

Population genetic analysis of forensic markers in the Kuwaiti population

By

Mahdi Haidar

Centre for Forensic Science

Department of Pure and Applied Chemistry

University of Strathclyde

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Forensic Science
Department of Pure and Applied Chemistry
University of Strathclyde

PhD Thesis
Mahdi Haidar
2020

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50.

Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Date: 30/11/2020

Sign:

A handwritten signature in black ink, consisting of a large, stylized 'X' shape with a smaller, more complex mark to its right.

Abstract

Located in the Arabian Gulf, the nation state of Kuwait is a relatively small country but represents a large geostrategic weight, regionally and globally. However, so far, its genetic diversity has not been extensively studied. This thesis investigates genetic diversity in this population, and evaluates the forensic utility of different classes of forensic markers using both traditional (capillary electrophoresis (CE)) and novel (massively parallel sequencing (MPS)) technologies. Thirty insertion-deletion polymorphism (indel) markers were typed in 150 unrelated Kuwaitis. The genotyping results showed that indel markers were highly discriminative, allowing them to be utilised as a stand-alone typing system for human identification in Kuwait. They can be used when additional markers are required, particularly in challenging relationship and paternity cases. Massively parallel sequencing (MPS) offers an improved system in which much larger numbers of different types of forensic markers can be combined in a single assay, with the additional benefit that each marker can be assayed using shorter amplicons, improving performance when DNA samples are degraded. The major benefit of MPS technology in forensic genetics is its ability to determine the actual nucleotide sequences of forensic markers, rather than just their length. The Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics MPS system was used to sequence 376 unrelated Kuwaiti individuals using 153 markers including 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 identity SNPs. The results showed an increased number of genetic variants detected for these markers based on their sequences, when compared to their lengths, which resulted in substantial increases in the discriminatory power of forensic DNA testing, providing high-resolution genotyping for the analysis of DNA evidence. This highlights the potential value of this sequencing technology. Comparison between CE and MPS systems was carried out by comparing the allele length-based genotypes generated using the ForenSeq™ DNA Signature Prep Kit with genotypes generated using the Promega PowerPlex® Fusion 6C kit, and the haplotypes generated using the Promega PowerPlex® Y23 System. The sequencing results were 100% in concordance with the results obtained using CE for both sets of markers, indicating that the profiles generated by the two technologies can be compared and utilised with confidence. The evaluation of the sequenced markers in this thesis provides

guidelines and information for data interpretation and the data produced delivers valuable reference datasets for forensic analysis in the Kuwaiti population.

Acknowledgments

I would like to express my very great appreciation to my supervisor Dr Penelope Haddrill for her guidance, encouragement and support throughout the duration of my PhD. A special thanks to my colleagues, Dr Hussain Alsaleh and Fatimah Abbas for their motivation and support. This project could not have been a success without all of you.

Most importantly, I would especially like to thank all my family. My parents deserve special thanks for their unconditional support and encouragement throughout this entire process. My wife, for her supportive countless sacrifices to help me get to this point. My children, providing me with the motivation to finish my degree with expediency. Without such a team behind me, I doubt that I would be in this place today.

Publications and presentations

Published articles

M. Haidar, H. Alsaleh, P.R. Haddrill, Evaluation of 30 insertion/deletion polymorphisms as forensic markers in the Kuwaiti population, *Forensic Science International: Genetics Supplement Series* 7(1) (2019) 737-738.

<https://doi.org/10.1007/S00414-019-02180-4>

M. Haidar, H. Alsaleh, P.R. Haddrill, Population genetics of 30 insertion/deletion polymorphisms in the Kuwaiti population, *International Journal of Legal Medicine* 134(3) (2020) 985-986.

<https://doi.org/10.1016/j.fsigss.2019.10.157>

M. Haidar, F. Abbas, H. Alsaleh, P.R. Haddrill, Population genetics and forensic utility of 23 autosomal PowerPlex Fusion 6C STR loci in the Kuwaiti population, *Scientific Reports*, 11(1) (2021) 1865.

<https://doi.org/10.1038/s41598-021-81425-y>

M. Almheiri, S. Sanqoor, M. Haidar, M. Nazir, R. Vijayan, Study of insertion-deletion polymorphisms (indels) in the UAE population, *Forensic Science International: Genetics Supplement Series* 7(1) (2019) 561-563.

