920816	rs2111980	0.4073	56.2715	124.5179	68.24635	0.438765443
93	Chr.12	q-q	SNP-SNP	rs2920816	rs10773760	0.01395	56.2715	168.4425	112.171	0.488869132
94	Chr.12	q-q	SNP-SNP	rs2111980	rs10773760	0.53318	124.5179	168.4425	43.92465	0.352831761
95	Chr.13	q-q	STR-SNP	D13S317	rs1335873	0.1836	79.83074	2.118193	77.71255	0.45724208
96	Chr.13	q-q	STR-SNP	D13S317	rs1886510	0.44997	79.83074	4.798954	75.03179	0.452631533
97	Chr.13	q-q	SNP-SNP	rs1335873	rs1886510	0.67946	2.118193	4.798954	2.680761	0.026781953
98	Chr.13	q-q	STR-SNP	D13S317	rs1058083	0.20642	79.83074	94.11131	14.28057	0.139045264
99	Chr.13	q-q	SNP-SNP	rs1335873	rs1058083	0.8001	2.118193	94.11131	91.99312	0.475390962
100	Chr.13	q-q	SNP-SNP	rs1886510	rs1058083	0.98344	4.798954	94.11131	89.31235	0.472681542
101	Chr.13	q-q	STR-SNP	D13S317	rs354439	0.05434	79.83074	107.2948	27.46404	0.249990473
102	Chr.13	q-q	SNP-SNP	rs1335873	rs354439	0.4077	2.118193	107.2948	105.1766	0.485328429
103	Chr.13	q-q	SNP-SNP	rs1886510	rs354439	0.18306	4.798954	107.2948	102.4958	0.483694821
104	Chr.13	q-q	SNP-SNP	rs1058083	rs354439	0.85193	94.11131	107.2948	13.18347	0.128862206
105	Chr.14	q-q	SNP-SNP	rs1454361	rs722290	0.40185	17.19934	47.50283	30.30349	0.270677336
106	Chr.14	q-q	SNP-SNP	rs1454361	rs873196	0.8474	17.19934	104.0042	86.80488	0.469886201
107	Chr.14	q-q	SNP-SNP	rs722290	rs873196	0.17717	47.50283	104.0042	56.50139	0.405514378
108	Chr.14	q-q	SNP-SNP	rs1454361	rs4530059	0.35381	17.19934	114.5175	97.31812	0.480017721

Table 7.3. continued.

109	Chr.14	q-q	SNP-SNP	rs722290	rs4530059	0.16601	47.50283	114.5175	67.01463	0.435871244
110	Chr.14	q-q	SNP-SNP	rs873196	rs4530059	0.29418	104.0042	114.5175	10.51324	0.103609962
111	Chr.15	q-q	STR-SNP	PentaE	rs1821380	0.90582	124.0505	53.23968	70.81086	0.444403566
112	Chr.15	q-q	STR-SNP	PentaE	rs8037429	0.05451	124.0505	64.45011	59.60043	0.415600384
113	Chr.15	q-q	SNP-SNP	rs1821380	rs8037429	0.15606	53.23968	64.45011	11.21043	0.110262881
114	Chr.15	q-q	STR-SNP	PentaE	rs1528460	0.01971	124.0505	66.37152	57.67902	0.409468346
115	Chr.15	q-q	SNP-SNP	rs1821380	rs1528460	0.62987	53.23968	66.37152	13.13185	0.128380147
116	Chr.15	q-q	SNP-SNP	rs8037429	rs1528460	0.72056	64.45011	66.37152	1.921413	0.019204678
117	Chr.16	p-p	SNP-SNP	rs729172	rs2342747	0.45601	11.31258	11.86134	0.548754	0.00548732
118	Chr.16	q-q	STR-SNP	D16S539	rs430046	0.93666	125.5782	97.20913	28.3691	0.256717036
119	Chr.16	q-q	STR-SNP	D16S539	rs1382387	0.48523	125.5782	103.7257	21.85252	0.205598289
120	Chr.16	q-q	SNP-SNP	rs430046	rs1382387	0.61647	97.20913	103.7257	6.516582	0.064799334
121	Chr.17	p-p	SNP-SNP	rs9905977	rs740910	0.09089	8.279761	13.40866	5.128894	0.051109803
122	Chr.17	q-q	STR-SNP	D17S1301	rs938283	0.05557	113.1115	120.3081	7.196665	0.071473761
123	Chr.17	q-q	STR-SNP	D17S1301	rs8078417	0.31548	113.1115	127.7513	14.6399	0.142354018
124	Chr.17	q-q	SNP-SNP	rs938283	rs8078417	0.58137	120.3081	127.7513	7.443234	0.073887346
125	Chr.18	p-p	SNP-SNP	rs1493232	rs9951171	0.42172	3.666872	28.53392	24.86705	0.230011673
126	Chr.18	q-q	STR-SNP	D18S51	rs1736442	0.2397	84.63976	74.55715	10.08261	0.09948127
127	Chr.18	q-q	STR-SNP	D18S51	rs1024116	0.0422	84.63976	112.7889	28.14917	0.255093811
128	Chr.18	q-q	SNP-SNP	rs1736442	rs1024116	0.6107	74.55715	112.7889	38.23177	0.321899622
129	Chr.19	q-q	STR-SNP	D19S433	rs719366	0.43138	51.72618	49.40652	2.319663	0.023180002
130	Chr.19	q-q	STR-SNP	D19S433	rs576261	0.41759	51.72618	63.83692	12.11074	0.118793304
131	Chr.19	q-q	SNP-SNP	rs719366	rs576261	0.93067	49.40652	63.83692	14.4304	0.140426575
132	Chr.20	p-p	STR-SNP	D20S482*	rs1031825	0.82481	13.25549	12.79543	0.460058	0.00460045
133	Chr.20	p-p	STR-SNP	D20S482*	rs445251	0.97978	13.25549	35.36648	22.11099	0.207741353
134	Chr.20	p-p	SNP-SNP	rs1031825	rs445251	0.53286	12.79543	35.36648	22.57105	0.211533151
135	Chr.20	q-q	SNP-SNP	rs1005533	rs1523537	0.49579	58.01538	77.58417	19.56879	0.186272852
136	Chr.21	q-q	STR-STR	D21S11	PentaD	1	14.64555	59.37591	44.73036	0.356830933
137	Chr.21	q-q	STR-SNP	D21S11	rs722098	0.28111	14.64555	4.539526	10.10602	0.099706169
138	Chr.21	q-q	STR-SNP	PentaD	rs722098	0.16683	59.37591	4.539526	54.83638	0.399660259
139	Chr.21	q-q	STR-SNP	D21S11	rs2830795	0.99158	14.64555	27.34826	12.70271	0.124362925
140	Chr.21	q-q	STR-SNP	PentaD	rs2830795	0.33265	59.37591	27.34826	32.02765	0.28263799
141	Chr.21	q-q	SNP-SNP	rs722098	rs2830795	0.16096	4.539526	27.34826	22.80873	0.213480666
142	Chr.21	q-q	STR-SNP	D21S11	rs2831700	0.78972	14.64555	29.39708	14.75153	0.143379188
143	Chr.21	q-q	STR-SNP	PentaD	rs2831700	0.57032	59.37591	29.39708	29.97883	0.268374108
144	Chr.21	q-q	SNP-SNP	rs722098	rs2831700	0.99355	4.539526	29.39708	24.85755	0.229936842
145	Chr.21	q-q	SNP-SNP	rs2830795	rs2831700	0.17265	27.34826	29.39708	2.048821	0.020476751
146	Chr.21	q-q	STR-SNP	D21S11	rs914165	0.99605	14.64555	50.55435	35.9088	0.30788909
147	Chr.21	q-q	STR-SNP	PentaD	rs914165	0.79375	59.37591	50.55435	8.821562	0.08731155

Table 7.3. continued.

148	Chr.21	q-q	SNP-SNP	rs722098	rs914165	0.49236	4.539526	50.55435	46.01482	0.363018811
149	Chr.21	q-q	SNP-SNP	rs2830795	rs914165	0.1051	27.34826	50.55435	23.20609	0.216718806
150	Chr.21	q-q	SNP-SNP	rs2831700	rs914165	0.26508	29.39708	50.55435	21.15727	0.199788458
151	Chr.21	q-q	STR-SNP	D21S11	rs221956	0.40982	14.64555	54.76922	40.12367	0.332708611
152	Chr.21	q-q	STR-SNP	PentaD	rs221956	0.13376	59.37591	54.76922	4.606694	0.045937032
153	Chr.21	q-q	SNP-SNP	rs722098	rs221956	0.14699	4.539526	54.76922	50.22969	0.381758347
154	Chr.21	q-q	SNP-SNP	rs2830795	rs221956	0.0073	27.34826	54.76922	27.42096	0.249667226
155	Chr.21	q-q	SNP-SNP	rs2831700	rs221956	0.94269	29.39708	54.76922	25.37214	0.233975149
156	Chr.21	q-q	SNP-SNP	rs914165	rs221956	0.80822	50.55435	54.76922	4.214868	0.042049126
157	Chr.22	q-q	STR-SNP	D22S1045	rs733164	0.6213	46.21362	31.36631	14.84731	0.14425778
158	Chr.22	q-q	STR-SNP	D22S1045	rs987640	0.93037	46.21362	37.65417	8.559449	0.084768042
159	Chr.22	q-q	SNP-SNP	rs733164	rs987640	0.90829	31.36631	37.65417	6.287865	0.06254926
160	Chr.22	q-q	STR-SNP	D22S1045	rs2040411	0.52838	46.21362	62.88724	16.67362	0.160818685
161	Chr.22	q-q	SNP-SNP	rs733164	rs2040411	0.62269	31.36631	62.88724	31.52093	0.2791702
162	Chr.22	q-q	SNP-SNP	rs987640	rs2040411	0.08637	37.65417	62.88724	25.23307	0.232887568
163	Chr.22	q-q	STR-SNP	D22S1045	rs1028528	0.35004	46.21362	64.13652	17.9229	0.171927565
164	Chr.22	q-q	SNP-SNP	rs733164	rs1028528	0.73114	31.36631	64.13652	32.77022	0.287648446
165	Chr.22	q-q	SNP-SNP	rs987640	rs1028528	0.95043	37.65417	64.13652	26.48235	0.24255562
166	Chr.22	q-q	SNP-SNP	rs2040411	rs1028528	0.20921	62.88724	64.13652	1.249286	0.012490261

* D20S482 was tested for LD as it did not show significant departure from HWE when the 87 samples were sequenced in Chapter 6.

Table 7.4. The results of the LD test for additional 50 syntenic (STR-STR and STR-SNP) pairs (at the same arm) resulted from combining GlobalFiler, SureID23 and ForenSeq DNA Signature Prep kits. The cumulative genetic map distance in cM of 12 STRs were reviewed from (Phillips 2017) and of D16S539 with the 94 iSNPs were estimated as described by Phillips *et al.* (2012) (Appendix 6, Section 10.6.2). The RFs were calculated by Kosambi mapping function using genetic map distance in cM that was estimated using high-density multi-point SNP data of HapMap as described by Phillips *et al.* (2012). None of the syntenic pairs showed LD after Bonferroni correction (P value = 0.00023). The Bonferroni correction was performed by dividing 0.05 by the number of tested pairs (the number of tests being performed), i.e. $0.05/216$ pairs (166 pairs from Table 7.3 and 50 from this table) = 0.00023. Shaded rows present pairs with RFs < 0.12) (49 pairs in total when using the 136 loci).

No.	Location		Pair type	Syntenic pair		LD P value	Cumulative genetic map distance in cM		Genetic map distance in cM	RFs from Kosambi mapping function
	Chr.	Arm		Locus 1	Locus 2		Locus 1	Locus 2		
1	Chr. 3	q-q	STR-SNP	D3S1744	rs2399332	0.90477	157.2413	120.1666	37.07471	0.315023543
2	Chr. 3	q-q	STR-SNP	D3S1744	rs1355366	0.34413	157.2413	209.7995	52.55814	0.391129009
3	Chr. 3	q-q	STR-SNP	D3S1744	rs6444724	0.10599	157.2413	214.0278	56.7865	0.406485625
4	Chr. 4	p-p	STR-SNP	D4S2366	rs2046361	0.71089	12.9467	26.4958	13.5491	0.132269204
5	Chr. 4	p-p	STR-STR	D4S2366	D4S2408	0.41066	12.9467	49.54939	36.60269	0.312160118
6	Chr. 4	p-p	STR-SNP	D4S2366	rs279844	0.36692	12.9467	68.75248	55.80578	0.403106769
7	Chr. 5	q-q	STR-STR	D5S2800	D5S818	0.15198	70.3208	126.6728	56.35204	0.405002025
8	Chr. 5	q-q	STR-SNP	D5S2800	rs13182883	0.80499	70.3208	139.7681	69.44726	0.441469306
9	Chr. 5	q-q	STR-STR	D5S2800	CSF1PO	0.85646	70.3208	154.434	84.11315	0.466577301
10	Chr. 5	q-q	STR-SNP	D5S2800	rs251934	0.12807	70.3208	191.9862	121.6654	0.492359464
11	Chr. 5	q-q	STR-SNP	D5S2800	rs338882	0.71809	70.3208	199.6403	129.3195	0.494363158
12	Chr. 6	q-q	STR-STR	D6S474	D6S1043	0.94319	118.6625	99.86628	18.7962	0.179581236
13	Chr. 6	q-q	STR-SNP	D6S474	rs1336071	0.28042	118.6625	100.6511	18.01138	0.172707239
14	Chr. 6	q-q	STR-SNP	D6S474	rs214955	0.49762	118.6625	159.8483	41.18584	0.338543909
15	Chr. 6	q-q	STR-SNP	D6S474	rs727811	0.24364	118.6625	180.0571	61.39459	0.420983376
16	Chr. 6	q-q	STR-STR	D6S474	SE33	0.99963	118.6625	95.44921	23.21327	0.216777122
17	Chr. 7	p-p	STR-SNP	D7S3048	rs6955448	0.6802	36.14071	6.912354	29.22836	0.26298849
18	Chr. 7	p-p	STR-SNP	D7S3048	rs917118	0.98508	36.14071	7.494464	28.64625	0.258752057
19	Chr. 8	q-q	STR-STR	D8S1132	D8S1179	0.23577	119.9623	136.4431	16.48085	0.159088296
20	Chr. 8	q-q	STR-SNP	D8S1132	rs2056277	0.17658	119.9623	156.441	36.47875	0.311402629
21	Chr. 8	q-q	STR-SNP	D8S1132	rs4606077	0.98885	119.9623	166.5671	46.60477	0.365784691
22	Chr. 11	p-p	STR-STR	D11S2368	TH01	0.20421	32.88891	4.48933	28.39958	0.256941387
23	Chr. 11	p-p	STR-SNP	D11S2368	rs1498553	0.01869	32.88891	11.57216	21.31675	0.201126911
24	Chr. 11	p-p	STR-SNP	D11S2368	rs901398	0.36799	32.88891	20.23465	12.65426	0.123908336

Table 7.4. continued.

25	Chr.13	q-q	STR-SNP	D13S325	rs1335873	0.62987	44.90825	2.118193	42.79006	0.347043992
26	Chr.13	q-q	STR-SNP	D13S325	rs1886510	0.96891	44.90825	4.798954	40.1093	0.332628523
27	Chr.13	q-q	STR-STR	D13S325	D13S317	0.97422	44.90825	79.83074	34.92249	0.301691441
28	Chr.13	q-q	STR-SNP	D13S325	rs1058083	0.57451	44.90825	94.11131	49.20306	0.377409289
29	Chr.13	q-q	STR-SNP	D13S325	rs354439	0.25147	44.90825	107.2948	62.38653	0.423823012
30	Chr.14	q-q	STR-SNP	D14S1434	rs1454361	0.30728	20.49462	17.19934	3.295282	0.032905192
31	Chr.14	q-q	STR-SNP	D14S1434	rs722290	0.18768	20.49462	47.50283	27.00821	0.24655614
32	Chr.14	q-q	STR-SNP	D14S1434	rs873196	0.46521	20.49462	104.0042	83.5096	0.465788531
33	Chr.14	q-q	STR-SNP	D14S1434	rs4530059	0.56093	20.49462	114.5175	94.02284	0.477266361
34	Chr.15	q-q	STR-SNP	D15S659	rs1821380	0.09532	49.51748	53.23968	3.722197	0.037153362
35	Chr.15	q-q	STR-SNP	D15S659	rs8037429	0.50019	49.51748	64.45011	14.93263	0.145039546
36	Chr.15	q-q	STR-SNP	D15S659	rs1528460	0.71734	49.51748	66.37152	16.85404	0.16243442
37	Chr.15	q-q	STR-STR	D15S659	PentaE	0.99899	49.51748	124.0505	74.53306	0.451723167
38	Chr. 18	q-q	STR-SNP	D18S1364	rs1736442	0.43181	91.21746	74.55715	16.66031	0.160699335
39	Chr. 18	q-q	STR-SNP	D18S1364	rs1024116	0.4926	91.21746	112.7889	21.57147	0.203257577
40	Chr.21	q-q	STR-SNP	D21S2055	rs722098	0.20135	49.46478	4.539526	44.92525	0.357784594
41	Chr.21	q-q	STR-STR	D21S2055	D21S11	1	49.46478	14.64555	34.81923	0.301033962
42	Chr.21	q-q	STR-SNP	D21S2055	rs2830795	0.71641	49.46478	27.34826	22.11652	0.20778713
43	Chr.21	q-q	STR-SNP	D21S2055	rs2831700	0.96164	49.46478	29.39708	20.0677	0.19055345
44	Chr.21	q-q	STR-SNP	D21S2055	rs914165	0.97298	49.46478	50.55435	1.089568	0.010893956
45	Chr.21	q-q	STR-SNP	D21S2055	rs221956	0.31349	49.46478	54.76922	5.304436	0.05284625
46	Chr.22	q-q	STR-SNP	D22GATA198B05	rs733164	0.90886	7.39585	31.36631	23.97046	0.222885135
47	Chr.22	q-q	STR-SNP	D22GATA198B05	rs987640	0.87136	7.39585	37.65417	30.25832	0.270357836
48	Chr.22	q-q	STR-STR	D22GATA198B05	D22S1045	0.98893	7.39585	46.21362	38.81777	0.325304911
49	Chr.22	q-q	STR-SNP	D22GATA198B05	rs2040411	0.49551	7.39585	62.88724	55.49139	0.402000746
50	Chr.22	q-q	STR-SNP	D22GATA198B05	rs1028528	0.1794	7.39585	64.13652	56.74067	0.40633012

None of the pairs showed LD allowing insignificant impact when using those pairs with RF values ~ 0.12 for most pedigrees (Gill *et al.* 2012). In addition, the RFs of the 220 syntenic pairs were estimated and showed that 49 pairs had RF values < 0.12 , four of which are STR-STR pairs, 22 STR-SNP pairs and 23 SNP-SNP pairs (Table 7.2, Table 7.3 and Table 7.4). Three out of the 49 pairs: vWA-D12S391 (0.117190251), D19S433-rs576261 (0.118793304) and rs1294331-rs10495407 (0.116048705) will have insignificant effect for most pedigrees as they had almost 0.12 RFs, while the rest (46 pairs) are expected to have a considerable effect on the LRs calculation. The effect within the 46 pairs will be varied, which was found to be influenced by the type of the pair (e.g. STR-STR or STR-SNP) and by the distance between the pairs (closer pairs have larger impact) (Tillmar and Phillips 2017). The effect STR-STR pairs was found to be the largest on LRs than STR-SNP pairs, which shows larger effect than SNP-SNP pairs, due to the increased level of heterozygosity of STRs (Tillmar and Phillips 2017). Therefore, it is expected to have significant impact from D6S1043-SE33 (RF= 0.044056152) D18S51-D18S1364 (RF= 0.065400163) and PentaD-D21S2055 (RF= 0.097833282) pairs more than other pairs due to the type of the pairs (STR-STR) and the close distance.

In real cases, the RF values estimated in this study can be used in FamLink software v.1.16 (Kling *et al.* 2012) to calculate LRs for two assumptions: ignoring linkage LR (unlinked) and considering linkage LR (linked). However, this version is limited in the number of pairs that can be run (i.e. the LR can be calculated and simulation can only be done for only one pair (2 markers) each run) (Kling *et al.* 2012). A new version v.2.1 (Beta) is being developed that will be able to handle any number of linked markers (see http://www.famlink.se/f_download.html).

Despite that this study was carried out under the assumption that no linkage between tested markers, a precise impact of linkage applicable for all possible scenarios cannot be achieved as it highly influenced by the case scenario itself (case-specific impact) (Tillmar and Phillips 2017). Here, the case scenario includes the type of relationship, the available members for testing, and the DNA profile of tested members, where the LRs are influenced by the amount of shared DNA (IBD) components between tested individuals that cannot be predictable.

7.5.5 Defining thresholds for kinship testing in Saudi Arabia

Although the Supreme Council of Magistracy of Saudi Arabia is the responsible authority of defining and enacting a specific LR threshold for kinship testing, this study can be used as a guide as it has defined the TP and FP that can be achieved at different thresholds for each relationship using different marker sets. Balance between the sensitivity (TP) and specificity (true negative (TN)) must be taken into consideration when defining the LR threshold (O'Connor *et al.* 2010) and uncertainty should be expected in some cases. It is also possible to use the grey zone approach (Giroti *et al.* 2007) rather than using a specific LR threshold, where an upper and a lower LR limits (LR rang) are defined as a grey zone for each type of relationship and LRs fall within this zone cannot eliminate uncertainty.

7.5.6 Defining the number of tested markers for each relationship

This study can also be used as a guide for genetic laboratories in Saudi Arabia regarding the number/type of markers that would allow sufficient differentiation between tested hypotheses. This study suggests 21 aSTRs (e.g. that included in the GlobalFiler kit) as a minimum number of markers for parent-child testing either trio or due cases, which would allow $\geq 99\%$ TP up to LR threshold of 1000 with 0% FP and 100%

TP up to LR threshold of 100,000 with 0% FP respectively. However, when the alleged fathers/mothers are relatives or when a mismatch was suspected to be a mutation, supplementary STR kits (e.g. SureID 23 kit) can be used to improve the certainty of the test. In more complex cases (e.g. when two or three mismatches were suspected), using ForenSeq DNA Signature Prep kit would allow much better resolution due to the inclusion of 94 iiSNPs and the lineage markers.

For the relationships of full-siblings, half-siblings and grand parent/child, using autosomal markers included the ForenSeq DNA Signature Prep kit alone would allow $\geq 97\%$ TP and $\leq 3.8\%$ FP at LR of 1, where lineage markers included in the kit can also improve these figures.

For more complex relationships like first-cousin, where even the 136 autosomal markers would allow the lowest TP and the highest FP comparing to other relationships, including as many as possible of relatives in the test would significantly improve the certainty. This has been shown when a grand-parent was added to the simulation that improved the TP to 99.4% and the FP to 0.2% (LR 1).

7.6 Conclusion

The performance of 136 autosomal DNA markers in kinship testing was assessed using seven different combinations of markers included in Identifiler Plus (currently used kit in Saudi Arabia), GlobalFiler, GlobalFiler and SureID 23, Fusion 6C and SureID 23, ForenSeq DNA Signature Prep kit (27 aSTRs and 94 iiSNPs), all markers (42 aSTRs and 94 iiSNPs (136 loci)) and 94 iiSNPs alone. Five types of relationships parent-child (duo pedigree), full-siblings (3 scenarios), half-siblings, first-cousins (2 scenarios) and grand-parent or grand-child were simulated under the assumption of no linkage between all markers.

The impact of testing additional markers was evaluated for all relationships tested that was found highly influenced by the relationship types. In addition, including more relatives had significant impact that was more than using more loci. It has shown that using 21 aSTRs as a minimum number of markers for parent-child relationship would provide confidence in most cases, but more supplementary aSTRs markers may be needed in some cases. The ForenSeq DNA Signature Prep kit showed the highest percentage of confidence due to number and the type of markers included in the system.

Potential TP and FP for each type of relationship using different marker sets can be used as a guide for the Supreme Council of Magistracy in Saudi Arabia in defining the LR threshold or a grey zone area for kinship testing in Saudi Arabia, and as a guide for the genetic laboratories in Saudi Arabia regarding the number/type of markers that would allow sufficient differentiation between tested hypotheses.

The genetic location of 95 markers in cM were estimated (41 markers were already published) using on the high-density multi-point SNP data of HapMap and RFs between syntenic markers located on the same arm were calculated using Kosambi function. The study highlighted 46 closely located syntenic pairs (3 STR-STR pairs, 21 STR-SNP pairs and 22 SNP-SNP) that would have significant impact ($RFs < 12$) on the LR estimation when using the 136 markers. The RFs values estimated here can be used to calculate the case specific LRs and to measure the case-specific impact of linkage.

With the increasing number of DNA markers that can be typed simultaneously, the need for a software that can calculate case-specific LR and includes the linkage effect for all linked pairs has become critical.

8 Chapter Eight: General Conclusion

The aims of the project were to evaluate a total of 42 aSTRs and 94 iiSNPs, which were generated using three commercially available kits, for kinship testing using samples from the population of Saudi Arabia. Five-hundred samples from unrelated individuals from the population of Saudi Arabia were collected after obtaining the ethical approvals from the SFHP (Saudi Arabia) and from UCLan Ethics Committee (STEMH 557).

Two typing systems (CE and MPS) were used in the project. GlobalFiler™ kit (AB) and SureID®23comp (Health Gene Technologies) were used to obtain the data of 38 aSTRs using the 500 samples. ForenSeq™ DNA Signature Prep Kit (Verogen) was used to obtain size and sequence-based data for 122 autosomal markers using 87 samples. The project allowed, in total, obtaining size-based data for 136 autosomal markers (42 aSTRs and 94 iiSNPs) and sequence-based data for 122 autosomal markers (28 aSTRs including SE33 and 94 iiSNPs).

The three kits were evaluated for the population of Saudi Arabia. For the GlobalFiler kit, as expected, the data of the 21 aSTRs included in the kit showed much higher CMP ($1.42E-26$) (Alsafiah *et al.* 2017) than the currently used kit in Saudi Arabia $2.23E-18$ (Identifiler plus kit). In addition, using the kit would improve the combined typical paternity index by 300-fold demonstrating the usefulness of adapting this kit in the forensic genetic laboratories of Saudi Arabia.

SureID®23comp kit is a supplementary STR kit that includes 22 aSTRs, 17 of which are non-CODIS STRs, developed for complex kinship testing. In this project, the kit has been evaluated following the minimum criteria for validation recommended by the ENFSI and by the SWGDAM (Alsafiah *et al.* 2019a). It was found that the kit met the criteria

commonly used in forensic genetics laboratories allowing the analysis of 17 non-CODIS loci that increases the number of aSTRs, when used in conjunction with any of commercially available kits, to 38-40 aSTRs. This would improve the resolution in kinship testing and thereby has the potential to increase the level of confidence in conclusions in kinship tests. The kit can benefit from some developments that were suggested by the evaluation study including adding as many common alleles found outside Chinese population were not included in the allelic ladder and increasing the concentration of the chemistry to allow more space for the DNA template. The data of the 17 non-CODIS STRs were approved by STRidER and were given a dataset reference number of STR000178.

ForenSeq™ DNA Signature Prep Kit was also used to sequence 87 samples on a MiSeq FGx instrument and the data were analysed using ForenSeq™ Universal Analysis Software (UAS) and STRait Razor v3.0 (SR). The system provided CMP of 1.97E-68 and 3.65E-77 for the size and sequence-based data respectively, where 1.24E-37 (size-based data) and 5.6E-41 (sequence-based data) were provided from the iiSNPs alone. Using the system allowed the data of additional four aSTRs (PentaE, PentaD, D6S1043, and D4S2408), which is not provided by CE-based kits used in the project, and of 94 iiSNPs. Analysing the data generated by the system and using SR also provided sequence-based Saudi population data for the most polymorphic well-characterised STR (SE33).

During the evaluation of the GlobalFiler kit, six allele variants were detected in the population of Saudi Arabia at the SE33 and D1S1656 that have not been characterised before. The SE33 variants were sequenced using Sanger sequencing while the D1S1656 variants were sequenced using the Verogen system. The sequence data of the six alleles were reported to STRBase to be included to the database. In addition, when the

ForenSeq™ DNA Signature Prep Kit was used, 13 novel sequences and 14 variants at the flanking region, which were not highlighted in the Flanking Region Report, were reported. The sequence-based data of the SE33 loci provided two novel motif patterns (D4 & D5) and seven novel sequences (Alsafiah *et al.* 2019b).

8.1 Human identification application in Saudi Arabia

Although one of the project aims was to evaluate the GlobalFiler kit, the project provided the data of 42 aSTRs that enabled expanding the evaluation to additional three commercially available CE-aSTR kits like PowerPlex Fusion 6C (Promega), VeriFiler Plus (AB) and Investigator 24plex (Qiagen) (Table 1.1).

As expected, PowerPlex Fusion 6C system showed the lowest CMP ($1.03E-29$) can be obtained from the kits that also includes two rapidly mutating STRs DYS570 and DYS576 that are useful for human identification application (Table 8.1). However, the other three kits (GlobalFiler (AB), VeriFiler Plus (AB) and Investigator 24plex (Qiagen)) benefit from including the Y-indel especially in determining male minor contribution in sexual assault cases. In addition, VeriFiler Plus ($9.26E-29$) and Investigator 24plex ($1.41E-26$) have an advantage of the presence of an internal quality control marker that enables differentiation between degraded samples and samples with inhibitors and thus allows more information about the sample's quality to decide further processing or not. Therefore, adopting either VeriFiler™ Plus (AB) or PowerPlex Fusion 6C system (Promega Corporation) as a standard analysis kit is highly encouraged for the Saudi laboratories which will provide much lower CMP using the same infrastructure and with almost the same cost. The flexibility of the Laboratory Information Management Systems (LIMS) used in Saudi Arabia will facilitate adopting any of the systems.

Table 8.1. The order of 42 aSTRs studied in this project based on their MP. The table also shows the CMP that can be obtained when using any of the latest four developed CE-based aSTR kits.

order	aSTRs	Matching Probability (MP)
1	SE33	0.007
2	D21S2055	0.016
3	PentaE	0.017
4	D12S391	0.026
5	D7S3048	0.027
6	D1S1656	0.030
7	D19S433	0.030
8	D18S51	0.031
9	FGA	0.033
10	D2S1338	0.035
11	D8S1132	0.041
12	PentaD	0.043
13	D22GATA198B05	0.045
14	D18S1364	0.046
15	D15S659	0.046
16	D8S1179	0.051
17	D21S11	0.055
18	D3S1744	0.060
19	D6S1043	0.063
20	D11S2368	0.068
21	D13S325	0.070
22	D4S2366	0.076
23	D7S820	0.076
24	D5S2800	0.079
25	vWA	0.082
26	D19S253	0.083
27	TH01	0.085
28	D16S539	0.087
29	D13S317	0.087
30	D3S1358	0.091
31	D2S441	0.091
32	D10S1248	0.098
33	D5S818	0.098
34	D6S474	0.102
35	D4S2408	0.112
36	CSF1PO	0.122
37	D22S1045	0.138
38	D9S1122	0.141
39	D20S482	0.143
40	D14S1434	0.148
41	TPOX	0.160
42	D17S1301	0.162
GlobalFiler		1.41E-26
Fusion 6C system		1.03E-29
VeriFiler Plus		9.26E-29
Investigator 24plex		1.41E-26

Allele frequencies of the aSTRs provided by the project can be used in estimating DNA profiles frequencies when using any of the commercially available kits. In addition, all autosomal markers were evaluated for forensic statistical parameters that can guide decision makers, in the forensic genetic laboratories, in creating population-specific aSTRs panel, for example, replacing some of the lowest informative loci with more

informative loci taking in account having the majority of loci shared with CODIS, ESS, and UK panel to allow sharing information between countries.

8.2 Kinship testing in Saudi Arabia

The data of 136 autosomal markers were also evaluated for kinship testing. The evaluation was carried out for seven different marker combinations that included in Identifiler Plus (15 aSTRs), GlobalFiler (21 aSTRs), GlobalFiler and SureID23 (38 aSTRs), Fusion 6C and SureID23 (40 aSTRs), ForenSeq DNA Signature Prep kit (27 aSTRs and 94 iiSNPs (121 loci), all markers (42 aSTRs and 94 iiSNPs (136 loci)) and 94 iiSNPs alone. Five types of relationships: parent-child (duo pedigree), full-siblings, half-siblings, first-cousins and grand-parent or grand-child, were simulated and the TP and FP were estimated at different LRs. Additional scenarios were included in the simulation study of full-siblings (3 scenarios) and first-cousin (2 scenarios) relationships to study the impact of testing more relatives for the same relationship.

The results supported previous work and showed that using 15 aSTRs will give sufficient confidence in most parents-child relationship cases (trio pedigrees). However, as recommended in the previous section (Section 8.1), if any of latest four developed CE-based aSTR kits (21-23 aSTRs) was adopted in Saudi Arabia as a standard kit, this would allow sufficient confidence for both types of pedigrees (trio and duo) in most cases. Supplementary STR kit may be used in more complex cases (e.g. when the alleged fathers or mothers are close relatives or when the 21-23 aSTRs showed inconclusive evidence). The SureID[®]23comp kit allows 38 or 40 aSTRs in conjunction with GlobalFiler or Fusion 6C respectively, providing 100% TP and 0% with LR thresholds up to 100,000 for parents-child (duo pedigree).

However, when two or three mismatches were suspected to be mutations or when testing distant relationships (e.g. full-siblings, half-siblings and grand parent/child or first-cousins), using ForenSeq DNA Signature Prep kit would allow much better resolution, due to the number (152 markers) of and the type (aSTRs, iiSNPs and lineage markers) of included markers, than CE systems analyse. In addition, this study supports previous work concluded that including more relatives to the test would significantly increase the resolution of kinship testing more than testing more DNA markers.

The study can be used by the Supreme Council of Magistracy of Saudi Arabia to define a specific threshold or a grey zone for kinship testing and by the genetic laboratories in Saudi Arabia to define the appropriate number of tested markers.

Using additional markers would increase the number markers located in the same chromosome (syntenic markers) and thus potential linkage should be considered in kinship testing. The RFs of a total of 220 syntenic pairs located at the same arm were estimated using the high-density multi-point SNP data of HapMap. As no LD was detected with the data set of the Saudi population, syntenic pairs with of ~ 0.12 will have almost zero effect for most pedigrees. Thus, the project has highlighted 46 syntenic pairs (3 STR-STR, 21 STR-SNP and 22 SNP-SNP) that would have significant impact on LR estimation due to lower RFs (< 0.12). The case-specific impact of linkage should be included in the estimation of LRs by using the RFs values estimated in this project. With the increasing number of DNA markers that can be typed simultaneously, the need to a software that can calculate case-specific LR and includes the linkage effect for all linked pairs has become critical and it is expected to be available in the near future.

8.3 Evidence of consanguinity in the population of Saudi Arabia

Previous studies, in the population of Saudi Arabia, which were conducted by either questionnaires (Wong and Anokute 1990, El-Hazmi *et al.* 1995) or by genetic analysis of forensically relevant markers (Khubrani *et al.* 2019b, Khubrani *et al.* 2019a), have demonstrated an increasing level of consanguinity.

In this project, lack of heterozygosity was observed in the size-based data of the majority of loci tested (20/21 loci in the GlobalFiler kit, 14/17 of the non- CODIS loci in the SureID23 kit, and 87/121 loci ForenSeq DNA Signature Prep kit(size-based data). In addition, the sequence-based data generated of the 122 markers (including the SE33 data) generated using ForenSeq DNA Signature Prep kit showed lack of heterozygosity in 92/122 loci. This was evidential by an increasing level of the inbreeding coefficient (F_{IS}) of 0.03560 (GlobalFiler kit), 0.02977 (SureID23 kit) and of 0.03924 (ForenSeq DNA Signature Prep kit).

Higher F_{IS} was also observed in the Middle Eastern samples included in the Human Genome Diversity Panel (HGDP-CEPH) that showed an averages of 0.041 (Bedouin) 0.032 (Druze) 0.014 (Mozabite) and 0.020 (Palestinian) (Leutenegger *et al.* 2011). These levels (including the Saudi data set examined here) are higher than other populations included in the HGDP-CEPH like Africans (7 populations with an average of 0.0032), Europeans (5 populations with an average of 0.003), East Asians (17 populations with an average of 0.0032) and Oceanians (2 populations with an average of 0.0025) (Leutenegger *et al.* 2011).

The data of the 21 aSTRs (GlobalFiler kit) showed an F_{IS} of 0.03560 that was increased to 0.03924 when more loci were tested (ForenSeq DNA Signature Prep kit). The higher inbreeding coefficient in the population of Saudi Arabia supports the need of expanding

of STR panel used in Saudi Arabia especially when relative are expected to be involved. In addition, less certainty would be expected in kinship testing comparing to other population with lower level of consanguinity.

8.4 Future work

As the hypothetical pedigree generated by the in-house Excel sheet does not reflect the type of samples that could be seen in real casework (F_{IS} of the 13 members was - 0.06343 that represents an excess of heterozygosity). It would be interesting to study the impact of higher inbreeding coefficient in real cases. Another option is by studying one of the large families in Saudi Arabia that are known to have a high level of consanguinity.

Sequencing more samples from the population of Saudi Arabia using MPS systems would allow establishing a representative sequence-based databased for both aSTRs and iiSNPs. It would be interesting to use the systems to study ancestry informative SNPs included in Primer Mix B.

9 Chapter Nine: References

AABB (2013) "Annual Report Summary",

(<http://www.aabb.org/sa/facilities/Documents/2013-relationship-testing-summary-report.pdf>).

AABB (2010a) "Annual Report Summary",

(<http://www.aabb.org/sa/facilities/Documents/rtannrpt10.pdf>).

AABB (2010b) Guidelines for Mass Fatality DNA Identification Operations.

(<http://www.aabb.org/programs/disasterresponse/Documents/aabbdnamassfatalityguidelines.pdf>).

AABB (2008) "Annual Report Summary for Testing In 2008,

(<https://www.aabb.org/sa/facilities/Documents/rtannrpt08.pdf>).

Abedalthagafi, M. (2019) "Precision Medicine of Monogenic Disorders: Lessons Learned from the Saudi Human Genome", *Frontiers in Bioscience (Landmark Ed)*, 24, pp. 870-889.

Abu-Amero, K.; Gonzalez, A.; Larruga, J.; Bosley, T. and Cabrera, V. (2007) "Eurasian and African Mitochondrial DNA Influences in the Saudi Arabian Population", *BMC Evolutionary Biology*, 7, pp. 32-46.

Abu-Amero, K.; Hellani, A.; Gonzalez, A.; Larruga, J.; Cabrera, V. and Underhill, P. (2009) "Saudi Arabian Y-Chromosome Diversity and its Relationship with Nearby Regions", *BMC Genetics*, 10, pp. 59-67.

Al-Enizi, M.; Ge, J.; Ismael, S.; Al-Enezi, H.; Al-Awadhi, A.; Al-Duaij, W.; Al-Saleh, B.; Ghulloom, Z. and Budowle, B. (2013) "Population Genetic Analyses of 15 STR Loci from Seven Forensically-Relevant Populations Residing in the State of Kuwait", *Forensic Science International: Genetics*, 7 (4), pp. e106-e107.

Allen, R.W. (2013) "Parentage Testing and Kinship Analysis - Encyclopaedia of forensic sciences" Second edition, Department of Forensic Sciences, Consolidated Forensic Laboratory, Washington, D.C., USA. edn, Elsevier, Amsterdam, pp. 287-294.

Almalki, N.; Chow, H.Y.; Sharma, V.; Hart, K.; Siegel, D. and Wurmbach, E. (2017) "Systematic Assessment of the Performance of Illumina's MiSeq FGx™ Forensic genomics System", *Electrophoresis*, 38 (6), pp. 846-854.

Almohammed, E. and Hadi, S. (2019) "The Study of Novel Sequence Alleles for Qatari Population using ForenSeq™ DNA Kit", *Forensic Science International: Genetics Supplement Series*. (In press).

Almohammed, E.; Zgonjanin, D.; Iyengar, A.; Ballard, D.; Devesse, L. and Sibte, H. (2017) "A Study of Degraded Skeletal Samples using ForenSeq™ DNA Signature Kit", *Forensic Science International: Genetics Supplement Series*, 6, pp. e410-e412.

Alonso, A.; Barrio, P.A.; Muller, P.; Kocher, S.; Berger, B.; Martin, P.; Bodner, M.; Willuweit, S.; Parson, W.; Roewer, L. and Budowle, B. (2018) "Current State-of-Art of STR Sequencing in Forensic Genetics", *Electrophoresis*, 39 (21), pp. 2655-2668.

Alsafiah, H.; Aljanabi, A.; Hadi, S.; Alturayef, S. and Goodwin, W. (2019a) "An Evaluation of the SureID 23comp Human Identification Kit for Kinship Testing", *Scientific Reports*, 9 (1), pp. 16859.

Alsafiah, H.; Goodwin, W.; Hadi, S.; Alshaikhi, M. and Wepeba, P. (2017) "Population Genetic Data for 21 Autosomal STR Loci for the Saudi Arabian Population using the GlobalFiler® PCR Amplification Kit", *Forensic Science International: Genetics*, 31 (Supplement C), pp. e59-e61.

Alsafiah, H.; Iyengar, A.; Hadi, S.; Alshlash, W. and Goodwin, W. (2018) "Sequence Data of Six Unusual Alleles at SE33 and D1S1656 STR Loci", *Electrophoresis*, 39, pp. 2471-2476.

Alsafiah, H.; Khubrani, Y.; Sibte, H. and Goodwin, W. (2019b) "Sequence-Based Saudi Population Data for the SE33 Locus", *Forensic Science International: Genetics Supplement Series*. (In press).

Alshamali, F.; Alkhayat, A.Q.; Budowle, B. and Watson, N.D. (2005) "STR Population Diversity in Nine Ethnic Populations Living in Dubai", *Forensic Science International*, 152 (2-3), pp. 267-279.

Amorim, A. and Pereira, L. (2005) "Pros and Cons in the use of SNPs in Forensic Kinship Investigation: A Comparative Analysis with STRs", *Forensic Science International*, 150 (1), pp. 17-21.

Applied Biosystems (2017) "Precision ID GlobalFiler NGS STR Panels", (<https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/AB-PrecisionID-GlobalFiler-NGS-STR-Panel-Flyer.pdf>).

Applied Biosystems (2015) "AmpFISTR® NGM SElect™ PCR Amplification Kit", (file:///lha-031/pers-K/00078795/Downloads/cms_089008.pdf).

Applied Biosystems (2004) "AmpFISTR Blue™, Green™, and Profiler® PCR Amplification Kits", (https://tools.thermofisher.com/content/sfs/brochures/cms_040284.pdf).

Applied Biosystems (2012) "AmpFISTR® Sinofiler™ PCR Amplification Kit, User Guide" (<Http://Tools.Thermofisher.Com/Content/Sfs/Manuals/4384256E.Pdf>).

Auton, A.; Abecasis, G.R.; Altshuler, D.M. *et al.* (2015) "A Global Reference for Human Genetic Variation", *Nature*, 526 (7571), pp. 68-74.

Bentley, D.R.; Balasubramanian, S.; Swerdlow, H.P. *et al.* (2008) "Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry", *Nature*, 456 (7218), pp. 53-59.

Bodner, M.; Bastisch, I.; Butler, J.M.; Fimmers, R.; Gill, P.; Gusmão, L.; Morling, N.; Phillips, C.; Prinz, M.; Schneider, P.M. and Parson, W. (2016) "Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on Quality Control of Autosomal Short Tandem Repeat Allele Frequency Databasing (STRidER)", *Forensic Science International: Genetics*, 24, pp. 97-102.

Børsting, C.; Rockenbauer, E. and Morling, N. (2009) "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*, 4 (1), pp. 34-42.

Børsting, C.; Tomas, C. and Morling, N. (2012) "Typing of 49 Autosomal SNPs by Single Base Extension and Capillary Electrophoresis for Forensic Genetic Testing", *Methods in Molecular Biology*, 830, pp. 87-107.

Børsting, C.; Mogensen, H.S. and Morling, N. (2013) "Forensic Genetic SNP Typing of Low-Template DNA and Highly Degraded DNA from Crime Case Samples", *Forensic Science International: Genetics*, 7 (3), pp. 345-352.

Børsting, C. and Morling, N. (2015) "Next Generation Sequencing and its Applications in Forensic Genetics", *Forensic Science International: Genetics*, 18, pp. 78-89.

Borsuk, L.A.; Gettings, K.B.; Steffen, C.R.; Kiesler, K.M. and Vallone, P.M. (2018) "Sequence-Based US Population Data for the SE33 Locus", *Electrophoresis*, 39, 2694–2701.

Brenner, C.H. (2018) "Forensic Mathematics- Mutations in Paternity" ([Http://Dna-View.Com/Mudisc.Htm](http://Dna-View.Com/Mudisc.Htm)).

Bright, J.; Hopwood, A.; Curran, J. and Buckleton, J. (2014) "A Guide to Forensic DNA Interpretation and Linkage". <https://www.promega.co.uk/resources/profiles-in-dna/2014/a-guide-to-forensic-dna-interpretation-and-linkage/> Updated 2014.

Brooker, R.J. (2012) "Genetics: analysis & principles", 4th edn, McGraw-Hill Education, New York.

Brookes, C.; Bright, J.A.; Harbison, S. and Buckleton, J. (2012) "Characterising Stutter in Forensic STR Multiplexes", *Forensic Science International: Genetics*, 6 (1), pp. 58-63.

Buckleton, J. and Triggs, C. (2006) "The Effect of Linkage on the Calculation of DNA Match Probabilities for Siblings and Half Siblings", *Forensic Science International*, 160 (2-3), pp. 193-199.

Budowle, B.; Moretti, T.R.; Niezgoda, S.J. and Brown, B.L. (1998) "CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools", <https://www.promega.com/~media/files/resources/conference%20proceedings/ish%2002/oral%20presentations/17.pdf> (Promega).

- Budowle, B.; Ge, J.; Chakraborty, R.; Eisenberg, A.J.; Green, R.; Mulero, J.; Lagace, R. and Hennessy, L. (2011) "Population Genetic Analyses of the NGM STR Loci", *International Journal of Legal Medicine*, 125 (1), pp. 101-109.
- Butler, J. (2015) "Advanced Topics in Forensic DNA Typing: Interpretation", Academic Press, San Diego.
- Canturk, K.M.; Emre, R.; Gurkan, C.; Komur, I.; Muslumanoglu, O. and Dogan, M. (2016) "An Incest Case with Three Biological Brothers as Alleged Fathers: Even 22 Autosomal STR Loci Analysis would Not Suffice without the Mother", *Medicine, Science and the Law*, 56 (3), pp. 210-212.
- Carboni, I.; Iozzi, S.; Nutini, A.L.; Torricelli, F. and Ricci, U. (2014) "Improving Complex Kinship Analyses with Additional STR Loci", *Electrophoresis*, 35 (21-22), pp. 3145-3151.
- Carothers, A.D. and Wright, A.F. (1992) "The Effect of Mutation on Linkage Disequilibrium", *Annals of Human Genetics*, 56 (2), pp. 155-158.
- Chakraborty, R. (1992) "Sample Size Requirements for Addressing the Population Genetic Issues of Forensic use of DNA Typing", *Human Biology*, 64 (2), pp. 141-159.
- Chen, F.; Dong, M.; Ge, M.; Zhu, L.; Ren, L.; Liu, G. and Mu, R. (2013) "The History and Advances of Reversible Terminators used in New Generations of Sequencing Technology", *Genomics Proteomics Bioinformatics*, 11 (1), pp. 34-40.
- Churchill, J.; Novroski, N.; King, J.; Seah, L. and Budowle, B. (2017) "Population and Performance Analyses of Four Major Populations with Illumina's FGx Forensic Genomics System", *Forensic Science International: Genetics*, 30, pp. 81-92.
- Churchill, J.; Schmedes, S.; King, J. and Budowle, B. (2016) "Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in Genetic Profiling", *Forensic Science International: Genetics*, 20, pp. 20-29.
- Collins, P.J.; Hennessy, L.K.; Leibel, C.S.; Roby, R.K.; Reeder, D.J. and Foxall, P.A. (2004) "Developmental Validation of a Single-Tube Amplification of the 13 CODIS STR Loci, D2S1338, D19S433, and Amelogenin: The AmpFISTR® Identifiler® PCR Amplification Kit", *Journal of Forensic Sciences*, 49 (6), pp. 1265-1277.

Cotton, E.A.; Allsop, R.F.; Guest, J.L.; Frazier, R.R.; Koumi, E.P.; Callow, I.P.; Seager, A. and Sparkes, R.L. (2000) "Validation of the AMPFISTR® SGM Plus™ System for use in Forensic Casework", *Forensic Science International*, 112 , pp. 151–161.

De La Puente, M.; Santos, C.; Fondevila, M.; Manzo, L.; Carracedo, Á; Lareu, M.V. and Phillips, C. (2016) "The Global AIMs Nano Set: A 31-Plex SNaPshot Assay of Ancestry-Informative SNPs", *Forensic Science International: Genetics*, 22, pp. 81-88.

Devesse, L.; Ballard, D.; Davenport, L.; Riethorst, I.; Mason-Buck, G. and Syndercombe Court, D. (2018) "Concordance of the ForenSeq™ System and Characterisation of Sequence-Specific Autosomal STR Alleles Across Two Major Population Groups", *Forensic Science International: Genetics*, 34, pp. 57-61.

DNA Resources-Forensic and Policy (2019) "Global Resources, Global DNA Databases" ([Http://Www.Dnaresource.Com/Resources.Html](http://www.dnaresource.com/resources.html)).

Dogan, M.; Kara, U.; Emre, R.; Fung, W.K. and Canturk, K.M. (2015) "Two Brothers' Alleged Paternity for a Child: Who is the Father?", *Molecular Biology Reports*, 42 (6), pp. 1025-1027.

Edge, M.; Algee-Hewitt, B.; Pemberton, T.; Li, J. and Rosenberg, N. (2017) "Linkage Disequilibrium Matches Forensic Genetic Records to Disjoint Genomic Marker Sets", *Proceedings of the National Academy of Sciences of the United States of America*, 114 (22), pp. 5671- 5676

El-Hazmi, M.; Al-Swailem, A, Warsy, A.; Sulaimani, R. and Al-Meshari, A. (1995) "Consanguinity among the Saudi Arabian Population", *Journal of Medical Genetics*, 32(8), pp. 623-626.

El-Mouzan, M.I.; Al-Salloum, A.A.; Al-Herbish, A.S.; Qurachi, M.M. and Al-Omar, A.A. (2007) "Regional Variations in the Prevalence of Consanguinity in Saudi Arabia", *Saudi Medical Journal*, 28 (12), pp. 1881-1884.

ENFSI (2010) "Recommended Minimum Criteria for the Validation of various Aspects of the DNA Profiling Process"([Http://Enfsi.Eu/Wp-Content/uploads/2016/09/Minimum_validation_guidelines_in_dna_profiling_-_v2010_0.Pdf](http://enfsi.eu/Wp-Content/uploads/2016/09/Minimum_validation_guidelines_in_dna_profiling_-_v2010_0.pdf)).

Ensenberger, M.G.; Lenz, K.A.; Matthies, L.K.; Hadinoto, G.M.; Schienman, J.E.; Przech, A.J.; Morganti, M.W.; Renstrom, D.T.; Baker, V.M.; Gawrys, K.M.; Hoogendoorn, M.; Steffen, C.R.; Martiñan, P.; Alonso, A.; Olson, H.R.; Sprecher, C.J. and Storts, D.R. (2016) "Developmental Validation of the PowerPlex® Fusion 6C System", *Forensic Science International: Genetics*, 21, pp. 134-144.

Excoffier, L.; Laval, G. and Schneider, S. (2007) "Arlequin (Version 3.0): An Integrated Software Package for Population Genetics Data Analysis", *Evolutionary bioinformatics online*, 1, pp. 47-50.

Faith, S. and Scheible, M. (2016) "Analyzing Data from Next Generation Sequencers using the PowerSeq® Auto/Mito/Y System".

[\(Http://Www.Promega.Co.Uk/Resources/Profiles-in-Dna/2016/Analyzing-Data-from-Next-Generation-Sequencers-using-the-Powerseq-Automitoy-System/\)](http://www.promega.co.uk/resources/profiles-in-dna/2016/analyzing-data-from-next-generation-sequencers-using-the-powerseq-automitoy-system/)

Federal Bureau of Investigation (accessed 2017), CODIS - NDIS Statistics, [\(<https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics>\)](https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics) (last updated August, 2016).

Fondevila, M.; Phillips, C.; Santos, C.; Freire Aradas, A.; Vallone, P.M.; Butler, J.M.; Lareu, M.V. and Carracedo, Á (2013) "Revision of the SNPforID 34-Plex Forensic Ancestry Test: Assay Enhancements, Standard Reference Sample Genotypes and Extended Population Studies", *Forensic Science International: Genetics*, 7 (1), pp. 63-74.

Fu, X.; Sun, S.; Liu, Y.; He, J.; Cai, J. and Lagabayila, Z. (2018) "The Validation of Goldeneye™ DNA ID 22NC Kit and the Genetic Polymorphism of 21 Short Tandem Repeat Loci in the Chinese Hunan Han Population", *Journal of Forensic Sciences Research*, 4 (3), pp. 122-128.

Gao, Z.; Chen, X.; Zhao, Y.; Zhao, X.; Zhang, S.; Yang, Y.; Wang, Y. and Zhang, J. (2018) "Forensic Genetic Informativeness of an SNP Panel Consisting of 19 Multi-Allelic SNPs", *Forensic Science International: Genetics*, 34, pp. 49-56.

Ge, J.; Sun, H.; Li, H.; Liu, C.; Yan, J. and Budowle, B. (2014) "Future Directions of Forensic DNA Databases", *Croatian Medical Journal*, 55 (2), pp. 163-166.

Gelardi, C.; Rockenbauer, E.; Dalsgaard, S.; Børsting, C. and Morling, N. (2014) "Second Generation Sequencing of Three STRs D3S1358, D12S391 and D21S11 in Danes and a New Nomenclature for Sequenced STR Alleles", *Forensic Science International: Genetics*, 12, pp. 38-41.

General Authority for Statistics in Saudi Arabia (2016) "Demographic Research Bulletin" (https://www.stats.gov.sa/sites/default/files/en-demographic-research-2016_4.pdf).

Gettings, K.; Aponte, R.; Vallone, P. and Butler, J. (2015) "STR Allele Sequence Variation: Current Knowledge and Future Issues", *Forensic Science International: Genetics*, 18, pp. 118-130.

Gettings, K.; Borsuk, L.; Ballard, D.; Budowle, B.; Devesse, L.; King, J.; Parson, W.; Phillips, C. and Vallone, P. (2017) "STRSeq: A Catalog of Sequence Diversity at Human Identification Short Tandem Repeat Loci", *Forensic Science International: Genetics*, 31, pp. 111-117.

Gettings, K.; Borsuk, L.; Steffen, C.; Kiesler, K. and Vallone, P. (2018) "Sequence-Based U.S. Population Data for 27 Autosomal STR Loci", *Forensic Science International: Genetics*, 37, pp. 106-115.

Gettings, K.; Kiesler, K.; Faith, S.; Montano, E.; Baker, C.; Young, B.; Guerrieri, R. and Vallone, P. (2016) "Sequence Variation of 22 Autosomal STR Loci Detected by Next Generation Sequencing", *Forensic Science International: Genetics*, 21, pp. 15-21.

Gettings, K.; Lai, R.; Johnson, J.; Peck, M.; Hart, J.; Gordish-Dressman, H.; Schanfield, M. and Podini, D. (2014) "A 50-SNP Assay for Biogeographic Ancestry and Phenotype Prediction in the U.S. Population", *Forensic Science International: Genetics*, 8 (1), pp. 101-108.

Gill, P.; Kimpton, C.; D'Aloja, E.; Andersen, J.; Bar, W.; Brinkmann, B.; Holgersson, S.; Johnsson, V.; Kloosterman, A. and Lareu, M. (1994) "Report of the European DNA Profiling Group (EDNAP)--Towards Standardisation of Short Tandem Repeat (STR) Loci", *Forensic Science International*, 65 (1), pp. 51-59.

Gill, P.; Phillips, C.; McGovern, C.; Bright, J. and Buckleton, J. (2012) "An Evaluation of Potential Allelic Association between the STRs vWA and D12S391: Implications in Criminal Casework and Applications to Short Pedigrees", *Forensic Science International: Genetics*, 6 (4), pp. 477-486.

Gill, P.; Fereday, L.; Morling, N. and Schneider, P. (2006) "New Multiplexes for Europe- Amendments and Clarification of Strategic Development", *Forensic Science International*, 163 (1), pp. 155-157.

Giroti, R.I.; Verma, S.; Singh, K.; Malik, R. and Talwar, I. (2007) "A Grey Zone Approach for Evaluation of 15 Short Tandem Repeat Loci in Sibship Analysis: A Pilot Study in Indian Subjects", *Journal of Forensic and Legal Medicine*, 14 (5), pp. 261-265.

Glaubitz, J. (2004) "CONVERT: A user Friendly Program to Reformat Diploid Genotypic Data for Commonly used Population Genetic Software Packages", *Molecular Ecology*, 4, pp. 309-310.

González-Andrade, F.; Sánchez, D.; Penacino, G. and Martínez-Jarreta, B. (2009) "Two Fathers for the Same Child: A Deficient Paternity Case of False Inclusion with Autosomic STRs", *Forensic Science International: Genetics*, 3 (2), pp. 138-140.

Goodwin, W. (2015) "DNA Profiling: The First 30 Years", *Science & Justice*, 55 (6), pp. 375-376.

Goodwin, W.; Ballard, D.; Simpson, K.; Thacker, C.; Syndercombe Court, D. and Gow, J. (2004) "Case Study: Paternity Testing-when 21 Loci are Not Enough", *International Congress Series*, 1261, pp. 460-462.

Greenspoon, S.A.; Lytle, P.J.; Turek, S.A.; Rolands, J.M.; Scarpetta, M.A. and Carr, C.D. (2000) "Validation of the PowerPlex 1.1™ Loci for use in Human Identification", *Journal of Forensic Sciences*, 45 (3), pp. 677-683.

Guo, F.; Yu, J.; Zhang, L. and Li, J. (2017) "Massively Parallel Sequencing of Forensic STRs and SNPs using the Illumina® ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System", *Forensic Science International: Genetics*, 31, pp. 135-148.

Gusmão, L.; Butler, J.M.; Linacre, A.; Parson, W.; Roewer, L.; Schneider, P.M. and Carracedo, A. (2017) "Revised Guidelines for the Publication of Genetic Population Data", *Forensic Science International: Genetics*, 30, pp. 160-163.

Hares, D.R. (2015) "Selection and Implementation of Expanded CODIS Core Loci in the United States", *Forensic Science International: Genetics*, 17, pp. 33-34.

Hering, S.; Nixdorf, R.; Edelmann, J.; Thiede, C. and Dreßler, J. (2006) "Further Sequence Data of Allelic Variants at the STR Locus ACTBP2 (SE33): Detection of a very Short Off Ladder Allele", *International Congress Series*, 1288 , pp. 810-812.

Hill, C.R.; Butler, J.M. and Vallone, P.M. (2009) "A 26plex Autosomal STR Assay to Aid Human Identity Testing*", *Journal of Forensic Sciences*, 54 (5), pp. 1008-1015.

Hill, C.R.; Kline, M.C.; Coble, M.D. and Butler, J.M. (2008) "Characterization of 26 miniSTR Loci for Improved Analysis of Degraded DNA Samples", *Journal of Forensic Sciences*, 53 (1), pp. 73-80.

Home office (2013), "Information fact sheet for the introduction of new DNA-17 profiling chemistries for the National DNA Database® (NDNAD) "
(<https://www.keyforensic.co.uk/docs/dna17-factsheet1.pdf>).

Huston, K. (1998) "Statistical Analysis of STR Data" (<https://www.Promega.Com/-/Media/Files/Resources/Profiles-in-Dna/103/Statistical-Analysis-of-Str-Data.Pdf?La=en>).

Ingwer, B. and Patrick J. (2005) "Modern Multidimensional Scaling Theory and Applications", Springer Series in Statistics (2nd Edition).

Interpol (2016) "Global DNA Profiling Survey Results"
([File:///C:/Users/Alsaf/Downloads/INTERPOL%20Global%20DNA%20Survey%20Results%202016%20\(Public%20Version\)%20\(2\).Pdf](File:///C:/Users/Alsaf/Downloads/INTERPOL%20Global%20DNA%20Survey%20Results%202016%20(Public%20Version)%20(2).Pdf))

Invitrogen (2019) "Qubit™ dsDNA HS Assay Kits User Guide, Pub. no. MAN0017455",
(https://Assets.Thermofisher.Com/TFS-Assets/LSG/Manuals/MAN0017455_Qubit_1X_dsDNA_HS_Assay_Kit_UG.Pdf).