<https://doi.org/10.1016/j.fsigss.2019.10.090>

Noora R. Al-Snan, Sabah Shabbir, Sahar S. Baksh, Mashael AlQerainees, Mahdi Haidar, Safia A. Messaoudi & Moiz Bakhiet, Population genetics of 30 insertion/deletion polymorphisms in the Bahraini population, *Scientific Reports*, 11(1) (2021) 6843.

<https://doi.org/10.1038/s41598-021-86386-w>

Poster presentations:

Population genetic data for insertion/deletion polymorphisms in a Kuwaiti population sample. The 4th Scottish Student Forensic Research Symposium, Dundee, UK (March 2018).

Population genetic data for insertion/deletion polymorphisms in a Kuwaiti population sample. Emirates International Forensic Conference & Exhibition, Dubai, UAE (April 2018).

Population genetic data for 27 autosomal STR markers in the Kuwaiti population using massively parallel sequencing. The 28th Congress of the International Society for Forensic Genetics, Czech Republic, Prague (September 2019).

Evaluation of 30 insertion/deletion polymorphisms in the Kuwaiti population. The 28th Congress of the International Society for Forensic Genetics, Czech Republic, Prague (September 2019).

Awards

Selected for Wellcome Trust Advanced Course of Next Generation Sequencing Bioinformatics, at the Wellcome Trust Genome Campus, Hinxton, UK. (October 2017).

Training on the MiSeq FGx Forensic Genomics at the Illumina Solutions Centre, Cambridge, UK. (March 2018).

Table of Contents

Abstract.....	i
Acknowledgments.....	iii
Publications and presentations	iv
Table of Contents.....	vi
List of Figures	xiv
List of Tables	xvi
List of equations.....	xvii
List of abbreviations.....	xviii
Chapter 1 Introduction.....	1
1.1 DNA in forensic science.....	1
1.2 The <i>Homo sapiens</i> genome.....	1
1.3 Genetic variation.....	3
1.4 A brief history of DNA fingerprinting	4
1.5 Short tandem repeat (STR) markers	5
1.5.1 Development of STR multiplexes.....	6
1.5.2 Mini-STRs	9
1.5.3 Basic steps of the polymerase chain reaction-capillary electrophoresis STR genotyping process	11
1.5.3.1 DNA Extraction.....	11
1.5.3.2 DNA quantification.....	12
1.5.3.3 DNA amplification	14
1.5.3.4 Capillary electrophoresis and data interpretation	15
1.5.4 STR artefacts and limitations	17
1.6 Single nucleotide polymorphism (SNP) markers	20
1.6.1 Types of forensic SNP markers	21
1.6.1.1 Individual identification SNPs	21
1.6.1.2 Ancestry informative SNPs.....	22
1.6.1.3 Phenotype informative SNPs	22
1.6.1.4 Lineage informative SNPs	23
1.6.2 SNP genotyping.....	23
1.7 Insertion and deletion (indel) markers	24

1.7.1 A brief history of the discovery of indels	24
1.7.2 Forensic indel multiplexes	26
1.7.3 Commercial indel markers	27
1.8 The Y chromosome	29
1.8.1 Y-chromosome markers.....	30
1.8.1.1 Y-chromosomal short tandem repeat (Y-STR) markers.....	31
1.8.1.2 Y-STRs kits	32
1.8.2 Y-chromosomal single nucleotide polymorphism (Y-SNP) haplogroups	33
1.8.2.1 Y-SNP haplogroup inference and prediction	33
1.9 The X chromosome	34
1.9.1 X-chromosome markers.....	35
1.10 Mitochondrial DNA (mtDNA) markers	36
1.11 DNA sequencing.....	38
1.11.1 Next-generation sequencing (NGS) in forensics	39
1.11.2 Forensic NGS platforms and commercial kits	41
1.11.3 Overview of the MiSeq FGx® workflow	41
1.11.3.1 Template and library preparation	42
1.11.3.2 Cluster generation by bridge amplification	42
1.11.3.3 Sequencing by synthesis	43
1.11.3.4 Indices and paired-end sequencing.....	44
1.11.3.5 Data analysis	45
1.11.4 ForenSeq DNA Signature Prep Kit studies	46
1.11.5 Forensic NGS improvements	47
1.12 Statistical analysis	48
1.12.1 Random match probability	49
1.12.2 Likelihood ratio (LR)	50
1.12.3 The counting method.....	50
1.13 The Kuwaiti population	51
1.13.1 Kuwaiti history and population genetic structure	52
1.13.2 Forensic studies in Kuwait	54
1.14 Aims and objectives of this study	54
Chapter 2 Materials and methods	58
2.1 Chapter overview.....	58