Iyavoo, S.; Afolabi, O.; Boggi, B.; Bernotaite, A. and Haizel, T. (2019) "Population Genetics Data for 22 Autosomal STR Loci in European, South Asian and African Populations using SureID® 23comp Human DNA Identification Kit", *Forensic Science International*, 301, pp. 174-181.

Jäger, A.C.; Alvarez, M.L.; Davis, C.P.; Guzmán, E.; Han, Y.; Way, L.; Walichiewicz, P.; Silva, D.; Pham, N.; Caves, G.; Bruand, J.; Schlesinger, F.; Pond, S.J.K.; Varlaro, J.; Stephens, K.M. and Holt, C.L. (2017) "Developmental Validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories", *Forensic Science International: Genetics*, 28, pp. 52-70.

James, N. (2015) "DNA Testing in Criminal Justice: Background, Current Law, and Grants (Report)"(<https://www.statewatch.org/news/2015/feb/usa-crs-dna.pdf>).

James, N. (2012) "DNA Testing in Criminal Justice: Background, Current Law, Grants, and Issues (Report)"(<https://fas.org/sgp/crs/misc/R41800.pdf>).

James, M. and Curran, T. (2017) "DNAtools: Tools for Empirical Testing of DNA Match Probabilities"(<https://CRAN.R-Project.Org/package=DNAtools>).

Jeffreys, A.J. (2005) "Genetic Fingerprinting", *Nature Medicine*, 11 (10), pp. 1035-1039.

Jia, Y.S.; Zhang, L.; Qi, L.Y.; Mei, K.; Zhou, F.L.; Huang, D.X. and Yi, S.H. (2015) "Multistep Microsatellite Mutation Leading to Father-Child Mismatch of FGA Locus in a Case of Non-Exclusion Parentage", *Legal Medicine*, 17 (5), pp. 364-365.

Jin, B.; Su, Q.; Luo, H.; Li, Y.; Wu, J.; Yan, J.; Hou, Y.; Liang, W. and Zhang, L. (2016) "Mutational Analysis of 33 Autosomal Short Tandem Repeat (STR) Loci in Southwest Chinese Han Population Based on Trio Parentage Testing", *Forensic Science International: Genetics*, 23, pp. 86-90.

Jobling, M.A. and Gill, P. (2004) "Encoded Evidence: DNA in Forensic Analysis", *Nature Reviews Genetics*, 5 (10), pp. 739-751.

Jones, R.; Tayyare, W.; Tay, G.; Alsafar, H. and Goodwin, W. (2017) "Population Data for 21 Autosomal Short Tandem Repeat Markers in the Arabic Population of the United Arab Emirates", *Forensic Science International: Genetics*, 28, pp. e41-e42.

Junge, A.; Brinkmann, B.; Fimmers, R. and Madea, B. (2006) "Mutations or Exclusion: An Unusual Case in Paternity Testing", *International Journal of Legal Medicine*, 120 (6), pp. 360-363.

Khubrani, Y.; Wetton, J. and Jobling, M. (2018) "Extensive Geographical and Social Structure in the Paternal Lineages of Saudi Arabia Revealed by Analysis of 27 Y-STRs", *Forensic Science International: Genetics*, 33, pp. 98-105.

Khubrani, Y.; Wetton, J. and Jobling, M. (2019a) "Analysis of 21 Autosomal STRs in Saudi Arabia Reveals Population Structure and the Influence of Consanguinity", *Forensic Science International: Genetics*, 39, pp. 97-102.

Khubrani, Y.; Hallast P.; and Jobling, M. and Wetton, J. (2019b) "Massively Parallel Sequencing of Autosomal STRs and Identity-Informative SNPs Highlights Consanguinity in Saudi Arabia", *Forensic Science International: Genetics*, 43, pp. 102164.

Kidd, K.K.; Pakstis, A.J.; Speed, W.C.; Grigorenko, E.L.; Kajuna, S.L.B.; Karoma, N.J.; Kungulilo, S.; Kim, J.; Lu, R.; Odunsi, A.; Okonofua, F.; Parnas, J.; Schulz, L.O.; Zhukova, O.V. and Kidd, J.R. (2006) "Developing a SNP Panel for Forensic Identification of Individuals", *Forensic Science International*, 164 (1), pp. 20-32.

Kim, E.H.; Lee, H.Y.; Kwon, S.Y.; Lee, E.Y.; Yang, W.I. and Shin, K. (2017) "Sequence-Based Diversity of 23 Autosomal STR Loci in Koreans Investigated using an in-House Massively Parallel Sequencing Panel", *Forensic Science International: Genetics*, 30, pp. 134-140.

Kimpton, C.; Fisher, D.; Watson, S.; Adams, M.; Urquhart, A.; Lygo, J. and Gill, P. (1994) "Evaluation of an Automated DNA Profiling System Employing Multiplex Amplification of Four Tetrameric STR Loci", *International Journal of Legal Medicine*, 106 (6), pp. 302-311.

King, J.; Churchill, J.; Novroski, N.; Zeng, X.; Warshauer, D.; Seah, L. and Budowle, B. (2018) "Increasing the discrimination power of ancestry- and identity-informative SNP

loci within the ForenSeq™ DNA Signature Prep Kit", *Forensic Science International: Genetics*, 36, pp. 60-76.

King, T.; Parkin, E.; Swinfield, G.; Cruciani, F.; Scozzari, R.; Rosa, A.; Lim, S.; Xue, Y.; Tyler-Smith, C. and Jobling, M.A. (2007) "Africans in Yorkshire? the Deepest-Rooting Clade of the Y Phylogeny within an English Genealogy", *European Journal of Human Genetics*, 15 (3), pp. 288-293.

Kline, M.C.; Hill, C.R.; Decker, A.E. and Butler, J.M. (2011) "STR Sequence Analysis for Characterizing Normal, Variant, and Null Alleles", *Forensic Science International: Genetics*, 5 (4), pp. 329-332.

Kling, D.; Egeland, T. and Tillmar, A.O. (2012) "FamLink-A User Friendly Software for Linkage Calculations in Family Genetics", *Forensic Science International: Genetics*, 6 (5), pp. 616-620.

Kling, D.; Tillmar, A.O. and Egeland, T. (2014) "Familias 3-Extensions and New Functionality", *Forensic Science International: Genetics*, 13, pp. 121-127.

Köcher, S.; Müller, P.; Berger, B.; Bodner, M.; Parson, W.; Roewer, L. and Willuweit, S. (2018) "Inter-Laboratory Validation Study of the ForenSeq™ DNA Signature Prep Kit", *Forensic Science International: Genetics*, 36, pp. 77-85.

Krenke, B.E.; Tereba, A.; Anderson, S.J.; Buel, E.; Culhane, S.; Finis, C.J.; Tomsey, C.S.; Zachetti, J.M.; Masibay, A.; Rabbach, D.R.; Amiott, E.A. and Sprecher, C.J. (2002) "Validation of a 16-Locus Fluorescent Multiplex System", *Journal of Forensic Sciences*, 47 (4), pp. 773-785.

Krenke, B.E.; Viculis, L.; Richard, M.L.; Prinz, M.; Milne, S.C.; Ladd, C.; Gross, A.M.; Gornall, T.; Frappier, J.R.; Eisenberg, A.J.; Barna, C.; Aranda, X.G.; Adamowicz, M.S. and Budowle, B. (2005) "Validation of Male-Specific, 12-Locus Fluorescent Short Tandem Repeat (STR) Multiplex", *Forensic Science International*, 151 (1), pp. 111-124.

Kutyavin, I.V.; Afonina, I.A.; Mills, A.; Gorn, V.V.; Lukhtanov, E.A.; Belousov, E.S.; Singer, M.J.; Walburger, D.K.; Lokhov, S.G.; Gall, A.A.; Dempcy, R.; Reed, M.W.; Meyer, R.B. and Hedgpeth, J. (2000) "3'-Minor Groove Binder-DNA Probes Increase Sequence

Specificity at PCR Extension Temperatures", *Nucleic Acids Research*, 28 (2), pp. 655-661.

Lan, Q.; Wang, H.; Shen, C.; Guo, Y.; Yin, C.; Xie, T.; Fang, Y.; Zhou, Y. and Zhu, B. (2018) "Mutability Analysis Towards 21 STR Loci Included in the AGCU 21 + 1 Kit in Chinese Han Population", *International Journal of Legal Medicine*, 132 (5), pp. 1287-1291.

Lederer, T.; Braunschweiger, G.; Dunkelmann, B. and Betz, P. (2008) "Characterization of Two Unusual Allele Variants at the STR Locus ACTBP2 (SE33)", *Forensic Science & Pathology*, 4 (3), pp. 164-166.

Levedakou, E.N.; Freeman, D.A.; Budzynski, M.J.; Early, B.E.; Damaso, R.C.; Pollard, A.M.; Townley, A.J.; Gombos, J.L.; Lewis, J.L.; Kist, F.G.; Hockensmith, M.E.; Terwilliger, M.L.; Amriott, E.; McElfresh, K.C.; Schumm, J.W.; Ulery, S.R.; Konotop, F.; Sessa, T.L.; Sailus, J.S.; Crouse, C.A.; Tomsey, C.S.; Ban, J.D. and Nelson, M.S. (2002) "Characterization and Validation Studies of PowerPlex® 2.1, a Nine-Locus Short Tandem Repeat (STR) Multiplex System and Penta D Monoplex", *Journal of Forensic Sciences*, 47 (4), pp. 757-772.

Leutenegger, A.; Sahbatou, M.; Gazal, S.; Cann, H. and Emmanuelle, G. (2011) "Consanguinity Around the World: What do the Genomic Data of the HGDP-CEPH Diversity Panel Tell Us?", *European Journal of Human Genetics*, 19 (5), pp. 583-587.

Li, H.; Zhao, X.; Ma, K.; Cao, Y.; Zhou, H.; Ping, Y.; Shao, C.; Xie, J. and Liu, W. (2017) "Applying Massively Parallel Sequencing to Paternity Testing on the Ion Torrent Personal Genome Machine", *Forensic Science International: Genetics*, 31, pp. 155-159.

Li, J.; Luo, H.; Song, F.; Zhang, L.; Deng, C.; Yu, Z.; Gao, T.; Liao, M. and Hou, Y. (2017) "Validation of the Microreader™ 23sp ID System: A New STR 23-Plex System for Forensic Application", *Forensic Science International: Genetics*, 27, pp. 67-73.

Li, R.; Li, H.; Dan, P.; H.; Wang, Z.; Huang, E.; Wu, R. and Sun, H. (2019) "Improved Pairwise Kinship Analysis using Massively Parallel Sequencing", *Forensic Science International: Genetics*, 38, pp. 77-85.

- Li, L.; Ge, J.; Zhang, S.; Guo, J.; Zhao, S.; Li, C.; Tang, H.; Davis, C.; Budowle, B.; Hou, Y. and Liu, Y. (2012) "Maternity Exclusion with a very High Autosomal STRs Kinship Index", *International Journal of Legal Medicine*, 126 (4), pp. 645-648.
- Lin, S.W.; Li, C. and Ip, S.C.Y. (2017) "A Selection Guide for the New Generation 6-Dye DNA Profiling Systems", *Forensic Science International: Genetics*, 30, pp. 34-42.
- Liu, Q.; Luo, H.; Zhao, H.; Huang, X.; Cheng, J. and Lu, D. (2014) "Recombination Analysis of Autosomal Short Tandem Repeats in Chinese Han Families", *Electrophoresis*, 35 (6), pp. 883-887.
- Liu, Y.; Guo, L.; Jin, H.; Li, Z.; Bai, R.; Shi, M. and Ma, S. (2017) "Developmental Validation of a 6-Dye Typing System with 27 Loci and Application in Han Population of China", *Scientific Reports*, 7 (1), pp. 4706-017-04548-1.
- Lobo, I. and Shaw, K. (2008) "Thomas Hunt Morgan, Genetic Recombination, and Gene Mapping. Nature Education 1(1):205.
(<https://www.nature.com/scitable/topicpage/thomas-hunt-morgan-genetic-recombination-and-gene-496>).
- Londin, E.R.; Keller, M.A.; Maista, C.; Smith, G.; Mamounas, L.A.; Zhang, R.; Madore, S.J.; Gwinn, K. and Corriveau, R.A. (2010) "CoAIMs: A Cost-Effective Panel of Ancestry Informative Markers for Determining Continental Origins", *PLoS One*, 5 (10), pp. e13443.
- Lou, C.; Cong, B.; Li, S.; Fu, L.; Zhang, X.; Feng, T.; Su, S.; Ma, C.; Yu, F.; Ye, J. and Pei, L. (2011) "A SNaPshot Assay for Genotyping 44 Individual Identification Single Nucleotide Polymorphisms", *Electrophoresis*, 32 (3-4), pp. 368-378.
- Luce, C.; Montpetit, S.; Gangitano, D. and O'Donnell, P. (2009) "Validation of the AMPFISTR MiniFiler PCR Amplification Kit for use in Forensic Casework*", *Journal of Forensic Sciences*, 54 (5), pp. 1046-1054.
- Ma, Y.; Kuang, J.; Nie, T.; Zhu, W. and Yang, Z. (2016) "Next Generation Sequencing: Improved Resolution for Paternal/Maternal Duos Analysis", *Forensic Science International: Genetics*, 24, pp. 83-85.

Machiela, M.J. and Chanock, S.J. (2015) "LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants", *Bioinformatics*, 31 (21), pp. 3555-3557.

Mathieson, I. and McVean, G. (2012) "Differential confounding of rare and common variants in spatially structured populations", *Nature genetic*, 44(3), pp.243–246.

Moller, A. and Brinkmann, B. (1994) "Locus ACTBP2 (SE33). Sequencing Data Reveal Considerable Polymorphism", *International Journal of Legal Medicine*, 106 (5), pp. 262-267.

Montano, E.A.; Bush, J.M.; Garver, A.M.; Larijani, M.M.; Wiechman, S.M.; Baker, C.H.; Wilson, M.R.; Guerrieri, R.A.; Benzinger, E.A.; Gehres, D.N. and Dickens, M.L. (2018) "Optimization of the Promega PowerSeq™ Auto/Y System for Efficient Integration within a Forensic DNA Laboratory", *Forensic Science International: Genetics*, 32 , pp. 26-32.

Mulero, J.J.; Chang, C.W.; Lagacé, R.E.; Wang, D.Y.; Bas, J.L.; McMahon, T.P. and Hennessy, L.K. (2008) "Development and Validation of the AmpFISTR® MiniFiler™ PCR Amplification Kit: A miniSTR Multiplex for the Analysis of Degraded and/Or PCR Inhibited DNA", *Journal of Forensic Sciences*, 53 (4), pp. 838-852.

Nachman, M.W. and Crowell, S.L. (2000) "Estimate of the Mutation Rate Per Nucleotide in Humans", *Genetics*, 156 (1), pp. 297-304.

National Police Chief's Council (2017), "National DNA Database Strategy Board Annual Report 2015/16"

(https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/594185/58714_Un-Num_Nat_DNA_DB_Accessible.pdf)

National Police Chief's Council (2015) "National DNA Database Strategy Board Annual Report 2014/15",

(https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/484937/52921_NPCC_National_DNA_Database_web_pdf.pdf)

National Research Council (1996) "The Evaluation of Forensic DNA Evidence", *The National Academies Press*, Washington, DC. (<https://doi.org/10.17226/5141>).

- Novroski, N.M.M., King, J.L., Churchill, J.D., Seah, L.H. & Budowle, B. (2016), "Characterization of genetic sequence variation of 58 STR loci in four major population groups", *Forensic Science International: Genetics*, 25, pp. 214-226.
- O'Connor, K.; Butts, E.; Hill, C.; Butler, J. and Vallone P. (2010) "Evaluating the Effect of Additional Forensic Loci on Likelihood Ratio Values for Complex Kinship Analysis", *Proceedings of the 21st International Symposium on Human Identification*.
- O'Connor, K. and Tillmar, A. (2012) "Effect of Linkage between vWA and D12S391 in Kinship Analysis", *Forensic Science International: Genetics*, 6 (6), pp. 840-844.
- Osman, A.; Alsafar, H.; Tay, G.; Theyab, J.; Mubasher, M.; Eltayeb-El Sheikh, N.; AlHarthi, H., Crawford, M. and El Ghazali G. (2015) "Autosomal Short Tandem Repeat (STR) Variation Based on 15 Loci in a Population from the Central Region (Riyadh Province) of Saudi Arabia", *Journal of Forensic Research*, 6, pp. 267-271.
- Pakstis, A.J.; Speed, W.C.; Fang, R.; Hyland, F.C.L.; Furtado, M.R.; Kidd, J.R. and Kidd, K.K. (2010) "SNPs for a Universal Individual Identification Panel", *Human Genetics*, 127 (3), pp. 315-324.
- Pakstis, A.J.; Speed, W.C.; Kidd, J.R. and Kidd, K.K. (2008) "SNPs for Individual Identification", *Forensic Science International: Genetics Supplement Series*, 1 (1), pp. 479-481.
- Pakstis, A.J.; Speed, W.C.; Kidd, J.R. and Kidd, K.K. (2007) "Candidate SNPs for a Universal Individual Identification Panel", *Human Genetics*, 121 (3-4), pp. 305-317.
- Parson, W.; Ballard, D.; Budowle, B.; Butler, J.M.; Gettings, K.B.; Gill, P.; Gusmão, L.; Hares, D.R.; Irwin, J.A.; King, J.L.; Knijff, P.d.; Morling, N.; Prinz, M.; Schneider, P.M.; Neste, C.V.; Willuweit, S. and Phillips, C. (2016) "Massively Parallel Sequencing of Forensic STRs: Considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on Minimal Nomenclature Requirements", *Forensic Science International: Genetics*, 22, pp. 54-63.
- Payseur, B.; Place, M. and Weber, J. (2008) "Linkage Disequilibrium between STRPs and SNPs Across the Human Genome", *American Journal of Human Genetics*, 82 (5), pp. 1039-1050.

Peakall, R. and Smouse, P.E. (2012) "GenAEx 6.5: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research _an Update", *Bioinformatics*, 28 (19), pp. 2537-2539.

Perez-Miranda, A.; Alfonso-Sanchez, M.; Pena, J. and Herrera, R. (2006) "Qatari DNA Variation at a Crossroad of Human Migrations", *Human Heredity*, 61 (2), pp. 67-79.

Phillips, C.; Devesse, L.; Ballard, D.; Van Weert, L.; De la Puente, M.; Melis, S.; Alvarez Iglesias, V.; Freire-Aradas, A.; Oldroyd, N.; Holt, C.; Syndercombe Court, D.; Carracedo, A. and Lareu, M.V. (2018a) "Global Patterns of STR Sequence Variation: Sequencing the CEPH Human Genome Diversity Panel for 58 Forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit", *Electrophoresis*, 39 (21), pp. 2708-2724.

Phillips, C. (2017) "A Genomic Audit of Newly-Adopted Autosomal STRs for Forensic Identification", *Forensic Science International: Genetics*, 29, pp. 193-204.

Phillips, C. (2015) "Forensic Genetic Analysis of Bio-Geographical Ancestry", *Forensic Science International: Genetics*, 18, pp. 49-65.

Phillips, C. (2012) "Applications of Autosomal SNPs and Indels in Forensic Analysis", *Forensic Science Review*, 24 (1), pp. 44-62.

Phillips, C.; Amigo, J.; Carracedo, Á and Lareu, M.V. (2015) "Tetra-Allelic SNPs: Informative Forensic Markers Compiled from Public Whole-Genome Sequence Data", *Forensic Science International: Genetics*, 19, pp. 100-106.

Phillips, C.; Ballard, D.; Gill, P.; Court, D.S.; Carracedo, A. and Lareu, M.V. (2012) "The Recombination Landscape Around Forensic STRs: Accurate Measurement of Genetic Distances between Syntenic STR Pairs using HapMap High Density SNP Data", *Forensic Science International: Genetics*, 6 (3), pp. 354-365.

Phillips, C.; Fernandez-Formoso, L.; Gelabert-Besada, M.; Garcia-Magariños, M.; Santos, C.; Fondevila, M.; Carracedo, A. and Lareu, M.V. (2013) "Development of a Novel Forensic STR Multiplex for Ancestry Analysis and Extended Identity Testing", *Electrophoresis*, 34 (8), pp. 1151-1162.

Phillips, C.; Gelabert-Besada, M.; Fernandez-Formoso, L.; García-Magariños, M.; Santos, C.; Fondevila, M.; Ballard, D.; Syndercombe Court, D.; Carracedo, A. and Victoria Lareu, M. (2014) "New Turns from Old STRs": Enhancing the Capabilities of Forensic Short Tandem Repeat Analysis", *Electrophoresis*, 35 (21-22), pp. 3173-3187.

Phillips, C.; Gettings, K.B.; King, J.L.; Ballard, D.; Bodner, M.; Borsuk, L. and Parson, W. (2018b) ""The Devil's in the Detail": Release of an Expanded, Enhanced and Dynamically Revised Forensic STR Sequence Guide", *Forensic Science International: Genetics*, 34, pp. 162-169.

Phillips, C.; Parson, W.; Amigo, J.; King, J.L.; Coble, M.D.; Steffen, C.R.; Vallone, P.M.; Gettings, K.B.; Butler, J.M. and Budowle, B. (2016) "D5S2500 is an Ambiguously Characterized STR: Identification and Description of Forensic Microsatellites in the Genomics Age", *Forensic Science International: Genetics*, 23, pp. 19-24.

Phillips, C.; Salas, A.; Sánchezb, J.J.; Fondevila, M.; Gómez-Tatoc, A.; Álvarez-Diosc, J.; Calaza, M.; de Cal, M.C.; Ballard, D.; Lareu, M.V. and Carracedo, A. (2007) "Inferring Ancestral Origin using a Single Multiplex Assay of Ancestry-Informative Marker SNPs", *Forensic Science International: Genetics*, 1 (3), pp. 273-280.

Poetsch, M.; Preusse-Prange, A.; Schwark, T. and von Wurmb-Schwark, N. (2013) "The New Guidelines for Paternity Analysis in Germany-how Many STR Loci are Necessary when Investigating Duo Cases?.", *International Journal of Legal Medicine*, 127 (4), pp. 731-734.

Promega Corporation (2017) "PowerPlex® ESI 17 Pro System"

<https://www.promega.co.uk/-/media/files/resources/protocols/technical-manuals/101/powerplex-esi-17-pro-system-protocol.pdf>

Promega Corporation (2016) "PowerPlex® CS7 System Technical Manual"

<https://www.promega.co.uk/products/genetic-identity/genetic-identity-workflow/str-amplification/powerplex-cs7-system/?catNum=DC6613>

Qiagen (2016) "QIAamp DNA Mini and Blood Mini Handbook"

https://moodle.ufsc.br/pluginfile.php/1379318/mod_resource/content/0/QIAamp_DNA_Mini_Blood.pdf

Qiagen (2012a) "Investigator HDplex Handbook"

<https://www.qiagen.com/gb/resources/resourcedetail?id=7d1661bd-a47b-4b19-a882-357a61b48c64&lang=en>.

Qiagen (2012b) "Investigator® ESSplex SE Plus Handbook"

<https://www.qiagen.com/ch/resources/resourcedetail?id=8e50645a-c0ca-4331-92c0-b580b8e07d6e&lang=en>

Rahikainen, A.; Palo, J.U.; de Leeuw, W.; Budowle, B. and Sajantila, A. (2016) "DNA Quality and Quantity from Up to 16 Years Old Post-Mortem Blood Stored on FTA Cards", *Forensic Science International*, 261, pp. 148-153.

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

Rogalla, U.; Rychlicka, E.; Derenko, M.V.; Malyarchuk, B.A. and Grzybowski, T. (2014) "Simple and Cost-Effective 14-Loci SNP Assay Designed for Differentiation of European, East Asian and African Samples", *Forensic Science International: Genetics*, 14, pp. 42-49.

Rolf, B.; Schurenkamp, M.; Junge, A. and Brinkmann, B. (1997) "Sequence Polymorphism at the Tetranucleotide Repeat of the Human Beta-Actin Related Pseudogene H-Beta-Ac-Psi-2 (ACTBP2) Locus", *International Journal of Legal Medicine*, 110 (2), pp. 69-72.

Rosenberg, N.A.; Li, L.M.; Ward, R. and Pritchard, J.K. (2003) "Informativeness of Genetic Markers for Inference of Ancestry*", *American Journal of Human Genetics*, 73 (6), pp. 1402-1422.

Rosenberg, N.A.; Pritchard, J.K.; Weber, J.L.; Cann, H.M.; Kidd, K.K.; Zhivotovsky, L.A. and Feldman, M.W. (2002) "Genetic Structure of Human Populations", *Science*, 298 (5602), pp. 2381.

RStudio Team (2016) "Integrated Development for R". RStudio, Inc., Boston, MA.

[Http://Www.Rstudio.Com/](http://www.Rstudio.com/)

Ruitberg, C.; Reeder, D. and Butler, J. (2001) "STRBase: A Short Tandem Repeat DNA Database for the Human Identity Testing Community", *Nucleic Acids Research*, 29, pp. 320-322.

Sanchez, J.J.; Phillips, C.; Børsting, C.; Balogh, K.; Bogus, M.; Fondevila, M.; Harrison, C.D.; Musgrave-Brown, E.; Salas, A.; Syndercombe-Court, D.; Schneider, P.M.; Carracedo, A. and Morling, N. (2006) "A Multiplex Assay with 52 Single Nucleotide Polymorphisms for Human Identification", *Electrophoresis*, 27 (9), pp. 1713-1724.

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

<https://www.promega.com/~media/Files/Resources/Profiles%20In%20DNA/1201/Expansion%20of%20the%20European%20Standard%20Set.ashx>.

Schneider, P.M. (2007) "Scientific Standards for Studies in Forensic Genetics", *Forensic Science International*, 165 (2-3), pp. 238-243.

Serrote, C; Reiniger, L.; Silva, K.; Rabaiolli, S. and Stefanel, C. (2020) "Determining the Polymorphism Information Content of a Molecular Marker", *Gene*, 726, pp. 144175-144175.

Sinha, S.; Amjad, M.; Rogers, C.; Hamby, J.E.; Tahir, U.A.; Balamurugan, K.; Al-Kubaidan, N.A.; Choudhry, A.R.; Budowle, B. and Tahir, M.A. (1999) "Typing of Eight Short Tandem Repeat (STR) Loci in a Saudi Arabian Population", *Forensic Science International*, 104 (2-3), pp. 143-146.

Sparkes, R.; Kimpton, C.; Watson, S.; Oldroyd, N.; Clayton, T.; Barnett, L.; Arnold, J.; Thompson, C.; Hale, R.; Chapman, J.; Urquhart, A. and Gill, P. (1996) "The Validation of a 7-Locus Multiplex STR Test for use in Forensic Casework (I). Mixtures, Ageing, Degradation and Species Studies", *International Journal of Legal Medicine*, 109 (4), pp. 186-194.

Stephenson, F.H. (2010) "Forensics and paternity- Calculations for molecular biology and biotechnology: a guide to mathematics in the laboratory (Second Edition)" in Academic Press, (Boston), pp. 423-446.

STRBase (2017a) "Overview of STR Fact Sheets (D1S1656) "

[Http://Strbase.Nist.Gov/str_D1S1656.Htm](http://Strbase.Nist.Gov/str_D1S1656.Htm)

STRBase (2017b) "Overview of STR Fact Sheets (SE33) "

[Http://Strbase.Nist.Gov/str_SE33.Htm](http://Strbase.Nist.Gov/str_SE33.Htm)

Sun, J.X.; Helgason, A.; Masson, G.; Ebenesersdottir, S.S.; Li, H.; Mallick, S.; Gnerre, S.; Patterson, N.; Kong, A.; Reich, D. and Stefansson, K. (2012) "A Direct Characterization of Human Mutation Based on Microsatellites", *Nature Genetics*, 44 (10), pp. 1161-1165.

SWGAM (2019) "Addendum to "SWGAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories" to Address Next Generation Sequencing"

(https://Docs.Wixstatic.Com/Ugd/4344b0_91f2b89538844575a9f51867def7be85.Pdf).

SWGAM (2016) "Validation Guidelines for DNA Analysis Methods"

(https://Docs.Wixstatic.Com/Ugd/4344b0_813b241e8944497e99b9c45b163b76bd.Pdf

)

Applied Biosystems (2016) "GlobalFiler™ PCR Amplification Kit User Guide",

<https://tools.thermofisher.com/content/sfs/manuals/4477604.pdf>.

Applied Biosystems (2014), "The TaqMan® SNP Genotyping Assays User Guide"

https://tools.thermofisher.com/content/sfs/manuals/TaqMan_SNP_Genotyping_Assays_man.pdf

Tillmar, A.O. and Phillips, C. (2017) "Evaluation of the Impact of Genetic Linkage in Forensic Identity and Relationship Testing for Expanded DNA Marker Sets", *Forensic Science International: Genetics*, 26, pp. 58-65.

Verogen (2018a) "ForenSeq™ DNA Signature Prep Reference Guide", Document # VD2018005 Rev. A.

Verogen (2018b) "ForenSeq™ Universal Analysis Software Guide", Document # VD2018007 Rev. A.

Verogen (2018c) "MiSeq FGx™ Instrument Reference Guide", Document # VD2018006 Rev. A.

Walsh, S.; Chaitanya, L.; Clarisse, L.; Wirken, L.; Draus-Barini, J.; Kovatsi, L.; Maeda, H.; Ishikawa, T.; Sijen, T.; de Knijff, P.; Branicki, W.; Liu, F. and Kayser, M. (2014) "Developmental Validation of the HirisPlex System: DNA-Based Eye and Hair Colour Prediction for Forensic and Anthropological Usage", *Forensic Science International: Genetics*, 9, pp. 150-161.

Wang, Q.; Fu, L.; Zhang, X.; Dai, X.; Bai, M.; Fu, G.; Cong, B. and Li, S. (2016) "Expansion of a SNaPshot Assay to a 55-SNP Multiplex: Assay Enhancements, Validation, and Power in Forensic Science", *Electrophoresis*, 37 (10), pp. 1310-1317.

Wang, Z.; Zhou, D.; Wang, H.; Jia, Z.; Liu, J.; Qian, X.; Li, C. and Hou, Y. (2017) "Massively Parallel Sequencing of 32 Forensic Markers using the Precision ID GlobalFiler™ NGS STR Panel and the Ion PGM™ System", *Forensic Science International: Genetics*, 31, pp. 126-134.

Wendt, F.R.; Churchill, J.D.; Novroski, N.M.M.; King, J.L.; Ng, J.; Oldt, R.F.; McCulloh, K.L.; Weise, J.A.; Smith, D.G.; Kanthaswamy, S. and Budowle, B. (2016) "Genetic Analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx™ Forensic Genomics System", *Forensic Science International: Genetics*, 24, pp. 18-23.