2.2 Ethical approval, sample collection and preparation	58
2.3 DNA extraction.....	59
2.4 DNA quantification.....	59
2.5 Indel marker genotyping using the Investigator® DIPplex kit	60
2.5.1 Amplification.....	60
2.5.2 Fragment analysis	61
2.6 Autosomal STR genotyping using the PowerPlex® Fusion 6C kit.....	61
2.6.1 Amplification.....	61
2.6.2 Fragment analysis	61
2.7 Y-chromosomal STR genotyping using the PowerPlex® Y23 System.....	62
2.7.1 Amplification.....	62
2.7.2 Fragment analysis	62
2.8 DNA sequencing using the MiSeq FGx® with the ForenSeq™ DNA Signature Prep kit	63
2.8.1 Template and library preparation using two-step PCR	63
2.8.2 Library purification and normalisation	63
2.8.3 Library denaturing and sequencing	64
2.9 CE data analysis.....	65
2.10 Sequencing data analysis	65
2.10.1 Universal Analysis Software (UAS).....	65
2.10.2 STRait Razor	65
2.10.3 Manual bioinformatic analysis.....	66
2.11 Statistical analysis	66
2.11.1 Genetic and forensic indices.....	66
2.11.1.1 Random match probability (RMP)	66
2.11.1.2 Power of discrimination (PD).....	66
2.11.1.3 Power of exclusion (PE)	67
2.11.1.4 Typical paternity index (TPI)	67
2.11.1.5 Polymorphism information content (PIC).....	67
2.11.2 Population analysis	68
2.11.3 Principal component analysis (PCA)	68
Chapter 3 Evaluation of 30 forensic indel markers in the Kuwaiti population	69
3.1 Introduction	69
3.2 Aims and objectives	70

3.3 Materials and methods.....	70
3.3.1 Sample collection and preparation.....	70
3.3.2 DNA extraction and quantification	70
3.3.3 Indel amplification and detection.....	71
3.3.4 Data analysis	71
3.3.4.1 Population and forensic statistical analysis	71
3.3.4.2 Inferring individuals' origin	71
3.3.4.3 Population structure analysis	72
3.3.4.4 Population differentiation	72
3.4 Results.....	73
3.4.1 Extraction and quantification	73
3.4.2 Indel genotyping	73
3.4.3 DIPplex artefacts.....	75
3.4.3.1 Allelic imbalance	75
3.4.3.2 Off-ladder allele	76
3.4.3.3 Drop-out alleles.....	76
3.4.4 Allele and genotype frequencies	77
3.4.5 Tests for Hardy-Weinberg equilibrium	79
3.4.6 Tests for linkage disequilibrium	81
3.4.7 Analysis of molecular variance	81
3.4.8 Forensic parameters	82
3.4.9 Inferring individuals' origins at the global level.....	84
3.4.10 Inferring individuals' origins at the regional level	86
3.4.11 Analysis of population structure.....	87
3.4.12 Population differentiation	87
3.5 Discussion.....	88
3.5.1 Allelic imbalance	89
3.5.2 Off-ladder alleles.....	89
3.5.3 Null alleles.....	90
3.5.4 Heterozygosity and Hardy-Weinberg equilibrium.....	90
3.5.5 Linkage disequilibrium	92
3.5.6 Forensic parameters	93
3.5.7 Genotype assignment and ancestry inference	94