Wenk, R.E. and Shao, A. (2012) "Empowering Sibship Analyses with Reference Pedigrees", *Transfusion*, 52 (12), pp. 2614-2619.

Westen, A.A.; Haned, H.; Grol, L.J.; Harteveld, J.; van der Gaag, K.J.; de Knijff, P. and Sijen, T. (2012) "Combining Results of Forensic STR Kits: HDplex Validation Including Allelic Association and Linkage Testing with NGM and Identifiler Loci", *International Journal of Legal Medicine*, 126 (5), pp. 781-789.

Wickham, H. (2016) "ggplot2: Elegant Graphics for Data Analysis" Springer-Verlag New York. <https://ggplot2.tidyverse.org>.

Wiegand, P.; Budowle, B.; Rand, S. and Brinkmann, B. (1993) "Forensic Validation of the STR Systems SE 33 and TC 11", *International Journal of Legal Medicine*, 105 (6), pp. 315-320.

Williamson, R. and Duncan, R. (2002) "DNA Testing for All", *Nature*, 418 (6898), pp. 585-586.

Wilson, I.G. (1997) "Inhibition and Facilitation of Nucleic Acid Amplification", *Applied and Environmental Microbiology*, 63 (10), pp. 3741-3751.

Woerner, A.E.; King, J.L. and Budowle, B. (2017) "Fast STR Allele Identification with STRait Razor 3.0", *Forensic Science International: Genetics*, 30, pp. 18-23.

Wong, S. and Anokute, C. (1990) "The Effect of Consanguinity on Pregnancy Outcome in Saudi Arabia", *Journal of The Royal Society of Health*, 110 (4), pp. 146–147.

Wu, W.; Hao, H.; Liu, Q.; Han, X.; Wu, Y.; Cheng, J. and Lu, D. (2014) "Analysis of Linkage and Linkage Disequilibrium for Syntenic STRs on 12 Chromosomes", *International Journal of Legal Medicine*, 128 (5), pp. 735-739.

Xavier, C. and Parson, W. (2017a) "Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit - MPS Forensic Application for the MiSeq FGx™ Benchtop Sequencer", *Forensic Science International: Genetics*, 28, pp. 188-194.

Zhang, S.; Bian, Y.; Tian, H.; Wang, Z.; Hu, Z. and Li, C. (2015) "Development and Validation of a New STR 25-Plex Typing System", *Forensic Science International: Genetics*, 17, pp. 61-69.

Zhu, B.; Zhang, Y.; Shen, C.; Du, W.; Liu, W.; Meng, H.; Wang, H.; Yang, G.; Jin, R.; Yang, C.; Yan, J. and Bie, X. (2015) "Developmental Validation of the AGCU 21+1 STR Kit: A Novel Multiplex Assay for Forensic Application", *Electrophoresis*, 36 (2), pp. 271-276.

10 Chapter Ten: Appendixes

10.1 Appendix 1

10.1.1 Scenarios specific equations that used for calculating the RI.

Table 10.1. Scenarios specific equations that used for calculating the RI. The table shows the equations that are used in calculating RI based on the genotypes of the tested individuals. The numerator (X) represents the probability of that the alleged father has passed the common allele with the disputed child. The denominator (Y) represent the probability of that random male from the same population is the source of shared allele. The RI equations are the result of numerator (X)/denominator (Y). The last five rows, show specific scenarios' equations that used for calculating the RI when the mother's genotype is not available (Stephenson 2010).

Mother	Genotypes		Numerator (X)	Denominator (Y)	RI (X/Y)
	Child	Alleged father			
AA	AA	AA	1	pA	1/ pA
AA	AA	AB	1/2	pA	1/2 pA
AA	AA	BC	0	pA	0
AB	AA	AA	1/2	pA/2	1/ pA
AB	AA	AB	1/4	pA/2	1/2 pA
AB	AA	AC	1/4	pA/2	1/2 pA
AB	AA	BC	0	pA/2	0
AA	AB	AB	1/4	pB/2	1/2 pB
AA	AB	BB	1	pB	1/ pB
AA	AB	BC	1/2	pB	1/2 pB
AA	AB	CD	0	PA	0
AB	AB	AA	1/2	(pA+pB)/2	1/(pA+pB)
AB	AB	AB	1/2	(pA+pB)/2	1/(pA+pB)
AB	AB	BC	1/4	(pA+pB)/2	1/[2(pA+pB)]
AB	AB	AC	1/4	(pA+pB)/2	1/[2(pA+pB)]
AB	AC	AC	1/2	pC	1/2 pC
AB	AC	CD	1/4	pC/2	1/ 2pC
AB	AC	BC	1/4	pC/2	1/2pC
AB	BC	CC	1/2	pC/2	1/ pC
AB	BB	AB	1/4	pB/2	1/ 2pB
AB	BC	BC	1/2	pC	1/2pC
AB	BC	CD	1/4	pC/2	1/ 2pC
AB	AB	CD	0	(pA+pB)/2	0
AC	AB	BB	1/2	pB/2	1/ pB
AC	AB	BD	1/4	pB/2	1/ 2pB
AC	AB	BC	1/4	pB/2	1/ 2pB
AC	AB	CD	0	pB/2	0
n/a	AA	AA	1	pA	1/pA
n/a	AA	AB	1	2pA	1/2pA
n/a	AB	AB	(pA+pB)	4pApB	(pA+pB)/4pApB)
n/a	AB	BB	1	2pB	1/2pB
n/a	AB	BC	1	4pB	1/4pB

A, B, C and D: the possible alleles of the tested individuals.

pA, pB and pC: the frequency of the alleles A, B and C respectively.

n/a: when the genotype is not available.

10.1.2 Scenarios specific equations that used for calculating the RI when the child is missing

Table 10.2. Scenarios specific equations that used for calculating the RI when the child is missing, and the genotypes of the parent are is available (AABB 2010b)

Genotypes			Numerator (X)	Denominator (Y)	RI (X/Y)
Mother	Alleged child	Father			
AA	AA	AB	1	$2pA^2$	$1/2pA^2$
AA	AB	AB	1	$4pApB$	$1/4pApB$
AA	AB	BC	1	$4pApB$	$1/4pApB$
AB	AA	AB	1	$4pA^2$	$1/4pA^2$
AB	AA	AC	1	$4pA^2$	$1/4pA^2$
BC	AB	AB	1	$8pApB$	$1/8pApB$
BC	AB	AC	1	$8pApB$	$1/8pApB$
BD	AB	AC	1	$8pApB$	$1/8pApB$
AA	AA	AA	1	pA^2	$1/pA^2$
AB	AA	AA	1	$2pA^2$	$1/2pA^2$
BB	AB	AA	1	$2pApB$	$1/2pApB$
BC	AB	AA	1	$4pApB$	$1/4pApB$
AB	AB	AC	1	$8pApB$	$1/8pApB$
AB	AB	AA	1	$4pApB$	$1/4pApB$
AB	AB	AB	1	$4pApB$	$1/4pApB$

A, B, C and D: the possible alleles of the tested individuals.

pA and pB: the frequency of the alleles A and B respectively.

RI: Relationship Index

10.1.3 Scenarios specific equations that used for calculating the SI and HSI.

Table 10.3. Scenarios specific equations that used for calculating the Sibling Index (SI) and the Half-Sibling Index (HSI) (AABB 2010b)

Genotypes			
Sibling / half sibling	Alleged Sibling/half sibling	Sibling Index (SI)	Half-Sibling Index (HSI)
AB	AB	$(1+pA+pB+2pApB)/8pApB$	$(pA+pB+4pApB)/8pApB$
AA	AA	$(1+pA)^2/(2pA)^2$	$(1+pA)/2pA$
AA	AB	$(1+pA)/4pA$	$(1+2pA)/4pA$
AB	AC	$(1+2pA)/8pA$	$(1+4pA)/8pA$
AB	CD	0.25	0.5

A, B, C and D: the possible alleles of the tested individuals.

pA and pB: the frequency of the alleles A and B respectively.

10.1.4 Including the mutation event into the RI-LR

First, by directly substituting the LR with the mutation rate of the locus. For example, the mutation rate of the CSF1PO is 0.002021 (AABB 2008), and when a mutation event is expected, the LR will be 0.002021 for this locus.

Second, by dividing the mutation rate of the locus by the power of exclusion of the same locus (Butler 2015), by which two hypotheses are considered: X (the alleged father is the true father and a mutation has occurred and has inherited to the child), which is equal to the mutation rate (μ), and Y (the alleged father is not the true father and the allele has inherited from unrelated man) which is equal to the power of exclusion (PE).

Using the above example:

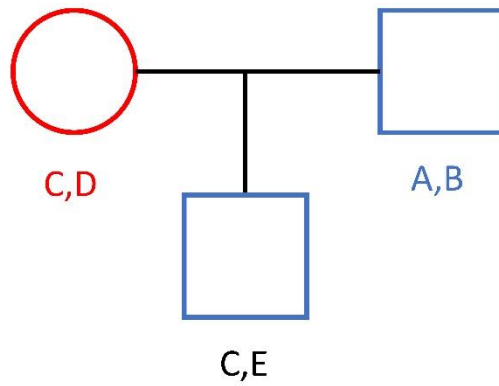
$$LR = X/Y = \mu \text{ of CSF1PO} / (\text{PE})$$

$$LR = 0.002021 / 0.431 \text{ (the PE of the locus for the Saudi population (Alsafiah } et al. \text{ 2017))}$$

$$LR = 0.00469.$$

However, the first way does not compare the two probabilities of the typical hypotheses, and the second one does not include the inheritance probability from the parent to the child (Allen 2013).

Third, compares two hypotheses, take in account the inheritance probability and uses allele-specific mutation rate (Figure 10.1)(Gjertson *et al.* 2007). Although the AABB has provides allele-specific mutation (paternal and maternal) rates for 15 STRs (AABB 2008), the data does not include the rest of commonly used STRs (e.g. D1S1656, D2S441, D10S1248, D12S391, D22S1045, SE33, D6S1043, Penta D and Penta E) (Gjertson *et al.* 2007).

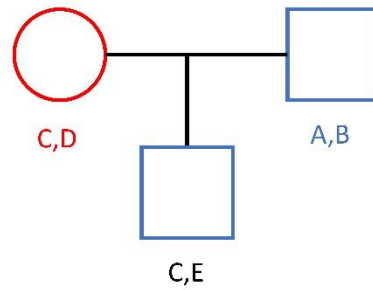


$$LR = \frac{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} \left(\frac{1}{2}\right)_{E \text{ is from the Alleged father}} (\mu_{B \rightarrow E} + \mu_{A \rightarrow E})}{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} (p_E)}$$

$$LR = \frac{(\mu_{B \rightarrow E} + \mu_{A \rightarrow E})}{2 p_E}$$

Figure 10.1. Incorporating the allele-specific mutation rates in the calculation of the RI-LR. This Figure explains the third way of incorporating the allele-specific mutation rates in the calculation of the RI-LR. The allele-specific mutation (paternal and maternal) rates for are provided in (AABB 2008). $\mu_{B \rightarrow E}$: is the mutation rate of allele B to allele E, $\mu_{A \rightarrow E}$: is the mutation rate of allele A to allele E, and p_E : the frequency of the allele E (An original figure).

Fourth, a more appropriate way for the STR markers (Gjertson *et al.* 2007), was suggested by Brenner (2018), which assumes a fixed probability for each type of mutation (0.5 for a single step increase/decrease, 0.05 for two steps increase/ decrease, 0.005 for three steps increase/ decrease.... etc), includes the average mutation rate of the locus (μ), and compares the two hypotheses (Figure 10.2).



If allele E = ± one repeat of the allele A or allele B

$$LR = \frac{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} (\mu)_{\text{Average mutation rate of the locus}} \left(\frac{1}{2}\right)_{\text{Probability of a single step increase or decrease}}}{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} (pE)}$$

If allele E = ± two repeats of the allele A or allele B

$$LR = \frac{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} (\mu)_{\text{Average mutation rate of the locus}} (0.05)_{\text{Probability of two steps increase or decrease}}}{\left(\frac{1}{2}\right)_{C \text{ is from the mother}} (pE)}$$

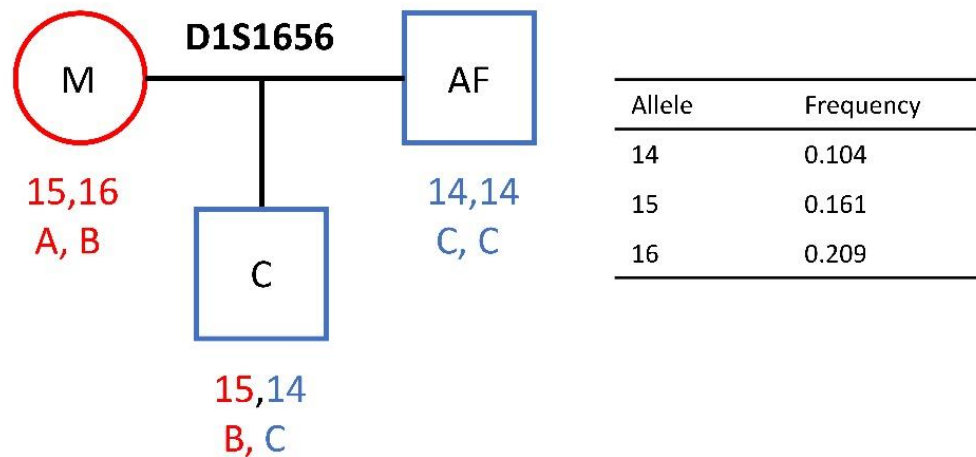
Figure 10.2. Incorporating the mutation event into the calculation of the RI-LR. This figure describes a way of including the mutation event into the calculation of the RI-LR using a fixed probability for each type of mutation that was suggested by Brenner (2018) (An original figure).

10.1.5 Including the prior probability (Pr) to the posterior probability (Po).

The Pr and the genetic evidence are included in the calculation of the posterior probability (Po) (i. e. relationship probability) as follows: $Po = (CRI \times Pr) / (CRI \times Pr + (1 - Pr))$.

10.1.6 RMNE calculation.

The RMNE (random man not excluded) can also be calculated using the frequency of the shared allele between the alleged father and the child. In other words, what the portion of the population that could have the shared allele. Assuming allele A is the shared allele and p is the frequency of the allele A, the genotypes (homozygous and heterozygous) that could have the allele A can be calculated by using the HW-equation ($p^2 + 2pq = RMNE$, where $q = 1 - p$). Subsequently, the power of exclusion (PE) can be estimated by using the equation $PE = 1 - RMNE$ (Allen 2013) (see Figure 10.3 and Figure 10.4).



$$PI = 1 / p_C \text{ (see Table 11.1)}$$

$$PI = 1 / 0.104 = 9.615$$

⇒ There is one chance in 9.615 that random unrelated man from the same population is the biological father.

Assuming $Pr = 0.5$

$$\text{Paternity probability} = (PI \times Pr) / (PI \times Pr + (1-Pr)) \times 100 = (9.615 \times 0.5) / (9.615 \times 0.5 + 1 - 0.5) \times 100$$

$$\text{Paternity probability} = 90.579\%$$

⇒ 90.6% is the chance that the AF is the source of allele 14.

$$RMNE = p_C^2 + 2p_C p_E \text{ (E represents all other possible alleles)}$$

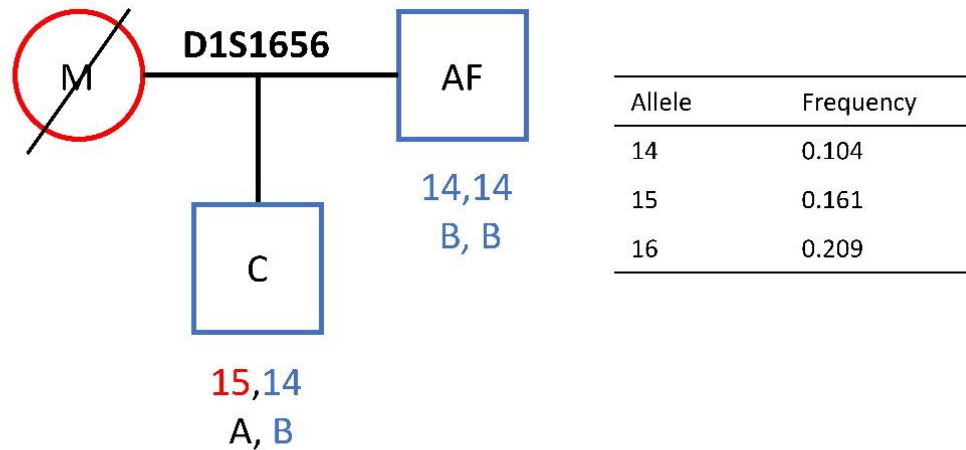
$$RMNE = (0.104)^2 + 2(0.104)(1 - 0.104) = 0.0108 + 2(0.104)(0.896) = 0.197 = 19.7\%$$

⇒ 19.7% of the population is expected to have the allele 14.

$$PE = 1 - RMNE = 1 - 0.197 = 0.803 = 80.3\%$$

⇒ 80.3% of the population is excluded from being the biological father.

Figure 10.3. An example of calculating the PI, paternity probability, RMNE and PE. This figure shows a typical parentage case and shows how the strength of evidence can be estimated. In this example, the specific equation was adopted from Table 10.1 based on the genotypes of the tested individuals and the frequencies of the D1S1656 alleles were adopted from (Alsafiah *et al.* 2017). By only one locus, the PI shows that there is 1/9.6 chance random unrelated man from the same population is the biological father. The paternity probability shows that 90.6% (posterior probability) is the chance that the AF is the source of the shared allele comparing to 50% (prior probability). Based on the RMNE, the PE is 80.3% that means 80.3% of the population is excluded from being the biological father of the disputed child (an original figure).



$$PI = 1/2p_B \text{ (see Table 11.1)}$$

$$PI = 1 / 2 (0.104) = 1/ 0.208 = 4.807$$

⇒ There is one chance in 4.807 that random unrelated man from the same population is the biological father.

Assuming Pr = 0.5

$$\text{Paternity probability} = (PI \times Pr) / (PI \times Pr + (1-Pr)) \times 100 = (4.807 \times 0.5) / (4.807 \times 0.5 + (1-0.5)) \times 100 = 82.77\%$$

⇒ 82.77% is the chance that the AF is the source of allele 14.

$$RMNE = p_B^2 + 2p_B p_E \text{ (E represents all other possible alleles)}$$

$$RMNE = (0.104)^2 + 2(0.104)(1 - 0.104) = 0.0108 + 2(0.104)(0.896) = 0.197 = 19.7\%.$$

⇒ 19.7% of the population is expected to have the allele 14.

$$PE = 1 - RMNE = 1 - 0.197 = 0.803 = 80.3\%$$

⇒ 80.3% of the population is excluded from being the biological father.

Figure 10.4. An example of calculating the PI, paternity probability, RMNE and PE in a mother-less case. This figure shows a typical parentage mother-less case and shows how the strength of evidence can be quantified and estimated. In this example, the specific equation was adopted from Table 10.1. (an original figure).

10.2 Appendix 2

10.2.1 Participant Information Sheet



University of Central Lancashire
School of Forensic and Applied Sciences
Preston PR1 2HE

Participant Information Sheet

Research title: forensically relevant polymorphisms in the population of Saudi Arabia

I am a student at the University of Central Lancashire in the UK undertaking research for a PhD in Forensic Genetics. You are being invited to take part in a research study. Before you decide whether or not to take part it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully.

About the study

This project aims to assess the application of different forensic DNA markers in the context of Saudi Arabia.

Some crime scene samples are challenging because there is not much DNA or it has broken down. It is therefore difficult to use the evidence for investigative or legal purposes. However, new types of DNA markers have been developed that can overcome some of the challenges.

In order to assess the usefulness of these new DNA markers within the population of Saudi Arabia, biological samples from Saudi population representatives are needed. DNA will be extracted from the blood samples and the DNA analysed. It is then possible to assess to what extent these markers can be used to identify individuals.

If these markers overcome the limitations of the conventional forensic tests, this will help with police investigations and ultimately facilitate the courts of law.

Why have I been invited to participate?

You have been asked to be a participant because you are from the population of Saudi Arabia.

Do I have to participate?

It is your decision whether or not to take part in this study. If you choose to participate, you will be asked to sign a consent form (Version 2: 19/10/2016), and you can keep this information sheet for future reference.

If you change your decision within the week following sampling, your sample can be withdrawn without giving a reason, and it would be destroyed as soon as practical, and within one week. You will be given a code to identify your sample—this will be needed in order to withdraw your sample from the study. Otherwise, your sample will only be identified to the researchers as male or female.

What are the benefits if I participate in this research?

Participant Information Sheet: Version 2. 19/10/2016

There are no direct benefits from participating in this research.

What is involved?

If you agree to help, you will need to reply to this e-mail and you will be given an appointment for sampling. You still have the time, starting from your response until the time of your appointment, to think about participating in this study. If you still willing to participate you will need to fill a consent form (attached), I will then collect your sample on the FTA card via the following process:

- 1) cleansing your finger with an alcohol swab.
- 2) pricking your finger by using a disposable (single-use) sterilised needle (1-2 mm).
- 3) applying 2-3 blood spots on the FTA card.
- 4) cleansing your finger again with an alcohol swab.

The FTA card will process your blood and keep your DNA for the study.

After you have given your sample, you will be given a code that will only be kept by you. If you wish to withdraw from the study you should contact the researchers listed below and request that your sample be removed. This can be done for up to one week after your donation, and we will inform you that your sample has been removed and destroyed as requested.

You will need to provide the code so that your sample can be identified—we will be unable to identify your sample other than through the code, so it is important that you keep it.

What will happen to the results of the study?

The samples provided and the data generated using them will only be used in this study. Data will be stored according to UCLan's data protection guidelines. At the end of the research, I will write a thesis based on the results obtained, but no participant will be identified. If you are interested, the thesis will be available on Central Lancashire's Online Knowledge (CLOK) site in approximately five years' time. The work may also be published in scientific journals and presented at scientific meetings. None of your personal details will be published. As for your sample, it will be destroyed during or on completion of this study.

Confidentiality and anonymity

The data we collect will not contain any personal information about you. No one will be able to link the biological sample you provide to the identifying information you supply (i.e., your name). The biological samples provided will be destroyed during or on completion of this study (estimated completion time: March 2020).

Who is carrying out the research?

Participant Information Sheet: Version 2. 19/10/2016

Hussain M. Alsafiah, a PhD student within the School of Forensic and Applied Sciences, is undertaking this research. The Saudi Arabian Cultural Bureau in London is funding this research.

How has the work been reviewed?

The work has been reviewed by the University of Central Lancashire STEMH Research Ethics Committee.

If you have any further doubts, please feel free to contact me or my supervisor about them. Concerns or complaints about this project should be addressed to the University Officer for Ethics at OfficerForEthics@uclan.ac.uk.

Thank you for your time.

Contact information

Research Student: Hussain M. Alsafiah

E-Mail: hmhalsafiah@uclan.ac.uk

Saudi contact Number: 00966558044100

Uk contact Number: 00447576595566

Research Supervisor: Dr. William Goodwin

E-Mail: whgoodwin@uclan.ac.uk



University of Central Lancashire
School of Forensic and Applied Sciences
Preston PR1 2HE

Consent Form

Research title: studying forensically relevant polymorphisms in the population of Saudi Arabia.

Researchers: Hussain M. Alsafiah, E-Mail: hmhalsafiah@uclan.ac.uk
Dr. William Goodwin, E-Mail: whgoodwin@uclan.ac.uk

I, the undersigned, confirm that (please tick box as appropriate):

1. I have read and understood the participant information sheet given to me with this form for the above study.
2. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
3. I understand that my participation in this study is completely voluntary.
4. I understand that I have at least one week (until the processing of the sample) after the collection of the sample to withdraw from the study without giving any reason, and my sample will be destroyed within one week.
5. The procedures regarding confidentiality have been clearly explained to me.
6. The use of my sample and data in this study, and publications, has been explained to me and I have no objection to that use.
7. I agree to participate in this study.

Name of Participant: _____

Signature: _____ Date ____/____/____

Name of Person collecting Permission _____

Signature: _____ Date ____/____/____

10.3 Appendix 3

10.3.1 Sample collection approval from the Security Forces Hospitals Programme.

الرقم :
التاريخ :
المرفقات :

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ
برنامج مستشفى قوى الأمن بالدمام
Security Forces Hospital Program - Dammam

المملكة العربية السعودية
وزارة الداخلية
الإدارة العامة للخدمات الطبية
برنامج مستشفى قوى الأمن بالدمام

Date : 27th June 2016

To : **Dr. William Goodwin**
Reader in Forensic Genetics
School of Forensic Applied Sciences
University of Central Lancashire
Preston PR1 2HE
United Kingdom

Subject : **Re: Biological Sample Collection for Mr. Hussain Mohammed Alsafiah's PhD Research**

Dear Dr. William,

With reference to your request for biological samples from the population of Kingdom of Saudi Arabia, I am pleased to inform you that the forensic medicine department would be happy to allow Mr. Hussain Mohammed Alsafiah access to its facilities to collect biological samples from volunteers from the population of Kingdom of Saudi Arabia.

We understand that information about the project will be provided to participants and that all volunteers will sign a consent from prior to donation a sample. We also understand that no pressure shall be applied to any individual to participate that does not wish to provide a sample.

Forensic Medicine Department also agreed to give Mr. Hussain Mohammed Alsafiah permission to transport these samples to preston for the purpose of carrying out his PhD research.



Colonel Dr. Mohammed Ahmed Alshaikhi
Forensic Medicine Consultant / Head of Forensic Department
Ministry of Interior
Mobile: +966551155524
Email: shkimbmd@hotmail.com

هاتف : ٠١٣/٨١٠٥٠٠٠ - فاكس : ٠١٣/٨١٠٣٦٠١ - ص.ب ٩٠٠٣ الرمز البريدي ٣١٤١٣ الدمام - المملكة العربية السعودية
Tel. : 013/8105000 - Fax : 013/8103601 - P.O.Box 9003 Dammam 31413 - Kingdom of Saudi Arabia
Website : www.dammam.sfh.med.sa Email : info@sfdh.med.sa

10.3.2 STEMH 557 ethical approval for the project



28 October 2016

Will Goodwin / Hussain Alsafiah
School of Forensic and Applied Sciences
University of Central Lancashire

Dear Will / Hussain

Re: STEMH Ethics Committee Application
Unique Reference Number: STEMH 557

The STEMH ethics committee has granted approval of your proposal application 'Forensically Relevant Polymorphisms (STRs/SNPs) in the population of Saudi Arabia'. Approval is granted up to the end of project date* or for 5 years from the date of this letter, whichever is the longer. It is your responsibility to ensure that

- the project is carried out in line with the information provided in the forms you have submitted
- you regularly re-consider the ethical issues that may be raised in generating and analysing your data
- any proposed amendments/changes to the project are raised with, and approved, by Committee
- you notify roffice@uclan.ac.uk if the end date changes or the project does not start
- serious adverse events that occur from the project are reported to Committee
- a closure report is submitted to complete the ethics governance procedures (Existing paperwork can be used for this purposes e.g. funder's end of grant report; abstract for student award or NRES final report. If none of these are available use [e-Ethics Closure Report Proforma](#)).

Additionally, STEMH Ethics Committee has listed the following recommendation(s) which it would prefer to be addressed. Please note, however, that the above decision will not be affected should you decide not to address any of these recommendation(s).

Should you decide to make any of these recommended amendments, please forward the amended documentation to roffice@uclan.ac.uk for its records and indicate, by completing the attached grid, which recommendations you have adopted. Please do not resubmit any documentation which you have **not** amended.

Yours sincerely

A handwritten signature in black ink, appearing to read 'A Chohan', is written above the printed name.

Ambreen Chohan
Chair
STEMH Ethics Committee

* for research degree students this will be the final lapse date

10.4 Appendix 4

10.4.1 STRidER final report for the data of the 17 non-CODIS loci.

Institute of Legal Medicine
Medical University of Innsbruck
Director: Prof. Richard Scheithauer, MD
Muellerstraße 44, A-6020 Innsbruck, Austria
Tel: +43/512/9003-70600, Fax: +43/512/9003-73600
e-mail: gmi@i-med.ac.at



Prof. Dr. Walther Parson
Head of High Throughput DNA Database Unit
Head of Forensic Genomics
walther.parson@gmail.com

Dr. Martin Bodner
Forensic Genomics
martin.bodner@i-med.ac.at

Innsbruck, 27.02.2019

STRidER QC Report – Dataset STR000178 Alsafiah SAU 500

Dear submitter,

Thank you for providing autosomal STR data for quality control (QC) to STRidER. Please find the results in this report.

1. Submitted datasets

origin of samples: Saudi Arabia
submitter: Hussain Alsafiah
School of Forensic and Applied Sciences, University of Central
Lancashire, UK
HMHAlsafiah@uclan.ac.uk
no. of genotypes: 501 (submitted), 500 (accepted)
autosomal STR loci: 17 (SureID 23comp Human DNA Identification kit)
format: genotype table, CE length based alleles

2. General information on the QC process

The dataset was re-submitted to STRidER after rejection of the original submission for reasons of quality concerns. The STR genotypes were scrutinized applying plausibility checks performed with the STRidER software suite and further manually scrutinized. Observations were made that included the invitation to send raw data. The reason for doing so is that we cannot know whether our observations indicate new variation or actual errors in the dataset. To clarify this, we invited contributors to check/confirm all observations.

The submission included the following loci: D18S1364, D13S325, D5S2800 (erroneously called D5S2500 in the original submission), D9S1122, D4S2366, D3S1744, D11S2368, D21S2055, D20S482, D8S1132, D7S3048, D19S253, D17S1301, D22GATA198B05,

STRidER_Report_Alsafiah_SAU_500_STR000178

D6S474, D14S1434, D15S659. The remaining loci contained in the kit were not sent to STRidER for quality control.

3. Results of QC: corrections made to the datasets

The data are of general good quality and seem to have been produced and analysed in high-quality DNA laboratories. The following corrections were made to the datasets during QC after communication with the submitter:

3.1. Identical but incomplete genotype pair

Genotype 511 was included twice in the submission. The incomplete copy was deleted (remaining sample number: 500).

3.2. Alleles at five loci in five genotypes were found erroneous after EPG inspection

sample	locus	submitted	corrected
237	D17S1301	12,12	11,12
266	D20S482	13,18	13,13
341	D22GATA198B05	18,18	18,19
501	D14S1434	10,10	10,14
346	D15S659	12,12	12,16

4. Summary

We thank the authors for providing autosomal STR population data for STRidER QC! The general quality of the data as determined by plausibility checks and inspection of raw data appears to meet forensic requirements.

Please find the **allele frequency table** calculated by STRidER and the **revised dataset** attached. When publishing the datasets, please indicate STRidER dataset reference **STR000178** in the manuscript(s) and provide this number to the editor during submission. If you encounter any inconsistencies in this dataset in the future, please let us know.

Please **cite STRidER** when publishing your research:

Bodner M., Bastisch, I., Butler, J.M., Fimmers, R., Gill, P., Gusmão, L., Morling, N., Phillips, C., Prinz, M., Schneider, P.M., Parson, W. (2016), 'Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)', Forensic Sci. Int. Genet. 24, 97-102

The STRidER online platform is work in progress. Additional datasets and features will continuously become available. To receive periodic news and stay updated about

Institute of Legal Medicine, Medical University of Innsbruck

STRidER, please register for the **STRidER newsletter** [<https://mailman.i-med.ac.at/mailman/listinfo/strider-l>].

Kind regards,

Dr. Walther Parson

Dr. Martin Bodner

Disclaimer: The applied quality control cannot be regarded as comprehensive independent evaluation of all raw data of the submitted dataset, but constitutes an optimized procedure for the detection of common data idiosyncrasies. The signatories cannot be made liable for correctness, completeness and topicality of the contents.

10.5 Appendix 5

Table 10.4. Sequence-based data for 27 aSTRs generated from Chapter 6.

STRs	Total Genotypes	Allele	Size Based Data	Repeat Region Sequence Data		Repeat and flanking regions	
			Frq.	Repeat Region Sequence	Frq.	Sequence	Frq.
D1S1656	174	10	0.00575	[TAGA]10	0.00575	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.00575
		11	0.05747	[TAGA]11	0.05747	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.05747
		12	0.12644	[TAGA]11 TAGG	0.08046	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.08046
				[TAGA]12	0.04598	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.04598
		13	0.05747	[TAGA]12 TAGG	0.02299	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.02299
				[TAGA]13	0.03448	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.03448
		14	0.11494	[TAGA]13 TAGG	0.11494	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.11494
		14.3	0.00575	[TAGA]4 TGA [TAGA]9 TAGG	0.00575	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.00575
		15	0.15517	[TAGA]14 TAGG	0.12069	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.12069
				[TAGA]15	0.02299	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.02299
				[TAGA]14 TAAG	0.01149	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.01149
		15.3	0.06897	[TAGA]4 TGA [TAGA]10 TAGG	0.06322	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.06322
				[TAGA]3 TGA [TAGA]11 TAGG	0.00575	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.00575
		16	0.22414	[TAGA]15 TAAG	0.01724	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.01724
				[TAGA]15 TAGG	0.2069	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.2069
		16.3	0.04598	[TAGA]4 TGA [TAGA]11 TAGG	0.04598	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.04598
		17	0.09195	[TAGA]16 TAGG	0.09195	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.09195
		17.3	0.02874	[TAGA]4 TGA [TAGA]12 TAGG	0.02874	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.02874
		18.3	0.00575	[TAGA]4 TGA [TAGA]13 TAGG	0.00575	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.00575
		19.3	0.01149	[TAGA]4 TGA [TAGA]14 TAGG	0.01149	TAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGTGTGTGTTAATTGTATGTATATATATTTGGTCCCTAGTGATTCTATTCTCTGAA	0.01149

Table 10.4. continued.

PentaE	172	5	0.02907	[AAAGA]5	0.02907	AAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.02907
		7	0.00581	[AAAGA]7	0.00581	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.00581
		8	0.0814	[AAAGA]8	0.0814	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.0814
		9	0.03488	[AAAGA]9	0.03488	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.03488
		10	0.05233	[AAAGA]10	0.05233	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.05233
		11	0.12791	[AAAGA]11	0.12791	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.12791
		12	0.19186	[AAAGA]12	0.19186	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.19186
		13	0.10465	[AAAGA]13	0.10465	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.10465
		14	0.06395	[AAAGA]14	0.06395	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.06395
		14.4	0.00581	AAGA [AAAGA]14	0.00581	AAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.00581
		15	0.05233	[AAAGA]14 AAATA	0.00581	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.00581
				[AAAGA]15	0.04651	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.04651
		16	0.05233	[AAAGA]16	0.05233	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.05233
		16.4	0.01744	AAGA [AAAGA]16	0.01744	AAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.01744
		17	0.05233	[AAAGA]17	0.04651	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.04651
				[AAAGA]10 AAATA [AAAGA]6	0.00581	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.00581
		18	0.0407	[AAAGA]18	0.03488	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.03488
				[AAAGA]16 [AAATA]2	0.00581	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.00581
		19	0.05233	[AAAGA]19	0.05233	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.05233
		20	0.02326	[AAAGA]20	0.02326	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.02326
		21	0.01163	[AAAGA]21	0.01163	AAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAATTGTAAGGAGTTTTCT	0.01163

Table 10.4. continued.

PentaD	174	2.2	0.05747	[AAAGA]5	0.05747	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAGAAAAAGAAAAAGAAA AAACGAAGGGGAAAAAAGAGAATCATAACATAAATGTAAAAATTTCTCAA	0.05747
		7	0.01149	[AAAGA]7	0.01149	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.01149
		8	0.0115	[AAAGA]8	0.0115	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.00575
						GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.00575
		9	0.17241	[AAAGA]9	0.17241	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.17241
		10	0.18966	[AAAGA]10	0.18966	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.18966
		11	0.21264	[AAAGA]11	0.21264	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.21264
		12	0.08621	[AAAGA]12	0.08621	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.08621
		13	0.17241	[AAAGA]13	0.17241	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.17241
		14	0.06322	[AAAGA]14	0.06322	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.06322
15	0.01724	[AAAGA]15	0.01724	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTCAA	0.01724		
16	0.00575	[AAAGA]16	0.00575	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGAGCCATGATCACACCACTACACTCCAGCCTAGGTGACAGAGCAAGACCCATCTCAAGAAAGAAAAAAGAAAAAGAAAAAG AAAAGAAAAAGAAAAAGAAAAAGAAAAACGAAGGGGAAAAAAGAGAATCATAAACATAAATGTAAAAATTTCTC AA	0.00575		
DZ51045	167	10	0.0061	[ATT]7 ACT [ATT]2	0.0061	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTACTATTATTATTATAAAAAATTGCCAAT	0.0061
		11	0.13415	[ATT]8 ACT [ATT]2	0.13415	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTACTATTATTATTATAAAAAATTGCCAAT	0.13415
		12	0.02439	[ATT]9 ACT [ATT]2	0.02439	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTACTATTATTATTATAAAAAATTGCCAAT	0.02439
		14	0.06098	[ATT]11 ACT [ATT]2	0.06098	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTATTATTATTACTATTATTATTATAAAAAATTGCCAAT	0.06098
		15	0.48171	[ATT]12 ACT [ATT]2	0.48171	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTATTATTATTACTATTATTATTATAAAAAATTGCCAAT	0.48171
		16	0.2378	[ATT]13 ACT [ATT]2	0.2378	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTATTATTATTACTATTATTATTATAAAAAATTGCCAAT	0.2378
		17	0.05488	[ATT]14 ACT [ATT]2	0.05488	CGTTGGAATTCCTCCAACTGGCCAGTTCCTCCACCCTATAGACCTGTCTAGCCTCTTATAGCTGCTATGGGGCTAGATTTTCCCAGATGATAGTGTCTATTATTATTATTATT ATTATTATTATTATTATTACTATTATTATTATAAAAAATTGCCAAT	0.05488

Table 10.5. HWE test for aSTRs for the data generated from Chapter 6. None of the analysed markers showed significant deviation from HWE after Bonferroni correction (P value>0.0004). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/121$ loci = 0.0004.

STRs	Size-based data			Repeat region			repeat and flanking regions		
	Ho	He	HW P-value	Ho	He	HW P-value	Ho	He	HW P-value
D1S1656	0.8046	0.87855	0.19358	0.82759	0.90439	0.1428	0.82759	0.90439	0.1428
TPOX	0.66667	0.66833	0.30143	0.66667	0.66833	0.30143	0.66667	0.66833	0.30143
D2S441	0.66667	0.73523	0.14683	0.67816	0.73802	0.17308	0.71264	0.75769	0.23924
D2S1338	0.77011	0.85921	0.0077	0.85057	0.91489	0.23498	0.85057	0.91489	0.23498
D3S1358	0.7931	0.77244	0.03995	0.87356	0.89044	0.49752	0.87356	0.89044	0.49752
D4S2408	0.66667	0.73949	0.49684	0.66667	0.73949	0.49684	0.66667	0.73949	0.49684
FGA	0.90805	0.86885	0.94851	0.91954	0.87124	0.9274	0.91954	0.87124	0.9274
D5S818	0.70115	0.76194	0.49358	0.70115	0.76194	0.49358	0.85057	0.85602	0.37169
CSF1PO	0.65517	0.7256	0.24374	0.67816	0.73836	0.19969	0.67816	0.73836	0.1997
D6S1043	0.73563	0.81131	0.56902	0.73563	0.81131	0.56902	0.73563	0.81131	0.56902
D7S820	0.75862	0.77922	0.45703	0.75862	0.77922	0.45703	0.83908	0.85569	0.96444
D8S1179	0.82759	0.84001	0.63187	0.88506	0.88984	0.67975	0.88506	0.88984	0.67975
D9S1122	0.75862	0.69271	0.84892	0.89655	0.83569	0.85082	0.89655	0.83569	0.85082
D10S1248	0.70115	0.72168	0.65711	0.70115	0.72168	0.65711	0.70115	0.72168	0.65711
TH01	0.65517	0.77038	0.18778	0.65517	0.77038	0.18778	0.65517	0.77038	0.18778
vWA	0.77011	0.79277	0.69823	0.82759	0.83284	0.08192	0.82759	0.83284	0.08192
D12S391	0.87356	0.89602	0.34996	0.94253	0.94412	0.35894	0.94253	0.94412	0.35894
D13S317	0.70115	0.7438	0.76506	0.70115	0.7438	0.76506	0.86207	0.86838	0.78149
PentaE	0.7907	0.91201	0.00217	0.80233	0.9135	0.00828	0.80233	0.9135	0.00828
D16S539	0.70115	0.76912	0.11584	0.70115	0.76912	0.11584	0.73563	0.80101	0.21452
D17S1301	0.6092	0.69225	0.00988	0.6092	0.69225	0.00988	0.6092	0.69225	0.00988
D18S51	0.85057	0.86679	0.4085	0.85057	0.86679	0.4085	0.85057	0.86679	0.4085
D19S433	0.88506	0.86008	0.33301	0.88506	0.86287	0.33811	0.88506	0.86287	0.33811
D20S482	0.74713	0.73942	0.03667	0.74713	0.73942	0.03667	0.8046	0.81975	0.09326
D21S11	0.81609	0.80466	0.52645	0.90805	0.89476	0.5972	0.90805	0.89542	0.62249
PentaD	0.8046	0.84891	0.16973	0.8046	0.84891	0.16973	0.8046	0.84898	0.2107
D22S1045	0.60976	0.69026	0.04232	0.60976	0.69026	0.04232	0.60976	0.69026	0.04232

Table 10.7. iiSNPs sequence-based data generated from Chapter 6.

iiSNP	Total Genotypes	CE data		Sequence data					
		iiSNPs Genotypes	SNPs Frequency	Detected SNPs/ Microhaplotype ^c	Microhaplotype/SNPs Frequency	iiSNP & Variant Reference SNP ^a	iiSNP & Variant GRCh37 Position ^b	Sequence ^d	Strand
rs1490413	174	G	0.55172	AG	0.55172	rs113167966_rs1490413	4367298_4367323	TCAGAACTGCCTGGTGTGGACTGGGCTGATGTGGTCTTTGCAGAACTGGCTGG	Forward
		A	0.44827	AA	0.43678	rs113167966_rs1490413	4367298_4367323	TCAGAACTGCCTGGTGTGGACTGGGCTGATGTGGTCTTTGCAAACTGGCTGG	Forward
				GA	0.01149	rs113167966_rs1490413	4367298_4367323	TCAGAACTGCCTGGTGTGGCTGGGCTGATGTGGTCTTTGCAAACTGGCTGG	Forward
rs560681	174	A	0.50575	ACA	0.48276	rs560681_rs186550433_rs60615385	160786670_160786675_160786688	TCCATCTCTATTTACTCAGGTCACAGGACCTTGGGGCCTCAAGAGTT	Forward
				ACG	0.02299	rs560681_rs186550433_rs60615385	160786670_160786675_160786688	TCCATCTCTATTTACTCAGGTCACAGGACCTTGGGGCCTCAAGAGTT	Forward
		G	0.49425	GCA	0.49425	rs560681_rs186550433_rs60615385	160786670_160786675_160786688	TCCATCTGTTTACTCAGGTCACAGGACCTTGGGGCCTCAAGAGTT	Forward
rs1294331	174	A	0.37356	A	0.37356	rs1294331	233448413	AGTATAGTTATGGATTTTATTGAATTTTTG	Reverse
		G	0.62644	G	0.62644	rs1294331	233448413	AGTGTAGTTATGGATTTTATTGAATTTTTG	Reverse
rs10495407	174	A	0.37356	AT	0.37356	rs10495407_rs187062753	238439308_238439314	TCTGCTTCTGGAGATCTCCACTTCTCTTGGTTGCATTGGATTCTCATTGAAAATCCTATTCATTTC	Forward
		G	0.62644	GT	0.62069	rs10495407_rs187062753	238439308_238439314	TCTGCTTCTGGAGATCTCCACTTCTCTTGGTTGCATTGGATTCTCATTGAAAATCCTATTCATTTC	Forward
				GG	0.00575	rs10495407_rs187062753	238439308_238439314	TCTGCTTCTGGAGATCTCCACTTCTCTTGGTTGCATTGGATTCTCATTGAAAATCCTATTCATTTC	Forward
rs891700	174	A	0.48276	GA	0.48276	rs12047255_rs891700	239881878_239881926	GTGTTAGCAGTAAACATTTTCATCAAATTCATTCTTTTTTTTTTGAAGCCTACTGCATAGTTCTAAGG	Forward
		G	0.51724	GG	0.51724	rs12047255_rs891700	239881878_239881926	GTGTTAGCAGTAAACATTTTCATCAAATTCATTCTTTTTTTTTTGAAGCCTGCTGCATAGTTCTAAGG	Forward
rs1413212	174	G	0.73563	G	0.73563	rs1413212	242806797	GGTGGAGCATGGGCATTTC	Reverse
		A	0.26437	A	0.26437	rs1413212	242806797	GGTGGAGCATGGGCATTTC	Reverse

Table 10.7. continued.

rs876724	174	C	0.76436	CCGT	0.01149	rs876724_rs77642176_rs114448669_rs300773	114974_114982_115033_115035	AGTACATTTTTCTCACACTGCTAACTGCCTGCTCATAGATATTCAAATTTA GTAGATGAGATA	Forward
				CGCT	0.01724	rs876724_rs77642176_rs114448669_rs300773	114974_114982_115033_115035	AGTACATTTTTGTACACTGCTAACTGCCTGCTCATAGATATTCAAATTT AGTAGATGACATA	Forward
				CGGC	0.21264	rs876724_rs77642176_rs114448669_rs300773	114974_114982_115033_115035	AGTACATTTTTGTACACTGCTAACTGCCTGCTCATAGATATTCAAATTT AGTAGATGAGACA	Forward
				CGGT	0.52299	rs876724_rs77642176_rs114448669_rs300773	114974_114982_115033_115035	AGTACATTTTTGTACACTGCTAACTGCCTGCTCATAGATATTCAAATTT AGTAGATGAGATA	Forward
		T	0.23563	TGGT	0.23563	rs876724_rs77642176_rs114448669_rs300773	114974_114982_115033_115035	AGTATATTTTTGTACACTGCTAACTGCCTGCTCATAGATATTCAAATTT AGTAGATGAGATA	Forward
rs1109037	174	A	0.46551	ACA	0.1954	rs1109037_rs183533496_rs1109038	10085722_10085753_10085786	CCAGTTTCTCCAGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
				ACG	0.27011	rs1109037_rs183533496_rs1109038	10085722_10085753_10085786	CCAGTTTCTCCAGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
		G	0.5345	GCA	0.1092	rs1109037_rs183533496_rs1109038	10085722_10085753_10085786	CCAGTTTCTCCGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
				GCG	0.40805	rs1109037_rs183533496_rs1109038	10085722_10085753_10085786	CCAGTTTCTCCGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
				GGCG	0.00575	rs1109037_NA_rs183533496_rs1109038	10085722_10085752_10085753_10085786	CCAGTTTCTCCGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
				TGCG	0.00575	rs999755320_rs1109037_rs183533496_rs1109038	10085721_10085722_10085753_10085786	CCAGTTTCTCGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
				TGCA	0.00575	rs999755320_rs1109037_rs183533496_rs1109038	10085721_10085722_10085753_10085786	CCAGTTTCTCGAGTGAAAGACTTTCATCTCGCACTGGCACGACCTTGAG ACCCCGGTTCTGATGAACTGGGAGG	Forward
rs993934	174	T	0.53449	TA	0.52874	rs993934_NA	124109213_124109179	TTATGAACTAAGCTAATATACTCTGGAGACTGTTATCACAATACTTTGCTCTA TTGCCTTACAAAGCAAA	Reverse
				TG	0.00575	rs993934_NA	124109213_124109179	TTATGAACTAAGCTAATATACTCTGGAGACTGTTATCACAATACTTTGCTCTA TTGCCTTACAAAGCAAA	Reverse
		C	0.46552	CA	0.46552	rs993934_NA	124109213_124109179	TTACGAACTAAGCTAATATACTCTGGAGACTGTTATCACAATACTTTGCTCTA TTGCCTTACAAAGCAAA	Reverse
rs12997453	174	G	0.72988	CG	0.72988	rs72883670_rs12997453	182413238_182413259	TTTTATGCTTTAAAGATACAGGTTATCTGTATTACATTGGGTTTTACCTACCT T	Forward
				CA	0.17815	rs72883670_rs12997453	182413238_182413259	TTTTATGCTTTAAAGATACAGGTTATCTGTATTACATTGAGTTTTACCTACCT T	Forward
		A	0.27011	TA	0.09195	rs72883670_rs12997453	182413238_182413259	TTTTATGCTTTAAAGATATAGGTTATCTGTATTACATTGAGTTTTACCTACCT T	Forward
rs907100	174	G	0.54023	GG	0.54023	rs907100_rs11689319	239563579_239563597	TGATGCCCTGGCATGAAAGAAGGCTCCAAGTGGCTCTTTCTGTGTTTTCC CAAGCTTGGAAAAG	Forward
				CG	0.24138	rs907100_rs11689319	239563579_239563597	TGATGCCCTGGCATCAAGAAGGCTCCAAGTGGCTCTTTCTGTGTTTTCC AAGCTTGGAAAAG	Forward
		C	0.45977	CA	0.21839	rs907100_rs11689319	239563579_239563597	TGATGCCCTGGCATCAAGAAGGCTCCAAGTGGCTCTTTCTGTGTTTTCC AAGCTTGGAAAAG	Forward

Table 10.7. continued.

rs1357617	172	T	0.80814	TC	0.80814	rs1357617_rs145504328	961782_961743	ATTGGGGTGCTGGTCATGTTCTTATCAGCTATCCCTATTCTAATCTCTAA TTGGGCTTTCAATTC	Reverse
		A	0.19186	AC	0.19186	rs1357617_rs145504328	961782_961743	ATTGGGGAGCTTGGTCATGTTCTTATCAGCTATCCCTATTCTAATCTCTAA TTGGGCTTTCAATTC	Reverse
rs4364205	174	T	0.4023	T	0.4023	rs4364205	32417644	TTCCAGTCTAGATATCCACCCATGAGAAATATATCCACAAAATATTGAGC AA	Forward
		G	0.5977	G	0.5977	rs4364205	32417644	TTCCAGTCTAGATATCCACCCATGAGAAATATATCCACAAAAGATTGAGC AA	Forward
rs2399332	174	A	0.35058	AAG	0.00575	rs2399332_rs2399333_rs2399334	110301126_110301062_110301025	ATTAAAAATCAGCAAATACATGAACAGATACTTCTCAAAAAGAGCCAAC AAACTTGAAAAAATGTTCAAAGTCACTAATCATGAGAGAAATGTAATCAA AACCCAGT	Reverse
				ACA	0.00575	rs2399332_rs2399333_rs2399334	110301126_110301062_110301025	ATTAAAAATCAGCAAATACATGAACAGATACTTCTCAAAAAGAGCCAAC AAACTTGAAAAAATGTTCAAAGTCACTAATCATGAGAGAAATGTAATCAA AACCCAAAT	Reverse
				ACG	0.33908	rs2399332_rs2399333_rs2399334	110301126_110301062_110301025	ATTAAAAATCAGCAAATACATGAACAGATACTTCTCAAAAAGAGCCAAC AAACTTGAAAAAATGTTCAAAGTCACTAATCATGAGAGAAATGTAATCAA AACCCAGT	Reverse
		C	0.64942	CAA	0.61494	rs2399332_rs2399333_rs2399334	110301126_110301062_110301025	ATTAAAAATCAGCAAATACATGAACAGATACTTCTCAAAAAGAGCCAAC AAACTTGAAAAAATGTTCAAAGTCACTAATCATGAGAGAAATGTAATCAA AACCCAAAT	Reverse
				CCG	0.03448	rs2399332_rs2399333_rs2399334	110301126_110301062_110301025	ATTAAAAATCAGCAAATACATGAACAGATACTTCTCAAAAAGAGCCAAC AAACTTGAAAAAATGTTCAAAGTCACTAATCATGAGAGAAATGTAATCAA AACCCAGT	Reverse
rs1355366	174	A	0.62069	AC	0.62069	rs1355366_rs573442621	190806108_190806084	TTCCAGGCCACTGGAGGCTCGAGGATGAGGACTTGCCAAAGCCAGTTGT GGCCTAAGCACATGTGCCAGCTGGAT	Reverse
		G	0.37931	GC	0.37931	rs1355366_rs573442621	190806108_190806084	TTCCAGGCCACTGGAGGCTCGAGGATGGGGACTTGCCAAAGCCAGTTGT GGCCTAAGCACATGTGCCAGCTGGAT	Reverse
rs6444724	174	C	0.3908	C	0.3908	rs6444724	193207380	TTGGAATGGGAGGAAAGGAAAGGACTAAATGTTGAACACTGGTTACCG TGCTAGGTATTACAAACTTG	Forward
		T	0.6092	T	0.6092	rs6444724	193207380	TTGGAATGGGAGGAAAGGAAAGGACTAAATGTTGAACACTGGTTACTG TGCTAGGTATTACAAACTTG	Forward
rs2046361	174	A	0.74138	CA	0.74138	rs116699455_rs2046361	10969083_10969059	TGAACGATCATTTTCAAATAAAAATTAAGAAAGGTGAAGTGTCAACAAT GACCAAAAATAGACAAATC	Reverse
		T	0.25862	CT	0.25862	rs116699455_rs2046361	10969083_10969059	TGAACGATCATTTTCAAATAAAAATTAAGAAAGGTGAAGTGTCAACAATG ACCAAAAATAGACAAATC	Reverse

Table 10.7. continued.

rs279844	174	T	0.3046	TCA	0.28161	rs279844_rs61621790_rs279845	46329655_46329665_46329723	AAGGATAACAGATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC AACTCTCTAAACTTCCTTGATATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC CTTCTGGAGATT	Forward
				TAT	0.02299	rs279844_rs61621790_rs279845	46329655_46329665_46329723	AAGGATAACAGATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC AACTCTCTAAACTTCCTTGATATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC CTTCTGGAGTTT	Forward
		A	0.6954	ACT	0.6954	rs279844_rs61621790_rs279845	46329655_46329665_46329723	AAGGATAACAGATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC AACTCTCTAAACTTCCTTGATATTAAGTTCAGTGTCAATTTGACCAGATATTAATCTCAC CTTCTGGAGTTT	Forward
rs6811238	174	G	0.41379	G	0.41379	rs6811238	169663615	CACTGAGAGGAGAAGACTGTGTGTTTAAAGCCAGGTTTGTAAAGGGTT ATGATAGTATTAA	Forward
		T	0.58621	T	0.58621	rs6811238	169663615	CACTGAGAGGAGAAGACTGTGTGTTTAAAGCCAGGTTTGTAAAGGTT ATGATAGTATTAA	Forward
rs1979255	174	G	0.33908	G	0.33333	rs1979255	190318080	GATGAGCAAGAGTTCCAACGTTCCATGCCTGACCAACACAAGCTA	Reverse
				GT	0.00575	rs1979255_rs190924736	190318080_190318065	GATGAGCAAGAGTTCCAATGTTCCATGCCTGACCAACACAAGCTA	Reverse
		C	0.66092	C	0.66092	rs1979255	190318080	GATCAGCAAGAGTTCCAACGTTCCATGCCTGACCAACACAAGCTA	Reverse
rs717302	174	A	0.54598	GA	0.54598	rs149072431_rs717302	2879382_2879395	AATAAGCTTTAGAAAGGCATATCGTATTAACTGTGTAGTGAACGTCTGTCTAT TAGGTTTAGCTC	Forward
		G	0.45402	GG	0.45402	rs149072431_rs717302	2879382_2879395	AATAAGCTTTAGAAAGGCATATCGTATTAACTGTGTGGTGAACGTCTGTCTAT TAGGTTTAGCTC	Forward
rs159606	174	G	0.78736	G	0.78736	rs159606	17374898	GTTTCTCATCCTGTTATTATTTGTTTACGCTGTCTCTATATTTTATTCTCTC	Forward
		A	0.21264	A	0.21264	rs159606	17374898	GTTTCTCATCCTGTTATTATTTGTTTACATCTGTCTCTATATTTTATTCTCTC	Forward
rs13182883	174	A	0.40805	A	0.40805	rs13182883	136633338	TGAGGGGAGGGTCCCTTCTGGCCTAGTAGAGGCCTGGCCTGCAGTGAG CATTCAAATCCTCAAGGAACAGGGTGGGAGGTGGGACAAAGGCAGGAA GAAAGTAACGGAGAGCCTGGGGAGACA	Forward
		G	0.59195	G	0.59195	rs13182883	136633338	TGAGGGGAGGGTCCCTTCTGGCCTAGTAGAGGCCTGGCCTGCAGTGAG CATTCAAATCCTCGAGGAACAGGGTGGGAGGTGGGACAAAGGCAGGAA GAAAGTAACGGAGAGCCTGGGGAGACA	Forward
rs251934	174	T	0.64368	T	0.64368	rs251934	174778678	GAGGCTTTAAGTAGAGTGGGACAGCCAGATATCTACTTCTCATGCCCCA GAC	Reverse
		C	0.35632	C	0.35632	rs251934	174778678	GAGGCTTTAAGTAGAGTGGGACAGCCAGATATCTACTTCTCATGCCCCA GAC	Reverse

Table 10.7. continued.

rs338882	174	C	0.47701	CC	0.47701	rs338882_rs746755126	178690725_178690719	GCCTGTGCACACACGTTTGGGACAAGGGCTGGATTCTTCGGCTGGGAT GTCTCTCAGAGCTCTTGACTTGGTCCCTTTGGCTGGGGCTTCCCGTGAGGT GTGGGCTGCGCCACG	Reverse
		T	0.52299	TC	0.52299	rs338882_rs746755126	178690725_178690719	GCCTGTGCATACACACGTTTGGGACAAGGGCTGGATTCTTCGGCTGGGAT GTCTCTCAGAGCTCTTGACTTGGTCCCTTTGGCTGGGGCTTCCCGTGAGGT GTGGGCTGCGCCACG	Reverse
rs13218440	174	G	0.62069	GT	0.62069	rs13218440_rs1011515	12059954_12060047	CCTCTGGGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTT GAGGTACAGCTCCCACTGCCTCTGAGTGGCCCTCCATGAAAATGCCTCATG TCTCTGTGCCCTAAACTGTAGG	Forward
		A	0.37931	AT	0.37931	rs13218440_rs1011515	12059954_12060047	CCTCTGAGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTT GAGGTACAGCTCCCACTGCCTCTGAGTGGCCCTCCATGAAAATGCCTCATG TCTCTGTGCCCTAAACTGTAGG	Forward
rs1336071	174	G	0.4023	GT	0.37931	rs1336071_rs73756355	94537255_94537233	AAAAAGCATCAGGTTAAAACAAAGATAAGAAAAATAGAAATTCAGATGG ACAAAACAG	Reverse
				GA	0.02299	rs1336071_rs73756355	94537255_94537233	AAAAAGCATCAGGTTAAAACAAAGATAAGAAAAATAGAAATTCAGATGG ACAAAACAG	Reverse
		A	0.5977	AT	0.5977	rs1336071_rs73756355	94537255_94537233	AAAAAGCATCAGATTA AAACAAAGATAAGAAAAATAGAAATTCAGATGG ACAAAACAG	Reverse
rs214955	174	G	0.43103	GC	0.43103	rs214955_rs117020413	152697706_152697675	CCTGCCATTAATTTTTGCGTCACCTTTTCAGTCTTTGTTGCAAGCATCAA GGGCTGCCAGAATAAAGAA	Reverse
		A	0.56897	AC	0.56322	rs214955_rs117020413	152697706_152697675	CCTGCCATTAATTTTTGCGTCACCTTTTCAGTCTTTGTTGCAAGCATCAA GGGCTGCCAGAATAAAGAA	Reverse
				AA	0.00575	rs214955_rs117020413	152697706_152697675	CCTGCCATTAATTTTTGCGTCACCTTTTCAGTCTTTGTTGCAAGCATCAA GGGCTGCCAGAATAAAGAA	Reverse
rs727811	174	A	0.42529	AA	0.42529	rs727811_rs1390470	165045334_165045290	CAACGACTTAAATCATCTGCATCTCCAGCAATCTCATGATTTCACTCACT TTATATTTTGCT	Reverse
		C	0.57472	CA	0.56897	rs727811_rs1390470	165045334_165045290	CAACGACTTAAATCATCTGCATCTCCAGCAATCTCATGATTTCACTCACT TATATTTTGCT	Reverse
				CG	0.00575	rs727811_rs1390470	165045334_165045290	CAACGACTTAAATCATCTGCATCTCCAGCAATCTCATGATTTCACTCACT TGATTTTGCT	Reverse
rs6955448	174	T	0.70115	AGAT	0.70115	rs6950322_rs140855431_rs143117431_r s6955448	4310317_4310327_4310349_4310365	TAAGTTGGATAGGTGATGCAAAGCCCTGCCGTTACTACTATTGTTGAAAA GTTTCATGGGAGAAAGTTGATAA	Forward
		C	0.29885	GGAC	0.29885	rs6950322_rs140855431_rs143117431_r s6955448	4310317_4310327_4310349_4310365	TAAGTTGGATAGGTGATGCAAAGCCCTGCCGTTACTACTATTGTTGAAAA GTTTCACGGGAGAAAGTTGATAA	Forward
rs917118	174	C	0.63793	C	0.63793	rs917118	4457003	CCATGAAGATGGAGTCAACATTTTACAAGACGCTCGTTGACCTCAGTCATC TCTTACCACCTTGCTC	Forward
		T	0.36207	T	0.35632	rs917118	4457003	CCATGAAGATGGAGTCAACATTTTACAAGATGCTCGTTGACCTCAGTCATC TCTTACCACCTTGCTC	Forward
				CT	0.00575	rs1431710768_rs917118	4456981_4457003	CCATGAAGCTGGAGTCAACATTTTACAAGATGCTCGTTGACCTCAGTCATC TCTTACCACCTTGCTC	Forward