3.5.8 Population differentiation	95
3.6 Conclusion	95
Chapter 4 Analysis of 23 Autosomal STR markers in the Kuwaiti population	97
4.1 Introduction	97
4.2 Aims and objectives	98
4.3 Materials and methods	98
4.3.1 Samples and genotyping	98
4.3.2 Statistical analysis	99
4.3.3 Intra-population genetic structure among Kuwaitis	99
4.3.4 Inter-population genetic structure and population relationship	101
4.4 Results and Discussion	102
4.4.1 STR genotyping	102
4.4.2 STR Off-ladder and novel alleles	103
4.4.3 Allele frequencies and forensic statistics.....	107
4.4.4 Statistical analysis of populations	111
4.4.5 Intra-population genetic structure	111
4.4.6 Genetic distance	114
4.5 Conclusion	117
Chapter 5 Massively parallel sequencing of short tandem repeat (STR) markers in the Kuwaiti population	118
5.1 Introduction	118
5.2 Aims and objectives	120
5.3 Materials and methods	120
5.3.1 Sample collection and preparation.....	120
5.3.2 Template and library preparation using two-step PCR	121
5.3.3 Library purification and normalisation	121
5.3.4 Library preparation and sequencing.....	121
5.3.5 Data analysis	122
5.3.6 Bioinformatics analysis	122
5.3.7 SE33 locus analysis.....	123
5.3.8 Concordance test	123
5.3.9 Forensic statistical analysis	123
5.4 Results and discussion	124
5.4.1 Sequencing quality.....	124

5.4.2 Sequencing performance	124
5.4.3 Sample qualification and genotyping	127
5.4.4 Sequence variation	128
5.4.5 Allele frequencies and novelty	129
5.4.6 SE33 locus	131
5.4.7 Concordance test	134
5.4.8 Forensic statistical analysis	134
5.5 Conclusion	138
Chapter 6 Massively parallel sequencing of 94 identity-informative SNPs in Kuwaiti populations	139
6.1 Introduction	139
6.2 Aims and objectives	140
6.3 Materials and methods	141
6.3.1 Data analysis	141
6.3.2 Bioinformatics analysis	142
6.3.3 Forensic statistical analysis	142
6.3.4 Genetic distance	142
6.4 Results and discussion	143
6.4.1 Sequencing performance	143
6.4.2 Sample qualification and genotyping	146
6.4.3 Sequence variation	148
6.4.4 Novel haplotypes	149
6.4.5 Statistical analysis	151
6.4.6 Population differentiation	155
6.5 Conclusion	158
Chapter 7 Analysis of sex chromosome STRs in the Kuwaiti population.	160
7.1 Introduction	160
7.2 Aims and objectives	161
7.3 Materials and methods	161
7.3.1 Y-STR PCR-CE genotyping	161
7.4 Data analysis	162
7.4.1 Bioinformatics analysis	163
7.5 Results and discussion	163
7.5.1 PowerPlex® Y23 kit haplotype genotyping	163

7.5.2 Sequencing performance of X and Y-STRs	166
7.5.3 X-STR allele frequencies and variation.	167
7.5.4 Null allele at the HPRTB X-STR	169
7.5.5 X-STR population genetic and forensic statistics.....	170
7.5.6 Concordance study and CE artefact issues	172
7.5.7 Haplotype diversity and haplogroup prediction	174
7.6 Conclusion	175
Chapter 8 General discussion	177
8.1 Future work.....	184
8.1.1 New indel markers	184
8.1.2 Y-chromosome data.....	186
References	187
Appendix 1 – Participant information sheet and consent	205
A1.1 DIPplex Project.....	205
A1.2 ForenSeq and Fusion 6C projects	209
Appendix 2 – DNA quantification results	213
A2.1 PowerQuant™ System Standard curves	213
A2.2 Autosomal DNA Concentrations.....	214
A2.3 Y-Chromosome Concentrations	214
Appendix 3 – The default stutter filter percentages used by AUS	215
Appendix 4 – Quality metrics from four MiSeq FGx® runs performed to sequence 376 Kuwaiti samples.	216
Appendix 5 – Average depth of coverage (DoC) of 27 autosomal STRs collected using four MiSeq FGx® runs to sequence 376 Kuwaiti samples.	218
Appendix 6 – Allele frequencies of ForenSeq autosomal STR loci	220
Appendix 7 – Allele frequencies of ForenSeq™ DNA Signature Prep Kit novel autosomal STR loci observed in literature.....	237
Appendix 8 – Allele Frequencies of 147 unique SE33 sequences identified in the Kuwaiti population.	238
Appendix 9. Allele frequencies for iiSNPs haplotypes.....	240
Appendix 10. Hardy-Weinberg Equilibrium (HWE) test.	251
Appendix 11. Forensic statistics for the target iiSNPs (iiSNP) and the target iiSNPs with flanking region variants (iiSNP-flank).....	253
Appendix 12. Pairwise F_{ST} genetic distance.	255