Table 10.7. continued.

rs321198	174	T	0.2931	TTT	0.2931	rs147098090_rs321198_rs186446125	137029829_137029838_137029840	ATACAATTCTCAAATGAAATAACTAAATAAGGAAGCTGTGTTCTTTCTCCT ACACACAGGCTTCAGGTTACCTGTTTTCTTTTGATTCCACTTCTGTGTGA AGCAAGCAGT	Forward
		C	0.7069	TCT	0.7069	rs147098090_rs321198_rs186446125	137029829_137029838_137029840	ATACAATTCTCAAATGAAATAACTAAATAAGGAAGCTGTGTTCTTTCTCCT ACACACAGGCTTCAGGTTACCTGTTTTCTTTTGATTCCACTTCCGTGTGA AGCAAGCAGT	Forward
rs737681	174	C	0.52874	CGC	0.52874	rs535138178_rs553538445_rs737681	155990783_155990787_155990813	CCGGGGAGCTCTGCACAAGCCGCAGGGTACATGTGAGGCCATCCACC CTCATCCCTGGGCCAGGTGCAGTCTCTC	Forward
		T	0.47126	CGT	0.47126	rs535138178_rs553538445_rs737681	155990783_155990787_155990813	CCGGGGAGCTCTGCACAAGCCGCAGGGTACATGTGAGGCCATCCACC TTCATCCCTGGGCCAGGTGCAGTCTCTC	Forward
rs763869	174	T	0.61494	T	0.61494	rs763869	1375610	TGTTTATATTATTCTAACTCAATTGCATTACATT	Reverse
		C	0.38506	C	0.38506	rs763869	1375610	TGTTTATATTATTCTAACTCAATTGCATTACATT	Reverse
rs10092491	174	T	0.2931	TG	0.2931	rs10092491_rs757115963	28411072_28411106	AAACTGAAGTTTTCTTATAGAGATTTATCCTAGTTAGTTGGGGGATACT GGTTGGGCCGAAA	Forward
		C	0.7069	CG	0.7069	rs10092491_rs757115963	28411072_28411106	AAACTGAAGTTTTCTTATAGAGATTTATCCTAGTTAGTTGGGGGATACT GGTTGGGCCGAAA	Forward
rs2056277	174	C	0.74138	CCG	0.68391	rs149318691_rs2056277_rs539552840	139399078_139399116_139399117	AAGCATGTGTCAACCGCCAACTGGGTGTTAGGGAGACAGGCATGAATGA GACGGGA	Forward
			0.05747	TCG	0.05747	rs149318691_rs2056277_rs539552840	139399078_139399116_139399117	AAGCATGTGTCAACTGCCAACTGGGTGTTAGGGAGACAGGCATGAATGA GACGGGA	Forward
		0.25862	CTG	0.25862	rs149318691_rs2056277_rs539552840	139399078_139399116_139399117	AAGCATGTGTCAACCGCCAACTGGGTGTTAGGGAGACAGGCATGAATGA GATGGGA	Forward	
rs4606077	174	T	0.37357	TTG	0.10345	rs4606077_rs58774517_rs1869434	144656754_144656764_144656765	TGGGATCTGACTCCCCACAGCCTACCCAAAGCTGGGGAACCTCACTGCC CTTCGGGCTTCAGCACAGGGCTGTCTCCACGCCGGCAGGGCCTGTGCTTT CACTGG	Forward
				TCG	0.26437	rs4606077_rs58774517_rs1869434	144656754_144656764_144656765	TGGGATCTGACTCCCCACAGCCTACCCAAAGCTGGGGAACCTCACTGCC CTTCGGGCTTCAGCACAGGGCTGTCTCCACGCCGGCAGGGCCTGTGCTTT CACTGG	Forward
				TCGC	0.00575	rs4606077_rs58774517_rs1869434_rs975955864	144656754_144656764_144656765_144656787	TGGGATCTGACTCCCCACAGCCTACCCAAAGCTGGGGAACCTCACTGCC CTTCGGGCTTCAGCACAGGGCTGTCTCCACGCCGGCAGGGCCTGTGCTTT ACTGG	Forward
		C	0.62644	CCA	0.62644	rs4606077_rs58774517_rs1869434	144656754_144656764_144656765	TGGGATCTGACTCCCCACAGCCACCCAAAGCTGGGGAACCTCACTGCC CTTCGGGCTTCAGCACAGGGCTGTCTCCACGCCGGCAGGGCCTGTGCTTT CACTGG	Forward

Table 10.7. continued.

rs1015250	174	C	0.2184	ACC	0.04598	rs6475200_rs1015250_rs145984676	1823749_1823774_1823792	AAAGGTTACTAAGTGATGGAGTTAGGAAAAGAACCAGGTGTTTTATTCTGTCCACGTGATTTTCA	Forward
				GCC	0.16667	rs6475200_rs1015250_rs145984676	1823749_1823774_1823792	AAAGGTTACTAAGTGATGGAGTTGGGAAAAGAACCAGGTGTTTTATTCTGTCCACGTGATTTTCA	Forward
				GCCC	0.00575	rs6475200_rs1015250_rs1307278892_rs145984676	1823749_1823774_1823783_1823792	AAAGGTTACTAAGTGATGGAGTTGGGAAAAGAACCAGGTGTTTTATTCTGTCCACGTGATTTTCA	Forward
		G	0.78161	AGC	0.78161	rs6475200_rs1015250_rs145984676	1823749_1823774_1823792	AAAGGTTACTAAGTGATGGAGTTAGGAAAAGAACCAGGTGTTTTATTCTGTCCACGTGATTTTCA	Forward
rs7041158	174	C	0.52299	CA	0.52299	rs7041158_rs11776421	27985938_27985976	TCAATTCCTCTCCAACCAAGACACTTCTACTGGAAAACCTCTCACAATCACATTTATAACAAGGAGAGG	Forward
		T	0.47701	TA	0.47701	rs7041158_rs11776421	27985938_27985976	TCAATTCCTCTCCAACCAAGACACTTCTACTGGAAAACCTCTCACAATCACATTTATAACAAGGAGAGG	Forward
rs1463729	174	A	0.54597	AAC	0.01149	rs553955089_rs1463729_rs114709250	126881464_126881448_126881423	CTTTGGCAGCATACACTCATAGCCACATGCAGCCATTCACCCAAAAGCAGCACAT	Reverse
				AAT	0.53448	rs553955089_rs1463729_rs114709250	126881464_126881448_126881423	CTTTGGCAGCATACACTCATAGCCACATGCAGCCATTCACCCAAAAGCAGCACAT	Reverse
		G	0.45402	AGT	0.45402	rs553955089_rs1463729_rs114709250	126881464_126881448_126881423	CTTTGGCAGCATACACTCATAGCCACATGCAGCCATTCACCCAAAAGCAGCACAT	Reverse
rs1360288	174	C	0.68391	GC	0.68391	rs116412791_rs1360288	128968017_128968063	GGGGAGAGGCTCCCTGCAGCCTCTGGGGAGTAGAGGAGACCTGGGAAGACTGGCTGCATCCCTCAACAGATGCCCC	Forward
		T	0.31609	GT	0.31609	rs116412791_rs1360288	128968017_128968063	GGGGAGAGGCTCCCTGCAGCCTCTGGGGAGTAGAGGAGACCTGGGAAGACTGGCTGCATCCCTCAACAGATGCCCC	Forward
rs10776839	174	T	0.55747	GT	0.55747	rs7037930_rs10776839	137417305_137417308	CAGCTGAGGAGCCCGAGTTTGCCGTGATCAGAGCCCAAGTTGCCCGGTC TGCCCGAGCTCT	Forward
		G	0.44253	AG	0.20115	rs7037930_rs10776839	137417305_137417308	CAGCTGAGGAGCCCAAGTTTGCCGTGATCAGAGCCCAAGTTGCCCGGTC TGCCCGAGCTCT	Forward
				GG	0.24138	rs7037930_rs10776839	137417305_137417308	CAGCTGAGGAGCCCGAGTTTGCCGTGATCAGAGCCCAAGTTGCCCGGTC TGCCCGAGCTCT	Forward
rs826472	174	T	0.4023	TT	0.4023	rs138540224_rs826472	2406617_2406631	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTGTCTCTGTTATATTAGTACCTATCTCTCACCAGGAT	Forward
		C	0.5977	TC	0.55747	rs138540224_rs826472	2406617_2406631	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTGTCTCTGTTATATTAGTACCTATCTCTCACCAGGAT	Forward
				CC	0.04023	rs138540224_rs826472	2406617_2406631	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTGTCTCTGTTATATTAGTACCTATCTCTCACCAGGAT	Forward

Table 10.7. continued.

rs735155	174	G	0.39081	GGGT	0.00575	rs79799511_rs735155_rs373487413_rs7905965	3374199_3374178_3374156_3374154	GACGCCGGTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCGGGAGAGGCTGCTGGGCCGCTGGAGAGGGAGACCTGCTGGCTTCCCGTTGAATTCGGTGACGCTC	Reverse
				GGAT	0.38506	rs79799511_rs735155_rs373487413_rs7905965	3374199_3374178_3374156_3374154	GACGCCGGTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCGGGAGAGGCTGCTGGGCCGCTGGAGAGGGAGACCTGCTGGCTTCCCGTTGAATTCGGTGACGCTC	Reverse
		A	0.6092	GAAT	0.60345	rs79799511_rs735155_rs373487413_rs7905965	3374199_3374178_3374156_3374154	GACGCCGGTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCGGGAGAGGCTGCTGGGCCGCTGGAGAGGGAGACCTGCTGACTTCCCGTTGAATTCGGTGACGCTC	Reverse
				AGAAT	0.00575	rs1003612513_rs79799511_rs735155_rs373487413_rs7905965	3374201_3374199_3374178_3374156_3374154	GACGCCGGTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCGGGAGAGGCTGCTGGGCCACGCTGGAGAGGGAGACCTGCTGACTTCCCGTTGAATTCGGTGACGCTC	Reverse
rs3780962	174	C	0.55172	C	0.55172	rs3780962	17193346	CTGTCCTCACGGGTGAAAGCTGATATCTTGACCTTGTTCATC	Reverse
		T	0.44828	T	0.44828	rs3780962	17193346	CTGTCCTACGGGTGAAAGCTGATATCTTGACCTTGTTCATC	Reverse
rs740598	174	A	0.48276	GAC	0.48276	rs151017734_rs740598_rs189367495	118506883_118506899_118506910	TCAAATAGCAATGGCTCGTCTATGGTTAGTCTCACAGCCACATTCTCAGAAC TGCTCAAACCTGGCCCTGC	Forward
		G	0.51724	GGC	0.51724	rs151017734_rs740598_rs189367495	118506883_118506899_118506910	TCAAATAGCAATGGCTCGTCTATGGTTAGTCTCGCAGCCACATTCTCAGAAC TGCTCAAACCTGGCCCTGC	Forward
rs964681	174	T	0.68966	T	0.68966	rs964681	132698419	GACATGGGCATTTGGGGCCACAGTGCTCAGACAGCAACCTCTGGTTCTTAC CAATCC	Forward
		C	0.31034	C	0.31034	rs964681	132698419	GACACGGGCATTTGGGGCCACAGTGCTCAGACAGCAACCTCTGGTTCTTAC CAATCC	Forward
rs1498553	174	C	0.44253	C	0.44253	rs1498553	5709028	ACTTCAGATGTTCAAAGCCAGACGAGAATAAAAGGATGGCTATGAGATCTA TGGAAGTGCTGAG	Forward
		T	0.55747	T	0.55747	rs1498553	5709028	ACTTCAGATGTTCAAAGCCAGACGAGAATAAAAGGATGGCTATGAGATCTA TGGAAGTGCTGAG	Forward
rs901398	174	T	0.6954	T	0.6954	rs901398	11096221	GTGCAAATAGCTGAATATCAGCCCTGTTGATAGCTAACATTAGT	Forward
		C	0.3046	C	0.3046	rs901398	11096221	GTGCAAATAGCTGAATATCAGCCCTGTTGATAGCTAACATTAGT	Forward
rs10488710	174	C	0.53448	C	0.53448	rs10488710	115207176	TTTACTGTATTAGGAGTCCCACTTGTCTTTTTCTGCAAATGTGGCACTCGG TTTATTTTTA	Reverse
		G	0.46552	G	0.46552	rs10488710	115207176	TTTACTGTATTAGGAGTCCCACTTGTCTTTTTCTGCAAATGTGGCACTCGG TTTATTTTTA	Reverse

Table 10.7. continued.

rs2076848	174	A	0.41954	AG	0.41954	rs2076848_rs7947725	134667546_134667524	GAAATTATTGATAATACACAGGTATCTGGCCTCACCACCAGAAATCAGGG CATGATGGACCTGAAGCGGTCCCGGG	Reverse
		T	0.58046	TG	0.51724	rs2076848_rs7947725	134667546_134667525	GAAATTATTGATAATACACAGGTATCTGGCCTCACCACCAGAAATCAGGG CTTGATGGACCTGAAGCGGTCCCGGG	Reverse
				TA	0.06322	rs2076848_rs7947725	134667546_134667526	GAAATTATTGATAATACACAGGTATCTGGCCTCACCACCAGAAATCAGGG CTTGATGGACCTGAAGCGGTCCCGGG	Reverse
rs2107612	174	A	0.63218	A	0.63218	rs2107612	888320	ACTAATTATGTGTTTTTCTAAATCATATTGTCTACTTTTCTCAAACA	Forward
		G	0.36782	G	0.36782	rs2107612	888320	ACTAATTATGTGTTTTTCTAAATCATATTGTCTACTTTTCTCAAACA	Forward
rs2269355	174	C	0.59195	C	0.59195	rs2269355	6945914	TCCCGAGTCTCTCCACAGTCCC	Forward
		G	0.40805	G	0.40805	rs2269355	6945914	TCCCGAGTCTCTGCACAGTCCC	Forward
rs2920816	166	T	0.60241	TA	0.60241	rs2920816_rs142684512	40863052_40863026	CTATTATCATCTGTAAATAGAAACTTTAATATTTTCTATTTCAAATCCA TGCTATCAATTGCTATATAATCAATTTATTAACTCCAGAAACTGTAAA TAAT	Reverse
		C	0.39759	CC	0.09639	rs2920816_rs142684512	40863052_40863026	CTATTATCATCTGTAAATAGAAACTTTAATATTTTCTATTTCAAATCCA TGCTACCAATTGCTATATAATCAATTTATTAACTCCAGAAACTGTAAA TAAT	Reverse
				CA	0.3012	rs2920816_rs142684512	40863052_40863026	CTATTATCATCTGTAAATAGAAACTTTAATATTTTCTATTTCAAATCCA TGCTACCAATTGCTATATAATCAATTTATTAACTCCAGAAACTGTAAA TAAT	Reverse
rs2111980	174	A	0.64368	AG	0.63793	rs2111980_rs79578959	106328254_106328228	CCTCAAGCTCCAGCCTGGTGCCTCCGCTCCGTGACTCACTGGCAAAGATCT	Reverse
				AA	0.00575	rs2111980_rs79578959	106328254_106328228	CCTCAAGCTCCAGCCTGGTGCCTCCGCTCCGTGACTCACTGGCAAAGATCT	Reverse
		G	0.35632	GG	0.35632	rs2111980_rs79578959	106328254_106328228	CCTTCAGCTCCAGCCTGGTGCCTCCGCTCCGTGACTCACTGGCAAAGATCT	Reverse
rs10773760	174	G	0.31034	AG	0.31034	rs185405753_rs10773760	130761684_130761696	TGTTAGCCGTGGGACCAGCTTCTGTCTGGAAGTTCGTCAAATTGCAGTTAG GTCC	Forward
		A	0.68966	AA	0.68966	rs185405753_rs10773760	130761684_130761696	TGTTAGCCGTGGGACCAGCTTCTGTCTGGAAGTTCGTCAAATTGCAGTTAA GTCC	Forward
rs1335873	174	T	0.57471	T	0.57471	rs1335873	20901724	AGGGTGCAGGTATGTATTGTTGCCGTGGTACTGAGTACATAGTAGGT ACCTGGTACGGGA	Reverse
		A	0.42529	A	0.41954	rs1335873	20901724	AGGGTGCAGGTATGTATTGTTGCCGTGGTAACTGAGTACATAGTAGGT ACCTGGTACGGGA	Reverse
				AT	0.00575	rs1335873_rs1021428287	20901724_20901709	AGGGTGCAGGTATGTATTGTTGCCGTGGTAACTGAGTACATAGTAGGT ACCTGGTACGGGA	Reverse

Table 10.7. continued.

rs1886510	174	C	0.61494	C	0.61494	rs1886510	22374700	GGGTAATTTTAGTAATCTTAAAAGAATAAGCAATATTCGCAAGTGTGGTTGTGAAATCCAGGCGTCA	Reverse
		T	0.38506	T	0.38506	rs1886510	22374700	GGGTAATTTTAGTAATCTTAAAAGAATAAGCAATATTCGCAAGTGTGGTTGTGAAATCCAGGCGTCA	Reverse
rs1058083	174	G	0.5977	G	0.5977	rs1058083	100038233	TTTGCTCAGAGTATCCGAGTTAGCCACTAGG	Forward
		A	0.4023	A	0.4023	rs1058083	100038233	TTTGCTCAGAGTATCCGAGTTAGCCACTAGG	Forward
rs354439	174	A	0.53448	AGA	0.53448	rs354439_rs564750466_rs144284297	106938411_106938406_106938390	TGTTCTGGTGGCTTCTTTCCCTTATGTATCTCTCATGTATCACATTCCTATTAAGCACAATATTCGAAATCATTCACTGTTTCTATCGCAACCTGCAATTGAGAGTTAAGAA	Reverse
		T	0.46552	TGA	0.46552	rs354439_rs564750466_rs144284297	106938411_106938406_106938390	TGTTCTGGTGGCTTCTTTCCCTTATGTATCTCTCATGTATCACATTCCTATTAAGCACAATATTCGAAATCATTCACTGTTTCTATCGCAACCTGCAATTGAGAGTTAAGAA	Reverse
rs1454361	174	A	0.56897	A	0.56897	rs1454361	25850832	GGGAGGAGGAAATACACCCTGAGTCATGTTGTTTCTAAATGGACTACTGAAAAGTGCTTACTGATGATGGAC	Reverse
		T	0.43103	T	0.43103	rs1454361	25850832	GGGAGGAGGAAATACACCCTGAGTCATGTTGTTTCTAAATGGACTACTGAAAAGTGCTTACTGATGATGGAC	Reverse
rs722290	174	G	0.48851	G	0.48851	rs722290	53216723	GTGTTTCAGATTTAGAAATATTGCATATACATACTGAGATGTCTTGGG	Reverse
		C	0.51149	C	0.51149	rs722290	53216723	GTGTTTCAGATTTAGAAATATTGCATATACATACTGAGATGTCTTGGG	Reverse
rs873196	174	T	0.79885	T	0.79885	rs873196	98845531	TGCCCTTTGTAATGTGAACATGCCTGATTGACTCCAACCTGCCAGCCTTGGCATGCTATTGTGAACC	Forward
		C	0.20115	C	0.20115	rs873196	98845531	TGCCCTTTGTAATGTGAACATGCCTGATTGACTCCAACCTGCCAGCCTTGGCATGCTATTGTGAACC	Forward
rs4530059	174	A	0.41379	ATT	0.3908	rs4530059_rs535737392_rs4450333	104769149_104769168_104769223	CAGAGCTCCAGAAGCAACTCCAGCACACAGAGGCGCTGATGTCCTGT CAGGTGCTGCTACTGAGGAAGCCGTTGGTCTCCGGAAGCTCTGTATCCTCAGGAGTGCCGCTGCCTGCCTCC	Forward
				ATC	0.02299	rs4530059_rs535737392_rs4450333	104769149_104769168_104769223	CAGAGCTCCAGAAGCAACTCCAGCACACAGAGGCGCTGATGTCCTGT CAGGTGCTGCTACTGAGGAAGCCGTTGGTCTCCGGAAGCTCTGTATC C C CAGGAGTGCCGCTGCCTGCCTCC	Forward
		G	0.58621	GTC	0.58621	rs4530059_rs535737392_rs4450333	104769149_104769168_104769223	CAGAGCTCCAGAAGCAACTCCAGCACACAGAGGCGCTGATGTCCTGT CAGGTGCTGCTACTGAGGAAGCCGTTGGTCTCCGGAAGCTCTGTATC C C CAGGAGTGCCGCTGCCTGCCTCC	Forward

Table 10.7. continued.

rs1821380	174	G	0.33333	GT	0.33333	rs1821380_rs76591310	39313402_39313380	ATGGAGCCACTGAACTGCAGTGCAAAAATGCAGTAAGGGATACAGATAGA AGAAGGAGAAATGTCAGGAAAAGACAG	Reverse
		C	0.66667	CT	0.66667	rs1821380_rs76591310	39313402_39313380	ATGGAGCCACTGAACTGCAGTGCAAAAATGCAGTAAGGCATACAGATAGA AGAAGGAGAAATGTCAGGAAAAGACAG	Reverse
rs8037429	174	T	0.40805	T	0.40805	rs8037429	53616909	GAGTTATGTAG	Forward
		C	0.59195	C	0.59195	rs8037429	53616909	GAGTTACGTAG	Forward
rs1528460	174	T	0.60345	TA	0.60345	rs1528460_rs764217792	55210705_55210710	CTTAACATATTTAAGATTGAAAATGTTACAGTAAAAGTTTTGTTAATCCT GCATTTGCCAAAC	Forward
		C	0.39655	CA	0.39655	rs1528460_rs764217792	55210705_55210710	CTTAACATATTTAAGACTGAAAATGTTACAGTAAAAGTTTTGTTAATCCT GCATTTGCCAAAC	Forward
rs729172	174	C	0.56897	C	0.56897	rs729172	5606197	AGCCTCATTAATATGACCAAGGCTCCTCGCAGACCGAATGTATGTAACCG	Reverse
		A	0.43103	A	0.43103	rs729172	5606197	AGCCTCATTAATATGACCAAGGCTCCTCGCAGACGAAATGTATGTAACCG	Reverse
rs2342747	174	A	0.27011	AA	0.27011	rs2342747_rs140745596	5868700_5868716	GGAGGAAGAAAACAGAGAGTCTTGACCGTAGAGGGGACAACAAGAAATG AGCTT	Forward
		G	0.72989	GA	0.72989	rs2342747_rs140745596	5868700_5868716	GGAGGAAGAAAACAGAGAGTCTTGACCGTAGAGGGGACAACAAGAAATG AGCTT	Forward
rs430046	174	C	0.59195	CCC	0.59195	rs409820_rs430044_rs430046	78017034_78017045_78017051	AAGGTCATACAATGAATGGTGTGATGTAACCGCTTGGGAGGCGATTCTGA GGGTAGGTGCTGGGTTT	Forward
		T	0.40805	ATT	0.40805	rs409820_rs430044_rs430046	78017034_78017045_78017051	AAGGTCATACAATGAATGGTGTGATGTAACCGCTTGGGAGGTGATTTTGA GGGTAGGTGCTGGGTTT	Forward
rs1382387	174	T	0.81034	TC	0.81034	rs1382387_rs551898660	80106361_80106359	GAAGGAGAAACACCTGAACTTTCAATTCCCTGCAGTGGGCAGATGC	Reverse
		G	0.18966	GC	0.18966	rs1382387_rs551898660	80106361_80106359	GAAGGAGAAACACCTGAACTTTCAAGTCCCTGCAGTGGGCAGATGC	Reverse

Table 10.7. continued.

rs9905977	174	G	0.72988	GGC	0.45977	rs9905977_rs28582109_rs73298992	2919393_2919430_2919461	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGCTCAGTCAGCGTCAAGATTCCTTTGCTTTCCCTCTGCCCTCCCTGGCTTGTCAGCTTTGCCCTCAGGCTTGGCCCTCGTGGCC	Forward
				GGT	0.25862	rs9905977_rs28582109_rs73298992	2919393_2919430_2919461	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGCTCAGTCAGCGTCAAGATTCCTTTGCTTTCCCTCTGCCCTCCCTGGCTTGTCAGCTTTGCCCTCAGGCTTGGCCCTCGTGGCC	Forward
				GAC	0.01149	rs9905977_rs28582109_rs73298992	2919393_2919430_2919461	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGCTCAGTCAGCGTCAAGATTCCTTTGCTTTCCCTCTGCCCTCCCTAGCTTGTCAGCTTTGCCCTCAGGCTTGGCCCTCGTGGCC	Forward
		A	0.27011	AGC	0.27011	rs9905977_rs28582109_rs73298992	2919393_2919430_2919461	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGCTCAGTCAGCATCAAGATTCCTTTGCTTTCCCTCTGCCCTCCCTGGCTTGTCAGCTTTGCCCTCAGGCTTGGCCCTCGTGGCC	Forward
rs740910	174	A	0.81035	AG	0.18966	rs60810599_rs740910	5706584_5706623	AAGTATAACAGTTTGCTAAGTAAGGTGAGTGGTATAATCATATGTTGTAA AAAGCAAAACAAA	Forward
				AA	0.70115	rs60810599_rs740910	5706584_5706623	AAGTATAACAGTTTGCTAAGTAAGGTGAGTGGTATAATCATATATTGTAAA AAGCAAAACAAA	Forward
				GA	0.1092	rs60810599_rs740910	5706584_5706623	AAGGTAAACAGTTTGCTAAGTAAGGTGAGTGGTATAATCATATATTGTAA AAAGCAAAACAAA	Forward
rs938283	174	T	0.77586	T	0.77586	rs938283	77468498	CATACATTGAAGTCTAACCCCTAGTACGTTAGATGTGACCGTATTTGGAGA T	Forward
		C	0.22414	C	0.22414	rs938283	77468498	CATACATTGAAGTCTAACCCCTAGTACGTTAGATGTGACCGCATTTGGAGA T	Forward
rs8078417	174	T	0.43678	CCCGGT	0.00575	rs78650971_rs182919351_rs567092265_rs138630479_rs559299986_rs8078417	80461871_80461880_80461904_80461905_80461913_80461935	GGGACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTCGTGCCCGGTGAA GCTTCAGCTCCCTCCCTGGCTGCTTGGAGCTTCTCAGCTCAGATGC	Forward
				GCCGGT	0.43103	rs78650971_rs182919351_rs567092265_rs138630479_rs559299986_rs8078417	80461871_80461880_80461904_80461905_80461913_80461935	GGGACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTCGTGCCCGGTGAA GCTTCAGCTCCCTCCCTGGCTGCTTGGAGCTTCTCAGCTCAGATGC	Forward
		C	0.56322	GCCAGC	0.01149	rs78650971_rs182919351_rs567092265_rs138630479_rs559299986_rs8078417	80461871_80461880_80461904_80461905_80461913_80461935	GGGACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTCATGCCCGGTGAA GCTTCAGCTCCCTCCCGCTGCTTGGAGCTTCTCAGCTCAGATGC	Forward
				GCCGGC	0.54598	rs78650971_rs182919351_rs567092265_rs138630479_rs559299986_rs8078417	80461871_80461880_80461904_80461905_80461913_80461935	GGGACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTCGTGCCCGGTGAA GCTTCAGCTCCCTCCCGCTGCTTGGAGCTTCTCAGCTCAGATGC	Forward
				GCCGTGC	0.00575	rs78650971_rs182919351_rs567092265_rs138630479_rs559299986_rs8078417	80461871_80461880_80461904_80461905_80461911_80461913_80461935	GGGACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTCGTGCCCGGTGAA GCTTCAGCTCCCTCCCGCTGCTTGGAGCTTCTCAGCTCAGATGC	Forward
rs1493232	174	A	0.75862	A	0.75862	rs1493232	1127986	TTTTGGGTGCTAGGCCACAAAATAACA	Forward
		C	0.24138	C	0.24138	rs1493232	1127986	TTTTGGGTGCTAGGCCACAAAATAACA	Forward
rs9951171	174	G	0.67241	AG	0.65517	rs145524126_rs9951171	9749820_9749879	GGGAGAAAGTCTCGTTGTTCTCTGGATGCAACATGAGAGAGCAGCA CACTGAGCCTTATGGTTGCCCTG	Forward
				CG	0.01724	rs145524126_rs9951171	9749820_9749879	GGGAGAAAGTCTCGTTGTTCTCTGGATGCAACATGAGAGAGCAGCAC ACTGAGCCTTATGGTTGCCCTG	Forward
		A	0.32759	AA	0.32759	rs145524126_rs9951171	9749820_9749879	GGGAGAAAGTCTCGTTGTTCTCTGGATGCAACATGAGAGAGCAGCA CACTGAGCCTTATGGATTGCCCTG	Forward

Table 10.7. continued.

rs1736442	162	A	0.48765	A	0.48765	rs1736442	55225777	TAAGTGGGACAGTTAAGAGAAGGCTGCTTTTGCCTGCCCTGTCAGCAGAGC TCAGCTTGATGTTTCTGTGTGTTGAGTGGGGGGTCTCCATTAGACAAGC G	Reverse
		G	0.51235	G	0.51235	rs1736442	55225777	TAAGTGGGACAGTTAAGAGAAGGCTGCTTTTGCCTGCCCTGTCGGCAGAGC TCAGCTTGATGTTTCTGTGTGTTGAGTGGGGGGTCTCCATTAGACAAGC G	Reverse
rs1024116	174	A	0.61494	AG	0.61494	rs1024116_rs545003555	75432386_75432370	CATACACTTAATAAAGTATGCCCTTTGATTTACTTTTGTCACTTCCCA	Reverse
		G	0.38506	GG	0.38506	rs1024116_rs545003555	75432386_75432370	CATGCACCTTAATAAAGTATGCCCTTTGATTTACTTTTGTCACTTCCCA	Reverse
rs719366	174	T	0.83333	CTAC	0.83333	rs719367_rs719366_rs769312704_rs5634 96574	28463416_28463337_28463332_28463326	GTCACCTTCTCGGCAGCATTAGCAGCTGTGACCACAGCATCTTTAACTCTTT ATTATCCTTTCTGCTTTTCTCTCCATTCTAGTAGCTACTCCTCTGGGGGC CTGCTCTTACTC	Reverse
		C	0.16667	CCAC	0.16667	rs719367_rs719366_rs769312704_rs5634 96574	28463416_28463337_28463332_28463326	GTCACCTTCTCGGCAGCATTAGCAGCTGTGACCACAGCATCTTTAACTCTTT ATTATCCTTTCTGCTTTTCTCTCCATTCTAGTAGCTACTCCTCTGGGGGC CTGCTCTTACTC	Reverse
rs576261	174	A	0.62644	A	0.62644	rs576261	39559807	GTCACCAACCTGGCCTCACAACTCTCTC	Forward
		C	0.37356	C	0.37356	rs576261	39559807	GTCACCAACCTGGCCTCACAACTCTCTC	Forward
rs1031825	174	A	0.41379	A	0.41379	rs1031825	4447483	GTCCTTAACCTATTAATTTTAAATGAGTATTTTATTTATCTAAACCCGAGCA TACTTGAAAGCAGTGATTATATCT	Forward
		C	0.58621	C	0.58621	rs1031825	4447483	GTCCTTAACCTATTAATTTTAAATGAGTATTTTATTTATCTAAACCCGAGCAT ACTTGAAAGCAGTGATTATATCT	Forward
rs445251	174	G	0.58046	CTGG	0.58046	rs117702247_rs535095356_rs445251_rs3 69438	15124957_15124953_15124933_15124893	CATGTGCATTGGAGTTTGTATCAGCAACCACTTGCAGTTTTACATTAATTTG AATTGTAGGCCGGT	Reverse
		C	0.41954	CTCA	0.41379	rs117702247_rs535095356_rs445251_rs3 69438	15124957_15124953_15124933_15124893	CATGTGCATTGGAGTTTGTATCAGCAACCACTTGCAGTTTTACATTAATTTG AATTGTAGGCCAGGT	Reverse
				TTCA	0.00575	rs117702247_rs535095356_rs445251_rs3 69438	15124957_15124953_15124933_15124893	TATGTGCATTGGAGTTTGTATCAGCAACCACTTGCAGTTTTACATTAATTTG AATTGTAGGCCAGGT	Reverse
rs1005533	174	A	0.55172	A	0.55172	rs1005533	39487110	GCAAAAAGCAAGAGCCGTGGAATTAAGTCGCCCTGTTTCAGGGGAGGCAT AAGGAGCTGGAGGACTGGGTGGCTCGGCAGCTTCCCTGGTCTTGCCCTG CACTCTCACCCAGC	Forward
		G	0.44828	G	0.44828	rs1005533	39487110	GCAAAAAGCAAGAGCCGTGGAATTAAGTCGCCCTGTTTCAGGGGAGGCAT AAGGAGCTGGAGGACTGGGTGGCTCGGCAGCTTCCCTGGTCTTGCCCTG CACTCTCACCCAGC	Forward

Table 10.7. continued.

rs1523537	174	C	0.37931	GAC	0.37356	rs538906241_rs77195753_rs1523537	51296121_51296123_51296162	TCTTAATACATTCACTTCTGCATGGGTGGGGTTTCAGTCTGCAACAAGATCT TGTAGGGACGCTATCGCT	Forward
				GGAC	0.00575	NA_rs538906241_rs77195753_rs1523537	51296113_51296121_51296123_51296162	TCTTAATACATGCACTTCTGCATGGGTGGGGTTTCAGTCTGCAACAAGATCT TGTAGGGACGCTATCGCT	Forward
		T	0.62069	GAT	0.60345	rs538906241_rs77195753_rs1523537	51296121_51296123_51296162	TCTTAATACATTCACTTCTGCATGGGTGGGGTTTCAGTCTGCAACAAGATCT TGTAGGGATGCTATCGCT	Forward
				GGT	0.01724	rs538906241_rs77195753_rs1523537	51296121_51296123_51296162	TCTTAATACATTCACTTCTGCGTGGGTGGGGTTTCAGTCTGCAACAAGATCT TGTAGGGATGCTATCGCT	Forward
rs722098	174	A	0.75287	A	0.75287	rs722098	16685598	GAAATATCCTTGATAAGGATTTAAATTTGGATGTGCTGAATTTCTT	Forward
		G	0.24713	G	0.24713	rs722098	16685598	GAAATATCCTTGGTAAGGATTTAAATTTGGATGTGCTGAATTTCTT	Forward
rs2830795	174	A	0.72414	CAA	0.0977	rs12626695_rs79319609_rs2830795	28608125_28608161_28608163	ACTGGGTTACCTCTATAGACATAGGACACACCATTTATTGTCTAAAGAGC AAAGAAGTCCTATTAT	Forward
				TAA	0.62644	rs12626695_rs79319609_rs2830795	28608125_28608161_28608163	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTATTGTCTAAAGAGC AAAGAAGTCCTATTAT	Forward
		G	0.27586	TAG	0.26437	rs12626695_rs79319609_rs2830795	28608125_28608161_28608163	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTATTGTCTAAAGGTC AAAGAAGTCCTATTAT	Forward
				CAG	0.01149	rs12626695_rs79319609_rs2830795	28608125_28608161_28608163	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTATTGTCTAAAGGTC AAAGAAGTCCTATTAT	Forward
rs2831700	174	A	0.46552	A	0.46552	rs2831700	29679687	ATTTGGCTAAACTATTGCCGGAGATAAGTTAGAA	Forward
		G	0.53448	G	0.53448	rs2831700	29679687	ATTTGGCTAAACTATTGCCGGAGATGAGTTAGAA	Forward
rs914165	174	A	0.51724	AC	0.47126	rs914165_rs755095	42415929_42415976	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAGAGGGCAACACCCCTC AGGCAGCTCTGCTGAGCCGCCCCACCCAGTGCAAAAAGAGGTGACTGGTC TGCACTC	Forward
				AG	0.04023	rs914165_rs755095	42415929_42415976	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAGAGGGCAACACCCCTC AGGCAGCTCTGCTGAGCCGCCCCACCCAGTGCAAAAAGAGGTGACTGGTC TGCACTC	Forward
				CAC	0.00575	rs192267746_rs914165_rs755095	42415913_42415929_42415976	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAGAGGGCAACACCCCTC AGGCAGCTCTGCTGAGCCGCCCCACCCAGTGCAAAAAGAGGTGACTGGTC TGCACTC	Forward
		G	0.48276	rs914165_rs755095	42415929_42415976	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAGAGGGCAACACCCCTC AGGCAGCTCTGCTGAGCCGCCCCACCCAGTGCAAAAAGAGGTGACTGGTC TGCACTC	Forward		

Table 10.7. continued.

rs221956	174	C	0.63793	CA	0.63793	rs221956_rs182328575	43606997_43607005	TTCCCTCCAGCTCTCCTCTCCCTTTCTGAGCCCTCAGCAAAGTACTTTAG	Forward
		T	0.36207	TA	0.36207	rs221956_rs182328575	43606997_43607005	TTCCCTCCAGCTCTCCTCTCCCTTTCTGAGCCCTCAGCAAATTGACTTTAG	Forward
rs733164	174	A	0.33908	A	0.33333	rs733164	27816784	CCAGGCTCAGCTTTCAGCCCCAGGTCCACCAACAGGCCATCCCCTTGGA AATTGCCTGACATTCTGAGCCGGGCC	Forward
				TA	0.00575	rs1361542862_rs733164	27816752_27816784	CCAGGCTCAGCTTTCAGCCCCGGTCCACCAACAGGCCATCCCCTTGGA AATTGCCTGACATTCTGAGCCGGGCC	Forward
		G	0.66092	G	0.66092	rs733164	27816784	CCAGGCTCAGCTTTCAGCCCCAGGTCCACCAACAGGCCATCCCCTTGGA AGTTGCCTGACATTCTGAGCCGGGCC	Forward
rs987640	174	T	0.53448	AT	0.53448	rs17793354_rs987640	33559474_33559508	ACAGGTACATTCACCTAACAGGCTCTTTCCACCCCTGTAGAAATACAAA ATAAGACTTAATACAGACGATGG	Forward
		A	0.46551	AA	0.3908	rs17793354_rs987640	33559474_33559508	ACAGGTACATTCACCTAACAGGCTCTTTCCACCCATGTAGAAATACAAA ATAAGACTTAATACAGACGATGG	Forward
				CA	0.07471	rs17793354_rs987640	33559474_33559508	ACCGGTACATTCACCTAACAGGCTCTTTCCACCCATGTAGAAATACAAA ATAAGACTTAATACAGACGATGG	Forward
rs2040411	174	A	0.62644	A	0.62644	rs2040411	47836412	AAGTGCATATTTTCATGA	Forward
		G	0.37356	G	0.37356	rs2040411	47836412	AAGTGCATATTTTCATGA	Forward
rs1028528	174	A	0.63218	A	0.63218	rs1028528	48362290	CTTACTCGACATCACTGTGTGCAGATCCGCGGAGGT	Forward
		G	0.36782	G	0.36782	rs1028528	48362290	CTTACTCGACATCACTGTGTGCAGATCCGCGGAGGT	Forward

Table 10.8. HWE test for iSNPs data generated from Chapter 6. None of the analysed markers showed significant deviation from HWE after Bonferroni correction (P value>0.0004). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. 0.05/121 loci = 0.0004.

iSNP	CE data			Sequence data		
	Ho	He	HW P-value	Ho	He	HW P-value
rs1490413	0.48276	0.49751	0.83053	0.48276	0.50761	0.00827
rs560681	0.54023	0.50289	0.52634	0.54023	0.52515	0.05416
rs1294331	0.47126	0.47073	1	0.47126	0.47073	1
rs10495407	0.35632	0.47073	0.03669	0.35632	0.47791	0.0142
rs891700	0.48276	0.50229	0.83035	0.48276	0.50229	0.83035
rs1413212	0.29885	0.3912	0.04966	0.29885	0.3912	0.04966
rs876724	0.33333	0.3623	0.55123	0.58621	0.62893	0.68571
rs1109037	0.50575	0.50123	1	0.67816	0.71444	0.64579
rs993934	0.44828	0.5005	0.39166	0.45977	0.50661	0.39343
rs12997453	0.26437	0.39658	0.00294	0.28736	0.42954	0.00061
rs907100	0.52874	0.49963	0.66566	0.5977	0.60567	0.43775
rs1357617	0.31395	0.31191	1	0.31395	0.31191	1
rs4364205	0.43678	0.48369	0.3842	0.43678	0.48369	0.3842
rs2399332	0.42529	0.45798	0.63671	0.49425	0.50854	0.51652
rs1355366	0.48276	0.47359	1	0.48276	0.47359	1
rs6444724	0.43678	0.47891	0.49917	0.43678	0.47891	0.49917
rs2046361	0.28736	0.38569	0.02415	0.28736	0.38569	0.02415
rs279844	0.33333	0.42608	0.04603	0.34483	0.43911	0.0971
rs6811238	0.50575	0.48794	0.82585	0.50575	0.48794	0.82585
rs1979255	0.48276	0.44701	0.48093	0.48276	0.45465	0.31
rs717302	0.49425	0.49864	1	0.49425	0.49864	1
rs159606	0.33333	0.33679	1	0.33333	0.33679	1
rs13182883	0.51724	0.48588	0.65618	0.51724	0.48588	0.65618
rs251934	0.43678	0.46136	0.64457	0.43678	0.46136	0.64457
rs338882	0.51724	0.50183	0.83124	0.51724	0.50183	0.83124
rs13218440	0.41379	0.47359	0.2609	0.41379	0.47359	0.2609
rs1336071	0.43678	0.48369	0.37753	0.43678	0.50123	0.14435
rs214955	0.42529	0.49332	0.27208	0.43678	0.49983	0.31959
rs727811	0.45977	0.49166	0.66078	0.47126	0.49824	0.80479
rs6955448	0.3908	0.4215	0.60739	0.3908	0.4215	0.60739
rs917118	0.49425	0.46462	0.64398	0.50575	0.46874	0.32782
rs321198	0.42529	0.41678	1	0.42529	0.41678	1
rs737681	0.55172	0.50123	0.39249	0.55172	0.50123	0.39249
rs763869	0.42529	0.47631	0.36764	0.42529	0.47631	0.36764
rs10092491	0.4023	0.41678	0.7966	0.4023	0.41678	0.7966
rs2056277	0.41379	0.3912	0.78198	0.49425	0.46475	0.86632
rs4606077	0.54023	0.47073	0.17941	0.5977	0.53	0.37152
rs1015250	0.34483	0.34337	1	0.35632	0.36124	0.25345
rs7041158	0.42529	0.50183	0.19486	0.42529	0.50183	0.19486
rs1463729	0.37931	0.49864	0.03223	0.4023	0.511	0.04016
rs1360288	0.33333	0.43485	0.04501	0.33333	0.43485	0.04501
rs10776839	0.47126	0.49625	0.66996	0.52874	0.59391	0.22383
rs826472	0.33333	0.48136	0.00642	0.3908	0.5288	0.02871
rs735155	0.41379	0.47891	0.26176	0.42529	0.49033	0.12284
rs3780962	0.50575	0.49751	1	0.50575	0.49751	1
rs740598	0.3908	0.50229	0.05308	0.3908	0.50229	0.05308
rs964681	0.43678	0.43054	1	0.43678	0.43054	1
rs1498553	0.44828	0.49625	0.38876	0.44828	0.49625	0.38876
rs901398	0.33333	0.42608	0.04716	0.33333	0.42608	0.04716
rs10488710	0.44828	0.5005	0.38993	0.44828	0.5005	0.38993
rs2076848	0.47126	0.48588	0.82688	0.50575	0.55564	0.41465
rs2107612	0.3908	0.46774	0.16255	0.3908	0.46774	0.16255
rs2269355	0.47126	0.48588	0.82701	0.47126	0.48588	0.82701
rs2920816	0.36145	0.48193	0.03786	0.40964	0.54034	0.013
rs2111980	0.45977	0.46136	1	0.47126	0.46874	1
rs10773760	0.29885	0.43054	0.00556	0.29885	0.43054	0.00556
rs1335873	0.52874	0.49166	0.51577	0.52874	0.49651	0.80296

Table 10.8. continued.

iiSNP	CE data			Sequence data		
	Ho	He	HW P-value	Ho	He	HW P-value
rs1886510	0.44828	0.47631	0.65242	0.44828	0.47631	0.65242
rs1058083	0.50575	0.48369	0.82408	0.50575	0.48369	0.82408
rs354439	0.47126	0.5005	0.66668	0.47126	0.5005	0.66668
rs1454361	0.56322	0.49332	0.19598	0.56322	0.49332	0.19598
rs722290	0.51724	0.50262	0.83107	0.51724	0.50262	0.83107
rs873196	0.26437	0.32323	0.09949	0.26437	0.32323	0.09949
rs4530059	0.43678	0.48794	0.37569	0.43678	0.50601	0.12226
rs1821380	0.43678	0.44701	1	0.43678	0.44701	1
rs8037429	0.51724	0.48588	0.65706	0.51724	0.48588	0.65706
rs1528460	0.47126	0.48136	1	0.47126	0.48136	1
rs729172	0.51724	0.49332	0.66668	0.51724	0.49332	0.66668
rs2342747	0.37931	0.39658	0.7848	0.37931	0.39658	0.7848
rs430046	0.37931	0.48588	0.0465	0.37931	0.48588	0.0465
rs1382387	0.31034	0.30915	1	0.31034	0.30915	1
rs9905977	0.35632	0.39658	0.4141	0.5977	0.65238	0.33312
rs740910	0.26437	0.30915	0.17705	0.41379	0.46316	0.12104
rs938283	0.26437	0.34981	0.03127	0.26437	0.34981	0.03127
rs8078417	0.42529	0.49332	0.27453	0.45977	0.5189	0.64078
rs1493232	0.32184	0.36835	0.2475	0.32184	0.36835	0.2475
rs9951171	0.35632	0.44309	0.08771	0.37931	0.46582	0.16165
rs1736442	0.38272	0.5028	0.04468	0.38272	0.5028	0.04468
rs1024116	0.44828	0.47631	0.65321	0.44828	0.47631	0.65321
rs719366	0.21839	0.27938	0.05311	0.21839	0.27938	0.05311
rs576261	0.51724	0.47073	0.49083	0.51724	0.47073	0.49083
rs1031825	0.48276	0.48794	1	0.48276	0.48794	1
rs445251	0.44828	0.48987	0.50983	0.44828	0.49465	0.49305
rs1005533	0.50575	0.49751	1	0.50575	0.49751	1
rs1523537	0.43678	0.47359	0.50004	0.44828	0.49884	0.48639
rs722098	0.33333	0.37426	0.3836	0.33333	0.37426	0.3836
rs2830795	0.36782	0.40183	0.42841	0.49425	0.53106	0.53606
rs2831700	0.49425	0.5005	1	0.49425	0.5005	1
rs914165	0.43678	0.50229	0.2832	0.48276	0.54634	0.0818
rs221956	0.37931	0.46462	0.10579	0.37931	0.46462	0.10579
rs733164	0.42529	0.4508	0.63726	0.42529	0.45465	0.75277
rs987640	0.47126	0.5005	0.66845	0.51724	0.55923	0.5596
rs2040411	0.47126	0.47073	1	0.47126	0.47073	1
rs1028528	0.41379	0.46774	0.35455	0.41379	0.46774	0.35455

Table 10.9. LD test for 122 autosomal markers. A total of 292 pairs (STR-STR, STR-SNP and SNP-SNP) of syntenic markers (q-q, p-p, and p-q) were tested and no LD was detected after Bonferroni correction (P value > Bonferroni-corrected P value 0.0001). The Bonferroni correction was performed by dividing 0.05 by the number of tested markers (the number of tests being performed), i.e. $0.05/292$ pairs = 0.0001.

Chr.1				
Locus 1	Chr	Locus 2	Chr	P value
D1S1656	1	rs1490413	1	0.49812
D1S1656	1	rs560681	1	0.92158
rs1490413	1	rs560681	1	0.21847
D1S1656	1	rs1294331	1	0.0004
rs1490413	1	rs1294331	1	0.26927
rs560681	1	rs1294331	1	0.20475
D1S1656	1	rs10495407	1	0.98907
rs1490413	1	rs10495407	1	0.2437
rs560681	1	rs10495407	1	0.95563
rs1294331	1	rs10495407	1	0.60182
D1S1656	1	rs891700	1	0.84749
rs1490413	1	rs891700	1	0.96219
rs560681	1	rs891700	1	0.34016
rs1294331	1	rs891700	1	0.11079
rs10495407	1	rs891700	1	0.37552
D1S1656	1	rs1413212	1	0.26974
rs1490413	1	rs1413212	1	0.72779
rs560681	1	rs1413212	1	0.66458
rs1294331	1	rs1413212	1	0.93969
rs10495407	1	rs1413212	1	0.4255
rs891700	1	rs1413212	1	0.92589
Chr.2				
Locus 1	Chr	Locus 2	Chr	P value
TPOX	2	D2S441	2	0.92419
TPOX	2	D2S1338	2	0.99902
D2S441	2	D2S1338	2	1
TPOX	2	rs876724	2	0.54976
D2S441	2	rs876724	2	0.94563
D2S1338	2	rs876724	2	0.95084
TPOX	2	rs1109037	2	0.66548
D2S441	2	rs1109037	2	0.97426
D2S1338	2	rs1109037	2	1
rs876724	2	rs1109037	2	0.4241
TPOX	2	rs993934	2	0.78769
D2S441	2	rs993934	2	0.15379
D2S1338	2	rs993934	2	0.99215
rs876724	2	rs993934	2	0.44642
rs1109037	2	rs993934	2	0.61803
TPOX	2	rs12997453	2	0.92084
D2S441	2	rs12997453	2	0.42381
D2S1338	2	rs12997453	2	0.96577
rs876724	2	rs12997453	2	0.42313
rs1109037	2	rs12997453	2	0.4308
rs993934	2	rs12997453	2	0.53139
TPOX	2	rs907100	2	0.32496
D2S441	2	rs907100	2	0.84765
D2S1338	2	rs907100	2	0.17878
rs876724	2	rs907100	2	0.32775
rs1109037	2	rs907100	2	0.09452
rs993934	2	rs907100	2	0.29721
rs12997453	2	rs907100	2	0.2164
Chr.3				
Locus 1	Chr	Locus 2	Chr	P value
D3S1358	3	rs1357617	3	0.69048
D3S1358	3	rs4364205	3	0.01625
rs1357617	3	rs4364205	3	0.21329
D3S1358	3	rs2399332	3	0.99998
rs1357617	3	rs2399332	3	0.203
rs4364205	3	rs2399332	3	0.5772
D3S1358	3	rs1355366	3	0.8982
rs1357617	3	rs1355366	3	0.43829
rs4364205	3	rs1355366	3	0.17517
rs2399332	3	rs1355366	3	0.0964
D3S1358	3	rs6444724	3	0.17982
rs1357617	3	rs6444724	3	0.33372
rs4364205	3	rs6444724	3	0.37217
rs2399332	3	rs6444724	3	0.14662
rs1355366	3	rs6444724	3	0.35106

Table 10.9. continued.

Chr.4					
Locus 1	Chr	Locus 2	Chr	P value	
D4S2408	4	FGA	4	0.97374	
D4S2408	4	rs2046361	4	0.0473	
FGA	4	rs2046361	4	0.58089	
D4S2408	4	rs279844	4	0.88737	
FGA	4	rs279844	4	0.58135	
rs2046361	4	rs279844	4	0.99515	
D4S2408	4	rs6811238	4	0.42173	
FGA	4	rs6811238	4	0.82269	
rs2046361	4	rs6811238	4	0.77136	
rs279844	4	rs6811238	4	0.39323	
D4S2408	4	rs1979255	4	0.86667	
FGA	4	rs1979255	4	0.98527	
rs2046361	4	rs1979255	4	0.47412	
rs279844	4	rs1979255	4	0.78028	
rs6811238	4	rs1979255	4	0.55961	
Chr.5					
Locus 1	Chr	Locus 2	Chr	P value	
D5S818	5	CSF1PO	5	0.61363	
D5S818	5	rs717302	5	0.47702	
CSF1PO	5	rs717302	5	0.77664	
D5S818	5	rs159606	5	0.50772	
CSF1PO	5	rs159606	5	0.92497	
rs717302	5	rs159606	5	0.70771	
D5S818	5	rs13182883	5	0.29308	
CSF1PO	5	rs13182883	5	0.95781	
rs717302	5	rs13182883	5	0.03423	
rs159606	5	rs13182883	5	0.06971	
D5S818	5	rs251934	5	0.63361	
CSF1PO	5	rs251934	5	0.51731	
rs717302	5	rs251934	5	0.42705	
rs159606	5	rs251934	5	0.53025	
rs13182883	5	rs251934	5	0.21562	
D5S818	5	rs338882	5	0.56016	
CSF1PO	5	rs338882	5	0.36063	
rs717302	5	rs338882	5	0.13621	
rs159606	5	rs338882	5	0.7189	
rs13182883	5	rs338882	5	0.40301	
rs251934	5	rs338882	5	0.68537	
Chr.6					
Locus 1	Chr	Locus 2	Chr	P value	
D6S1043	6	rs13218440	6	0.25599	
D6S1043	6	rs1336071	6	0.03743	
rs13218440	6	rs1336071	6	0.92278	
D6S1043	6	rs214955	6	0.51	
rs13218440	6	rs214955	6	0.72748	
rs1336071	6	rs214955	6	0.2689	
D6S1043	6	rs727811	6	0.94169	
rs13218440	6	rs727811	6	0.55577	
rs1336071	6	rs727811	6	0.76128	
rs214955	6	rs727811	6	0.376	
D6S1043	6	SE33	6	1	
rs13218440	6	SE33	6	0.03156	
rs1336071	6	SE33	6	0.87309	
rs214955	6	SE33	6	0.99894	
rs727811	6	SE33	6	0.98595	
Chr.7					
Locus 1	Chr	Locus 2	Chr	P value	
D7S820	7	rs6955448	7	0.27511	
D7S820	7	rs917118	7	0.33909	
rs6955448	7	rs917118	7	0.5883	
D7S820	7	rs321198	7	0.61806	
rs6955448	7	rs321198	7	0.88742	
rs917118	7	rs321198	7	0.70173	
D7S820	7	rs737681	7	0.37979	
rs6955448	7	rs737681	7	0.62853	
rs917118	7	rs737681	7	0.40061	
rs321198	7	rs737681	7	0.98621	
Chr.8					
Locus 1	Chr	Locus 2	Chr	P value	
D8S1179	8	rs763869	8	0.5584	
D8S1179	8	rs10092491	8	0.80943	
rs763869	8	rs10092491	8	0.94969	
D8S1179	8	rs2056277	8	0.06427	
rs763869	8	rs2056277	8	0.58136	
rs10092491	8	rs2056277	8	0.03819	
D8S1179	8	rs4606077	8	0.73678	
rs763869	8	rs4606077	8	0.05069	
rs10092491	8	rs4606077	8	0.54565	
rs2056277	8	rs4606077	8	0.37239	

Table 10.9. continued.

Chr.9				
Locus 1	Chr	Locus 2	Chr	P value
D9S1122	9	rs1015250	9	0.82415
D9S1122	9	rs7041158	9	0.54512
rs1015250	9	rs7041158	9	0.58814
D9S1122	9	rs1463729	9	0.78652
rs1015250	9	rs1463729	9	0.56888
rs7041158	9	rs1463729	9	0.9958
D9S1122	9	rs1360288	9	0.58519
rs1015250	9	rs1360288	9	0.12583
rs7041158	9	rs1360288	9	0.48289
rs1463729	9	rs1360288	9	0.13896
D9S1122	9	rs10776839	9	0.35676
rs1015250	9	rs10776839	9	0.17131
rs7041158	9	rs10776839	9	0.12548
rs1463729	9	rs10776839	9	0.7046
rs1360288	9	rs10776839	9	0.95171
Chr.10				
Locus 1	Chr	Locus 2	Chr	P value
D10S1248	10	rs826472	10	0.71276
D10S1248	10	rs735155	10	0.81945
rs826472	10	rs735155	10	0.9813
D10S1248	10	rs3780962	10	0.84618
rs826472	10	rs3780962	10	0.82519
rs735155	10	rs3780962	10	0.008
D10S1248	10	rs740598	10	0.61305
rs826472	10	rs740598	10	0.07554
rs735155	10	rs740598	10	0.59965
rs3780962	10	rs740598	10	0.18169
D10S1248	10	rs964681	10	0.24918
rs826472	10	rs964681	10	0.11341
rs735155	10	rs964681	10	0.64347
rs3780962	10	rs964681	10	0.55506
rs740598	10	rs964681	10	0.85486
Chr.11				
Locus 1	Chr	Locus 2	Chr	P value
TH01	11	rs1498553	11	0.44699
TH01	11	rs901398	11	0.40305
rs1498553	11	rs901398	11	0.43619
TH01	11	rs10488710	11	0.06893
rs1498553	11	rs10488710	11	0.7525
rs901398	11	rs10488710	11	0.77747
TH01	11	rs2076848	11	0.10053
rs1498553	11	rs2076848	11	0.1509
rs901398	11	rs2076848	11	0.33796
rs10488710	11	rs2076848	11	0.19635
Chr.12				
Locus 1	Chr	Locus 2	Chr	P value
vWA	12	D12S391	12	1
vWA	12	rs2107612	12	0.092
D12S391	12	rs2107612	12	0.11141
vWA	12	rs2269355	12	0.34054
D12S391	12	rs2269355	12	0.01118
rs2107612	12	rs2269355	12	0.09493
vWA	12	rs2920816	12	0.75296
D12S391	12	rs2920816	12	0.91764
rs2107612	12	rs2920816	12	0.26062
rs2269355	12	rs2920816	12	0.27157
vWA	12	rs2111980	12	0.99755
D12S391	12	rs2111980	12	0.99653
rs2107612	12	rs2111980	12	0.24372
rs2269355	12	rs2111980	12	0.32378
rs2920816	12	rs2111980	12	0.4073
vWA	12	rs10773760	12	0.45454
D12S391	12	rs10773760	12	0.54357
rs2107612	12	rs10773760	12	0.46767
rs2269355	12	rs10773760	12	0.2639
rs2920816	12	rs10773760	12	0.01395
rs2111980	12	rs10773760	12	0.53318
Chr.13				
Locus 1	Chr	Locus 2	Chr	P value
D13S317	13	rs1335873	13	0.1836
D13S317	13	rs1886510	13	0.44997
rs1335873	13	rs1886510	13	0.67946
D13S317	13	rs1058083	13	0.20642
rs1335873	13	rs1058083	13	0.8001
rs1886510	13	rs1058083	13	0.98344
D13S317	13	rs354439	13	0.05434
rs1335873	13	rs354439	13	0.4077
rs1886510	13	rs354439	13	0.18306
rs1058083	13	rs354439	13	0.85193

Table 10.9. continued.

Chr.14				
Locus 1	Chr	Locus 2	Chr	P value
rs1454361	14	rs722290	14	0.40185
rs1454361	14	rs873196	14	0.8474
rs722290	14	rs873196	14	0.17717
rs1454361	14	rs4530059	14	0.35381
rs722290	14	rs4530059	14	0.16601
rs873196	14	rs4530059	14	0.29418
Chr.15				
Locus 1	Chr	Locus 2	Chr	P value
PentaE	15	rs1821380	15	0.90582
PentaE	15	rs8037429	15	0.05451
rs1821380	15	rs8037429	15	0.15606
PentaE	15	rs1528460	15	0.01971
rs1821380	15	rs1528460	15	0.62987
rs8037429	15	rs1528460	15	0.72056
Chr.16				
Locus 1	Chr	Locus 2	Chr	P value
D16S539	16	rs729172	16	0.66784
D16S539	16	rs2342747	16	0.40729
rs729172	16	rs2342747	16	0.45601
D16S539	16	rs430046	16	0.93666
rs729172	16	rs430046	16	0.96647
rs2342747	16	rs430046	16	0.32501
D16S539	16	rs1382387	16	0.48523
rs729172	16	rs1382387	16	0.37103
rs2342747	16	rs1382387	16	0.49138
rs430046	16	rs1382387	16	0.61647
Chr.17				
Locus 1	Chr	Locus 2	Chr	P value
D17S1301	17	rs9905977	17	0.31891
D17S1301	17	rs740910	17	0.76952
rs9905977	17	rs740910	17	0.09089
D17S1301	17	rs938283	17	0.05557
rs9905977	17	rs938283	17	0.77258
rs740910	17	rs938283	17	0.21571
D17S1301	17	rs8078417	17	0.31548
rs9905977	17	rs8078417	17	0.86012
rs740910	17	rs8078417	17	0.93825
rs938283	17	rs8078417	17	0.58137
Chr.18				
Locus 1	Chr	Locus 2	Chr	P value
D18S51	18	rs1493232	18	0.16849
D18S51	18	rs9951171	18	0.85345
rs1493232	18	rs9951171	18	0.42172
D18S51	18	rs1736442	18	0.2397
rs1493232	18	rs1736442	18	0.27082
rs9951171	18	rs1736442	18	0.30702
D18S51	18	rs1024116	18	0.0422
rs1493232	18	rs1024116	18	0.69953
rs9951171	18	rs1024116	18	0.22616
rs1736442	18	rs1024116	18	0.6107
Chr.19				
Locus 1	Chr	Locus 2	Chr	P value
D19S433	19	rs719366	19	0.43138
D19S433	19	rs576261	19	0.41759
rs719366	19	rs576261	19	0.93067
Chr.20				
Locus 1	Chr	Locus 2	Chr	P value
D20S482	20	rs1031825	20	0.82481
D20S482	20	rs445251	20	0.97978
rs1031825	20	rs445251	20	0.53286
D20S482	20	rs1005533	20	0.52687
rs1031825	20	rs1005533	20	0.58398
rs445251	20	rs1005533	20	0.20619
D20S482	20	rs1523537	20	0.90156
rs1031825	20	rs1523537	20	0.29564
rs445251	20	rs1523537	20	0.60129
rs1005533	20	rs1523537	20	0.49579

Table 10.9. continued.