Appendix 13. The default stutter filter percentage used for the Y and X-STR markers in the UAS.....	256
Appendix 14. Allele frequencies for X-STR loci.	257
Appendix 15- population statistics for 7 X-STRs in the Kuwaiti population.	269
Appendix 16- Haplotype frequencies for four linkage groups in the Kuwaiti population.	270
Appendix 17. Y-STR haplotypes and their major group.....	272

List of Figures

Figure 1.1. Inheritance of mitochondrial DNA (mtDNA) and sex chromosomes.....	2
Figure 1.2. An outline of the main history of DNA fingerprinting development in the 20 th –21 st century.....	5
Figure 1.3. Classification and variability of short tandem repeats (STRs).....	6
Figure 1.4. Expanded CODIS core short tandem repeat (STR) loci.	8
Figure 1.5. Recently developed commercial short tandem repeat (STR) kits containing the expanded CODIS STR loci.	9
Figure 1.6. AmpFISTR® MiniFiler™ and AmpFISTR® Identifiler™ loci.....	10
Figure 1.7. Four commonly used extraction protocols.	12
Figure 1.8. Quantitative polymerase chain reaction (qPCR) using a TaqMan assay.....	13
Figure 1.9. DNA amplification with polymerase chain reaction (PCR).	15
Figure 1.10. The process of separation and detection of short tandem repeat (STR) fragments by capillary electrophoresis.....	16
Figure 1.11. A segment of a DNA profile represented by an electropherogram generated by capillary electrophoresis.	17
Figure 1.12. A segment of a DNA profile generated from typing old bone sample.....	19
Figure 1.13. Potential nucleotide substitutions that can cause single nucleotide polymorphisms (SNPs).....	20
Figure 1.14. Comparison of typing methodologies of short tandem repeats (STRs), indels and single nucleotide polymorphisms (SNPs).....	24
Figure 1.15. Indel polymorphism classes.....	25
Figure 1.16. Electropherogram of indel typing of a degraded femur bone sample.....	26
Figure 1.17. Electropherogram of indel typing using the DIPplex kit.....	28
Figure 1.18. The structure of the Y chromosome.....	29
Figure 1.19. The locations of commonly used Y chromosome short tandem repeat (Y-STR) markers.	32
Figure 1.20. X chromosome short tandem repeat (X-STR) markers in the Qiagen Investigator® Argus X-12 QS kit.....	36
Figure 1.21. The human mtDNA genome with genes and the control region.	37
Figure 1.22. MiSeq FGx® target amplification and library preparation.	42
Figure 1.23. Cluster generation by bridge amplification.	43
Figure 1.24. Sequencing by synthesis using reversible terminators.	44
Figure 1.25. Indices and paired-end sequencing.....	45
Figure 1.26. The map of Kuwait.	52
Figure 3.1. Electropherogram of the internal size standard (BTO).....	74
Figure 3.2. Electropherogram of the DIPplex allelic ladder.	74
Figure 3.3. Electropherogram of the DIPplex positive control.	75
Figure 3.4. Allelic imbalance at the HLD128 locus.	76
Figure 3.5. Off-ladder allele at the HLD58 locus.	76
Figure 3.6. The partial drop-out allele of HLD97.....	77
Figure 3.7. DIPplex allele frequencies in the Kuwaiti population.	78
Figure 3.8. Pentagonal plot showing the classification of individuals into their population of origin for 30 indel loci in five global populations.	85
Figure 3.9. Principal component analysis (PCA) of five populations using 30 indel markers.....	85