Chr.21				
Locus 1	Chr	Locus 2	Chr	P value
D21S11	21	PentaD	21	1
D21S11	21	rs722098	21	0.28111
PentaD	21	rs722098	21	0.16683
D21S11	21	rs2830795	21	0.99158
PentaD	21	rs2830795	21	0.33265
rs722098	21	rs2830795	21	0.16096
D21S11	21	rs2831700	21	0.78972
PentaD	21	rs2831700	21	0.57032
rs722098	21	rs2831700	21	0.99355
rs2830795	21	rs2831700	21	0.17265
D21S11	21	rs914165	21	0.99605
PentaD	21	rs914165	21	0.79375
rs722098	21	rs914165	21	0.49236
rs2830795	21	rs914165	21	0.1051
rs2831700	21	rs914165	21	0.26508
D21S11	21	rs221956	21	0.40982
PentaD	21	rs221956	21	0.13376
rs722098	21	rs221956	21	0.14699
rs2830795	21	rs221956	21	0.0073
rs2831700	21	rs221956	21	0.94269
rs914165	21	rs221956	21	0.80822
Chr.22				
Locus 1	Chr	Locus 2	Chr	P value
D22S1045	22	rs733164	22	0.6213
D22S1045	22	rs987640	22	0.93037
rs733164	22	rs987640	22	0.90829
D22S1045	22	rs2040411	22	0.52838
rs733164	22	rs2040411	22	0.62269
rs987640	22	rs2040411	22	0.08637
D22S1045	22	rs1028528	22	0.35004
rs733164	22	rs1028528	22	0.73114
rs987640	22	rs1028528	22	0.95043
rs2040411	22	rs1028528	22	0.20921

10.6 Appendix 6

10.6.1 Combined exceedance probability Figures

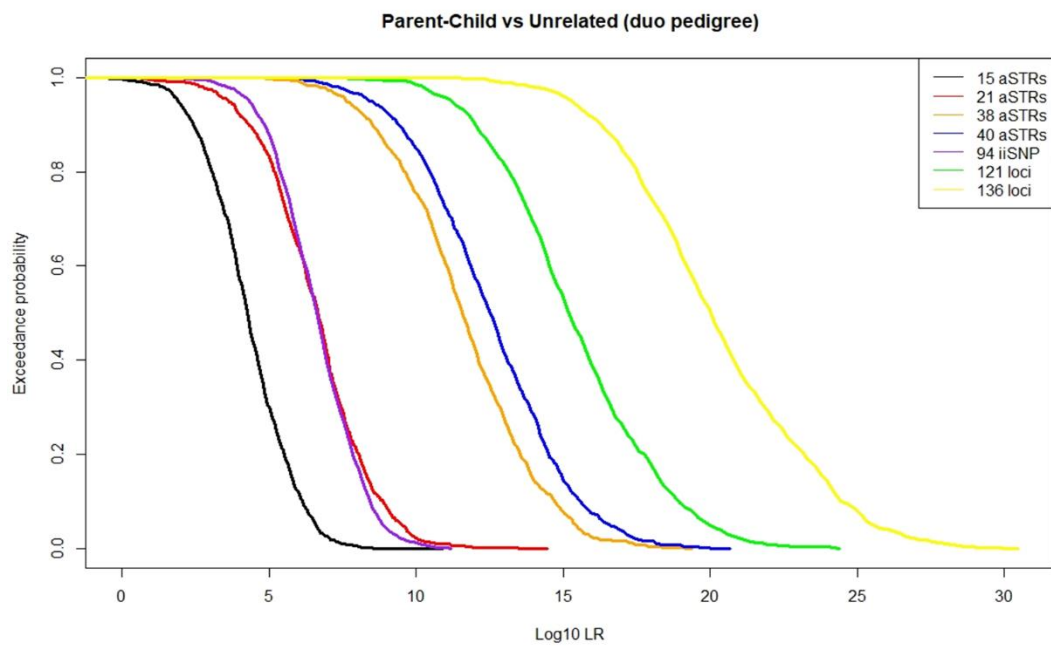


Figure 10.5. Exceedance probability for Parent-child relationship when using seven different marker combinations.

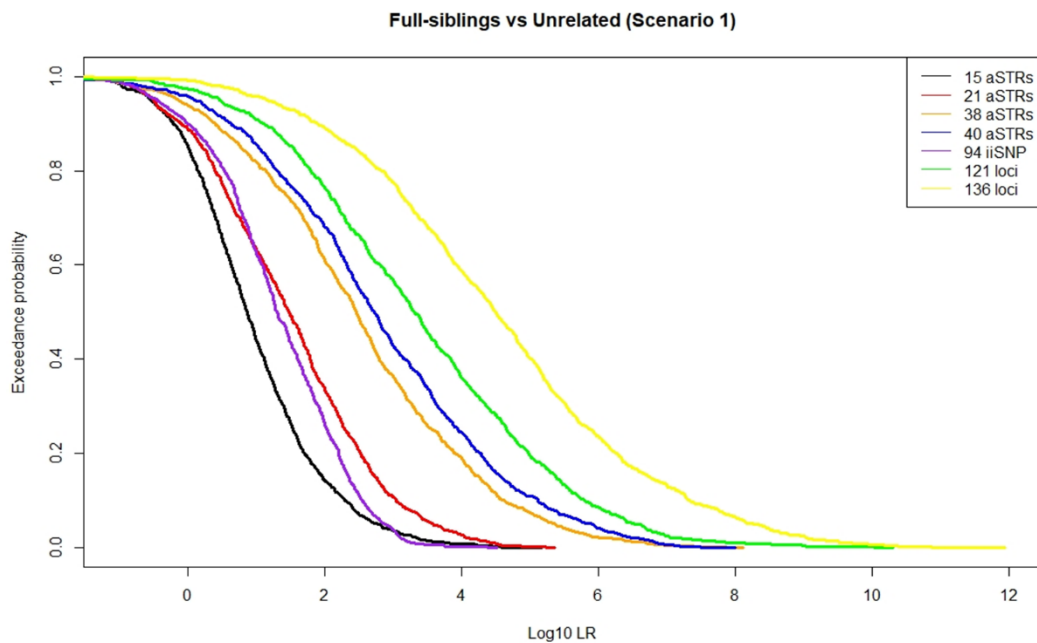


Figure 10.6. Exceedance probability for full-siblings (Scenario 1) relationship when using seven different marker combinations.

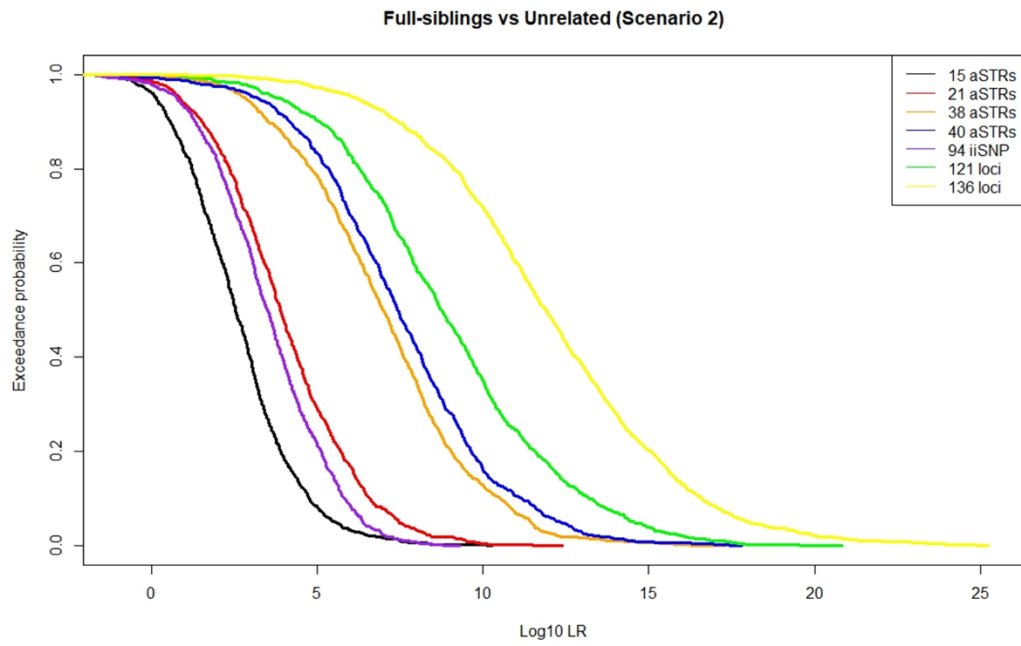


Figure 10.7. Exceedance probability for full-siblings (Scenario 2) relationship when using seven different marker combinations.

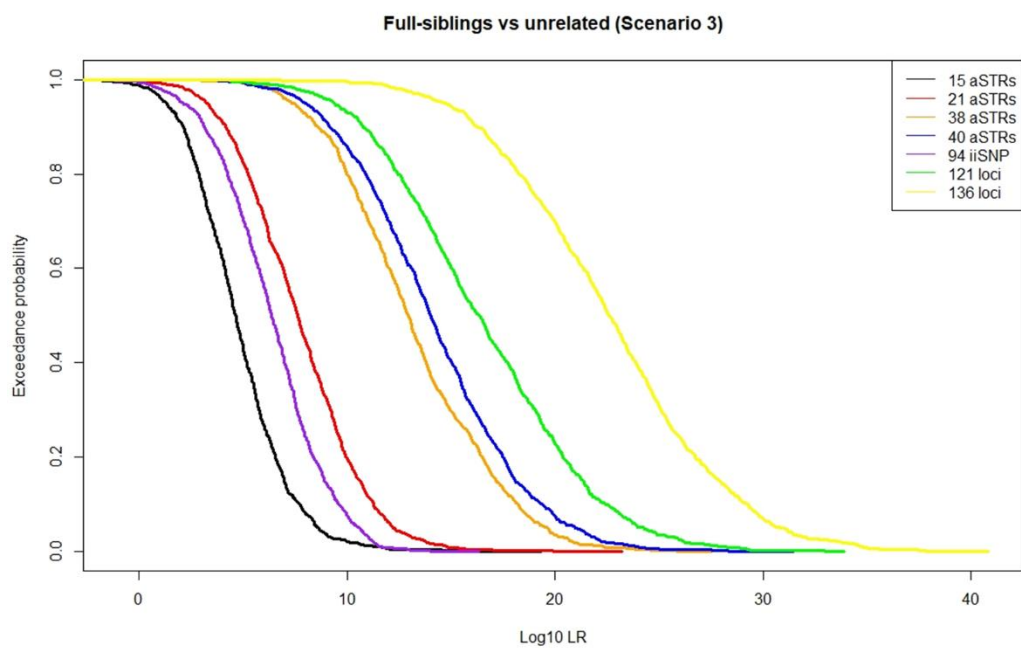


Figure 10.8. Exceedance probability for full-siblings (Scenario 3) relationship when using seven different marker combinations.

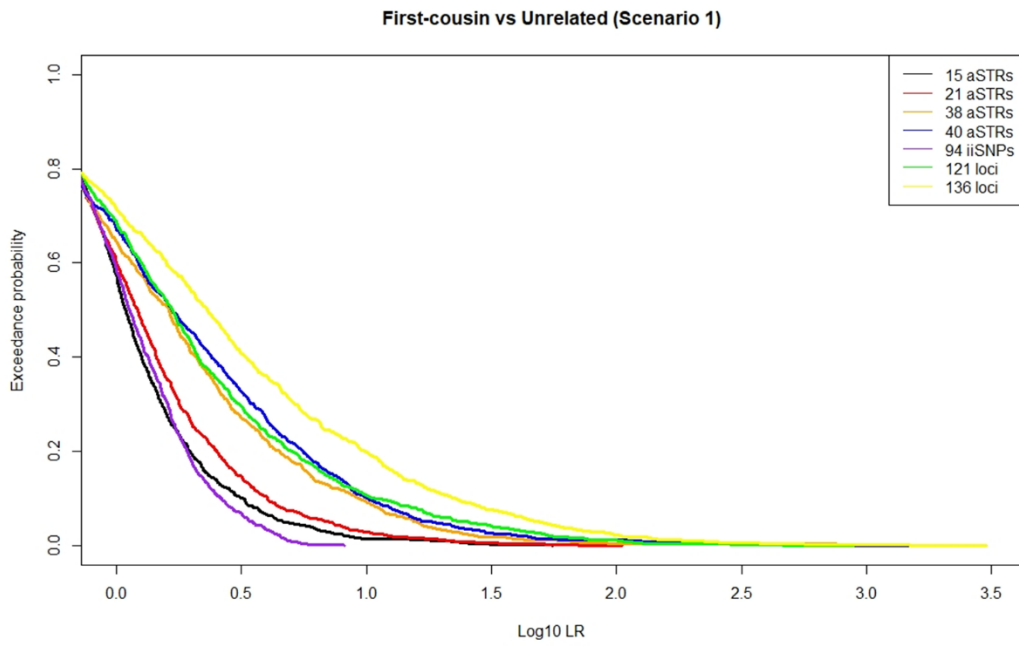


Figure 10.9. Exceedance probability for first-cousin (Scenario 1) relationship when using seven different marker combinations.

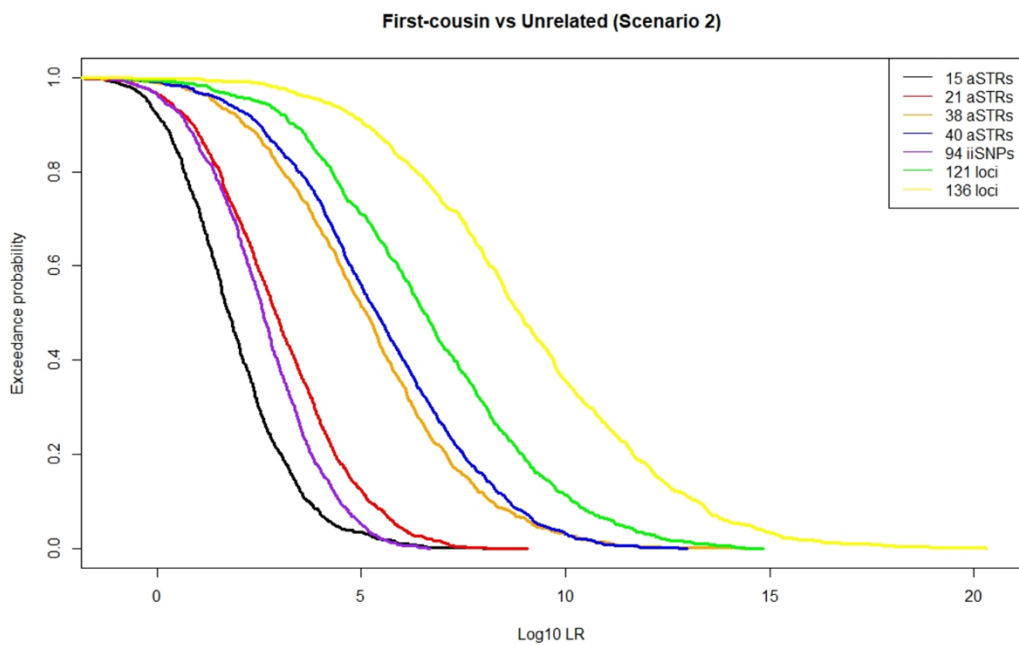


Figure 10.10. Exceedance probability for first-cousin (Scenario 2) relationship when using seven different marker combinations.

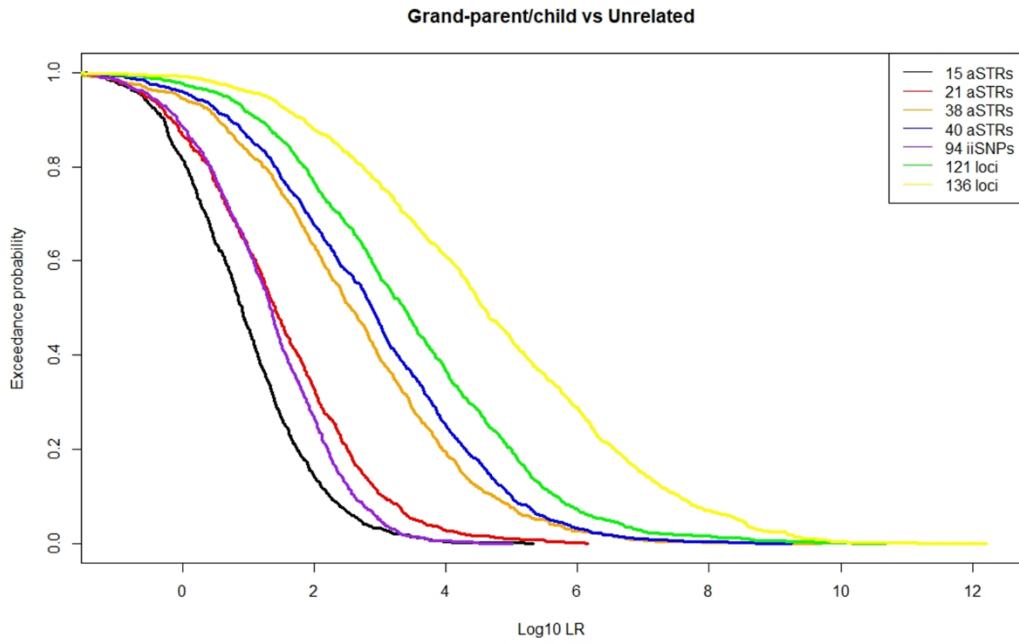


Figure 10.11. Exceedance probability for grand parent/child relationship when using seven different marker combinations.

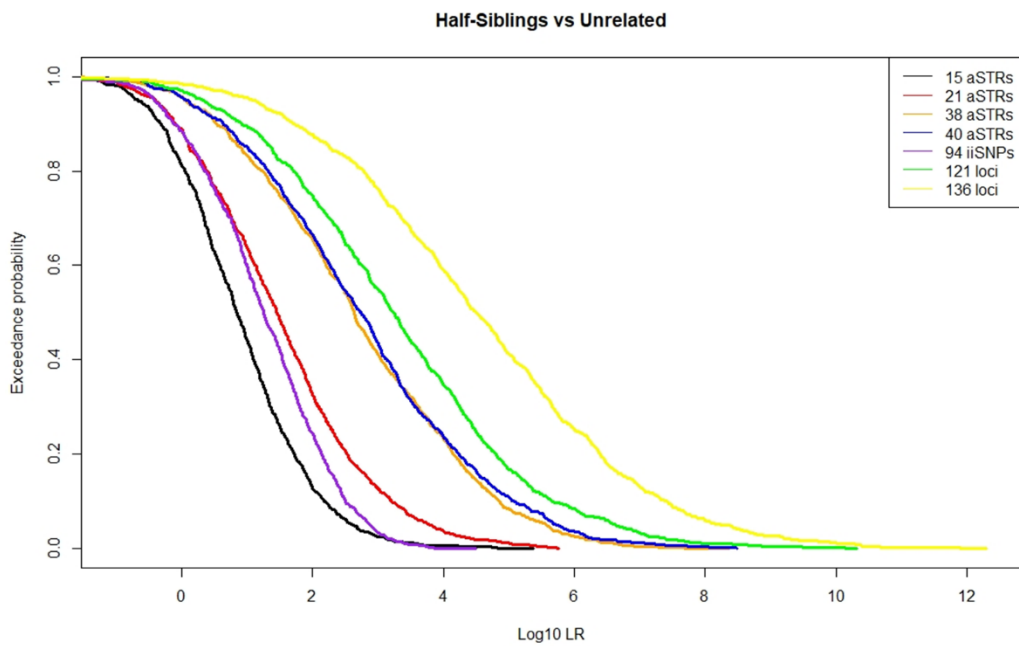


Figure 10.12. Exceedance probability for half-sibling relationship when using seven different marker combinations.

10.6.2 Cumulative genetic map distances (cM) of 95 SNPs.

Table 10.10. The cumulative genetic map distances of 95 SNPs estimated in this study. The 95 SNPs includes 94 iiSNPs and (rs925658351) for D16S539 STR (shaded row). The cumulative genetic map distances were estimated as described by (Phillips *et al.* 2012). The SNP position (bp) on 1000 Genome Browser was used to find the approximate HAP MAP Position (bp) and then to give the cumulative genetic map distance estimation that were eventually used to calculate the RFs.

Chr.	SNP	Position (bp) based on 1000 Genome Browser	HAP MAP data			
			Approximate HAP MAP Position (bp)	Rate (cM/Mb)	Cumulative genetic map distances (cM)	Difference between the two potions
chr1	rs1490413	4367323	4367389	59.435141	10.344332	66
chr1	rs1294331	233448413	233448413	0.118928	252.689344	0
chr1	rs891700	239881926	239881926	0.045405	266.756507	0
chr1	rs1413212	242806797	242806797	2.302109	275.111572	0
chr1	rs10495407	238439308	238439308	0.335417	264.509602	0
chr1	rs560681	160786670	160787725	0.273362	173.518665	1055
chr2	rs876724	114974	115035	0.139591	0.054278	61
chr2	rs1109037	10085722	10085722	27.277036	25.845894	0
chr2	rs993934	124109213	124109213	0.130837	143.138842	0
chr2	rs907100	239563579	239563579	0.037258	261.36756	0
chr2	rs12997453	182413259	182413259	0.506536	196.66925	0
chr3	rs1357617	961782	961782	0.501977	1.267142	0
chr3	rs4364205	32417644	32417644	0.735566	56.460103	0
chr3	rs2399332	110301126	110301126	8.794905	120.166599	0
chr3	rs1355366	190806108	190806108	1.004389	209.79945	0
chr3	rs6444724	193207380	193207380	0.461293	214.02781	0
chr4	rs2046361	10969059	10969059	0.478327	26.495803	0
chr4	rs279844	46329655	46329655	0.093115	68.752478	0
chr4	rs6811238	169663615	169663615	0.232386	174.391264	0
chr4	rs1979255	190318080	190318080	0.141926	213.05529	0
chr5	rs717302	2879395	2879395	0.608829	6.711702	0
chr5	rs159606	17374898	17374898	3.172765	33.526135	0
chr5	rs13182883	136633338	136633338	0.347841	139.768061	0
chr5	rs251934	174778678	174778678	0.150587	191.98624	0
chr5	rs338882	178690725	178690725	1.698655	199.640261	0
chr6	rs13218440	12059954	12059954	0.167491	26.504673	0
chr6	rs1336071	94537255	94537255	1.416482	100.651101	0
chr6	rs214955	152697706	152697706	3.948237	159.848323	0
chr6	rs727811	165045334	165045334	2.845524	180.057073	0
chr7	rs6955448	4310365	4310365	1.246937	6.912354	0
chr7	rs917118	4457003	4457003	0.257833	7.494464	0
chr7	rs321198	137029838	137029838	0.456228	145.377873	0
chr7	rs737681	155990813	155990813	3.482274	181.919589	0
chr8	rs763869	1375610	1376074	0.261994	1.957165	464
chr8	rs10092491	28411072	28411072	0.00252	56.016662	0
chr8	rs2056277	139399116	139399116	2.752499	156.441029	0
chr8	rs4606077	144656754	144656754	1.288918	166.567054	0
chr9	rs1015250	1823774	1823774	0.783254	4.30155	0
chr9	rs7041158	27985938	27985938	0.024507	53.005534	0
chr9	rs1463729	126881448	126881448	1.545209	136.052552	0
chr9	rs1360288	128968063	128968063	1.159908	137.914483	0
chr9	rs10776839	137417308	137417308	0.548701	155.845308	0
chr10	rs826472	2406631	2406750	0.255393	3.568912	119

Table 10.10. continued.

chr10	rs735155	3374178	3374178	0.159617	6.796451	0
chr10	rs3780962	17193346	17193346	0.003591	38.18664	0
chr10	rs740598	118506899	118507219	3.475211	143.730145	320
chr10	rs964681	132698419	132698419	10.412997	175.669404	0
chr11	rs1498553	5709028	5709028	0.669088	11.572163	0
chr11	rs901398	11096221	11096221	1.199891	20.234646	0
chr11	rs10488710	115207176	115207176	4.03968	119.995733	0
chr11	rs2076848	134667546	134667546	4.646889	157.84371	0
chr12	rs2107612	888320	888320	0.206495	2.139891	0
chr12	rs2269355	6945914	6945914	3.626098	17.707304	0
chr12	rs2920816	40863052	40863052	0.003186	56.271503	0
chr12	rs2111980	106328254	106328254	6.280602	124.517852	0
chr12	rs10773760	130761696	130761696	0.542549	168.442499	0
chr13	rs1335873	20901724	20901724	0.179491	2.118193	0
chr13	rs1886510	22374700	22374700	25.742717	4.798954	0
chr13	rs1058083	100038233	100038233	0.350926	94.111308	0
chr13	rs354439	106938411	106938411	0.880698	107.294777	0
chr14	rs1454361	25850832	25850832	0.070597	17.199338	0
chr14	rs722290	53216723	53216723	0.002611	47.502832	0
chr14	rs873196	98845531	98845531	0.012235	104.004219	0
chr14	rs4530059	104769149	104769149	1.968272	114.517458	0
chr15	rs1821380	39313402	39313402	3.849343	53.239677	0
chr15	rs8037429	53616909	53616909	0.533222	64.450111	0
chr15	rs1528460	55210705	55210705	0.105941	66.371524	0
chr16	D16S539 (rs925658351)	86386300	86386367	10.340142	125.578237	67
chr16	rs729172	5606197	5606197	0.494381	11.312582	0
chr16	rs2342747	5868700	5868700	1.253583	11.861336	0
chr16	rs430046	78017051	78017051	0.925856	97.209133	0
chr16	rs1382387	80106361	80106361	0.292771	103.725715	0
chr17	rs9905977	2919393	2919393	9.110377	8.279761	0
chr17	rs740910	5706623	5706623	0.273417	13.408655	0
chr17	rs938283	77468498	77467821	11.920114	120.308115	677
chr17	rs8078417	80461935	80461935	31.907279	127.751349	0
chr18	rs1493232	1127986	1127986	0.079291	3.666872	0
chr18	rs9951171	9749879	9749879	1.956372	28.533917	0
chr18	rs1736442	55225777	55225777	0.321859	74.557154	0
chr18	rs1024116	75432386	75432386	0.054529	112.788928	0
chr19	rs719366	28463337	28463337	2.41979	49.406517	0
chr19	rs576261	39559807	39559807	0.128519	63.836919	0
chr20	rs1031825	4447483	4447483	1.041895	12.795432	0
chr20	rs445251	15124933	15124933	0.001705	35.366478	0
chr20	rs1005533	39487110	39487110	1.169482	58.015382	0
chr20	rs1523537	51296162	51296162	0.565814	77.584173	0
chr21	rs722098	16685598	16686158	1.932588	4.539526	560
chr21	rs2830795	28608163	28608163	0.373499	27.348259	0
chr21	rs2831700	29679687	29679687	0.099294	29.39708	0
chr21	rs914165	42415929	42415929	1.71141	50.554348	0
chr21	rs221956	43606997	43606997	0.900948	54.769216	0
chr22	rs733164	27816784	27816784	2.701817	31.366306	0
chr22	rs987640	33559508	33559508	2.465237	37.654171	0
chr22	rs2040411	47836412	47836412	0.747932	62.887237	0
chr22	rs1028528	48362290	48362290	1.610734	64.136523	0

11 Chapter Eleven: Publications and Participations

11.1 Publications

- 1- Alsafiah, H.; Goodwin, W.; Hadi, S.; Alshaikhi, M. and Wepeba, P. (2017) "Population Genetic Data for 21 Autosomal STR Loci for the Saudi Arabian Population using the GlobalFiler® PCR Amplification Kit", *Forensic Science International: Genetics*, 31 (Supplement C), pp. e59-e61. (Chapter 3).
- 2- Alsafiah, H.; Iyengar, A.; Hadi, S.; Alshlash, W. and Goodwin, W. (2018) "Sequence Data of Six Unusual Alleles at SE33 and D1S1656 STR Loci", *Electrophoresis*, 39, pp. 2471-2476. (Chapter 4).
- 3- Alsafiah, H.; Aljanabi, A.; Hadi, S.; Alturayeif, S. and Goodwin, W. (2019a) "An Evaluation of the SureID 23comp Human Identification Kit for Kinship Testing", *Scientific Reports*, 9 (1), pp. 16859. (Chapter 5).
- 4- Goodwin, W.; Alsafiah, H. and Al-Janabi, A. (2020) "Chapter 31: Short tandem repeat markers applied to the identification of human remains" in *Forensic Science and Humanitarian Action: Interacting with the Dead and the Living*, eds. Parra, R.; Zapico, S. and Ubelaker, D., Wiley. (In press).
- 5- Chapters 6 and 7 will be submitted to *Forensic Science International: Genetics*.

11.2 Participations

- 1- Chapter 6 (Section 6.5.7) was presented as a poster in ISFG2019, Prague. The work was also published in the supplement series of the *Forensic Science International: Genetics*.
Alsafiah, H.; Khubrani, Y.; Sibte, H. and Goodwin, W. (2019b) "Sequence-Based Saudi Population Data for the SE33 Locus", *Forensic Science International: Genetics Supplement Series*. (In press).