Figure 3.10. Triangular plot showing the classification of individuals into their population of origin for 30 indel loci in three closely related populations.	86
Figure 3.11. Analysis of the structure of seven populations using 30 indel markers.	87
Figure 3.12. 3D-PCA plot of genetic distance based on DIPplex allele frequencies in 28 populations.	88
Figure 4.1. Electropherogram of the PowerPlex® Fusion 6C positive control 2800M.	103
Figure 4.2. PentaE off-ladder alleles.	104
Figure 4.3. SE33 off-ladder alleles.	105
Figure 4.4. PentaD off-ladder allele.	106
Figure 4.5. D22S1045 off-ladder allele.	106
Figure 4.6. D18S51 off-ladder allele.	107
Figure 4.7. Principal component analysis (PCA) of three Kuwaiti subpopulations using 23 autosomal STR markers.	112
Figure 4.8. Mean likelihood and Delta K values generated from <i>STRUCTURE</i>	113
Figure 4.9. Analysis of the structure of the Kuwaiti population using 23 STR markers.	114
Figure 4.10. Principal component analysis of nine populations genotyped for 23 STR markers. ...	115
Figure 4.11. Principal component analysis of 13 populations using 13 autosomal STR markers. ..	116
Figure 4.12. Neighbour joining (NJ) tree based on autosomal STR markers.	117
Figure 5.1. Average DoC for the ForenSeq™ STR loci.	125
Figure 5.2. ForenSeq™ positive control profile.	127
Figure 5.3 Increase in allele number at autosomal STRs as a result of sequence variants.	128
Figure 5.4. A plot of DoC and ACR in heterozygote allele pairs at the SE33 locus.	131
Figure 5.5. IGV display of SE33 read alignment of three samples sequenced using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® system.	133
Figure 6.1. Depth of Coverage for ForenSeq™ DNA Signature Prep Kit iiSNPs.	145
Figure 6.2. ForenSeq™ DNA Signature Prep Kit positive control profile for 94 loci SNP.	147
Figure 6.3. Number of observed haplotypes at each SNP locus and the increased discriminatory power associated with sequence variants.	149
Figure 6.4. Effective number of alleles (Ae) for 45 iiSNP microhaplotypes.	153
Figure 6.5. Analysis of the structures of nine population groups using the 87 iiSNP loci.	156
Figure 6.6. F_{ST} -based genetic distance matrix among global population groups.	157
Figure 6.7. F_{ST} -based genetic distance matrix among Middle East populations.	158
Figure 7.1 Electropherogram of the internal size standard (WEN 500 Y23).	164
Figure 7.2. Electropherogram of the Y23 allelic ladder.	165
Figure 7.3. Electropherogram of the Y23 positive control.	166
Figure 7.4. Average DoC for the ForenSeq™ DNA Signature Prep kit sex-chromosome STR loci.	167
Figure 7.5. Increase in allele number at autosomal STRs as a result of sequence variants.	168
Figure 7.6. Illustration of a sequence variant at HPRTB locus.	170
Figure 7.7. Appearance of an allele peak classified as a DYS392 off-ladder allele and a DYS643 null allele.	173
Figure 7.8. IGV display of DYS643 read alignment of a sample sequenced using the ForenSeq™ DNA Signature Prep Kit.	174

List of Tables

Table 3.1. Genotype frequencies for 30 DIPplex indel loci in the Kuwaiti population.	79
Table 3.2. Results of the Hardy-Weinberg equilibrium (HWE) tests.	80
Table 3.3. Results of the inbreeding coefficient test.	82
Table 3.4. Forensic parameters for the DIPplex markers in 150 Kuwaiti individuals.	83
Table 3.5. Summary statistics for population genetic and forensic parameters in the Kuwaiti population.	84
Table 3.6. The success rate of inferring the population of origin of individuals based on 30 indel markers.	84
Table 3.7. The success rate of inferring the population of origin, based on 30 indel loci, of individuals from three closely related populations.	86
Table 3.8. Comparison of <i>p</i> -values in tests of the Hardy-Weinberg equilibrium (HWE) between the Kuwait population and other populations.	92
Table 4.1. Off-ladder alleles identified using the Verogen ForenSeq™ DNA Signature Prep kit.	107
Table 4.2. Allele frequencies and forensic statistics among 400 Kuwaiti individuals typed at 23 autosomal STR loci in the PowerPlex® Fusion 6C kit.	109
Table 5.1. Novel autosomal STR sequence variants.	130
Table 5.2. SE33 sequence motif patterns observed in 365 Kuwaiti individuals.	132
Table 5.3. Hardy-Weinberg Equilibrium (HWE) test	135
Table 5.4. Forensic statistics in the Kuwaiti population using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® system.	136
Table 5.5. Summary statistics for forensic parameters in the Kuwaiti population.	136
Table 5.6. Comparison of combined match probabilities (CMP) in the Kuwaiti population.	137
Table 6.1. Novel alleles associated with iiSNP in the Kuwaiti population.	150
Table 6.2. Summary statistics for forensic parameters among the Kuwaiti population.	154
Table 7.1. The PowerPlex® Y23 System dye-labels and allelic Ladder Information.	164
Table 7.2. Novel X-STR sequence variants.	169
Table 7.3. Forensic parameters calculated using pooled male and female allele frequencies for the Kuwaiti population.	171
Table 7.4. Forensic parameters calculated using male haplotype frequencies for the Kuwaiti population.	172

List of equations

Equation 2.1. Random match probability (RPM)	66
Equation 2.2. Power of exclusion (PE)	67
Equation 2.3. Combined power of exclusion (CPE)	67
Equation 2.4. Combined paternity index (CPI)	67
Equation 2.5. Polymorphism information content (PIC)	68

List of abbreviations

AIM	Ancestry informative marker
AMOVA	Analysis of Molecular Variance
bp	Base pair
CCD	Charge-coupled device
CE	Capillary electrophoresis
CMP	Combined match probability
CODIS	Combined DNA Index System
CPD	Combined power of discrimination
CPE	Combined power of exclusion
CPI	Combined Paternity index
C _q	Quantification cycle
CRS	Cambridge reference sequence
CNV	Copy number variation
ddNTP	Dideoxynucleoside triphosphate
DEL	Deletion
Dip	Deletion and insertion polymorphism
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EVC	Externally visible characteristic
ExoSAP	Exonuclease Shrimp alkaline phosphate
FBI	Federal Bureau of Investigation
FSS	Forensic Science Service
GRCh38	Genome Reference Consortium Human genome build 38
GWAS	Genome-wide association studies
hg19	Human Genome version 19
HGDP-CEPH	Human Genome Diversity Project - Centre d'Etude du Polymorphism Humain
HV	Hypervariable region
HWE	Hardy-Weinberg Equilibrium
ID-SNP	Identity testing SNP
indel	Insertion-deletion polymorphism
INS	Insertion-deletion polymorphism
ISFG	International Society of Forensic Genetics
KB	kilobase
LD	Linkage Disequilibrium
LE	Linkage equilibrium
LI-SNP	Identity testing SNP
LR	Likelihood ratio
M	Molar
MPS	Massively parallel sequencing

MSY	Male-specific region of the Y chromosome
mtDNA	Mitochondrial DNA
ng	nanogram
NGS	Next generation sequencing
NYR	Nonrecombining region
PAR	Pseudoautosomal region
PCA	Principal component analysis
PCR	Polymerase chain reaction
PD	Discrimination power
PE	Power of exclusion
PI	Paternity index
PIC	Polymorphism information content
qPCR	Quantitative polymerase chain reaction
R ²	Regression coefficient
RM	Rapidly mutating
RMP	Random match probability
ROS	Reactive oxygen species
RPM	Revolutions per minute
SAP	Shrimp alkaline phosphate
SGM	Second generation multiplex
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
SWGDM	Scientific Working Group on DNA Analysis Methods
TPI	Typical paternity index
VNTR	Variable number tandem repeat
YHRD	Y Chromosome Haplotype Reference Database
μL	Microliter

Chapter 1 Introduction

1.1 DNA in forensic science

Since the discovery of DNA as the genetic material encoding the blueprint for the majority of living organisms on earth, scientists have been seeking to untangle its secrets. This has given rise to rapid improvements in methodologies and technologies in forensic genetics. The field was revolutionised by the introduction of DNA testing for human identification (also known as DNA fingerprinting, DNA profiling or DNA typing) in the mid-1980s [1]. Since this development, criminal justice systems across the globe have been able to use this new scientific information that gives them greater power to resolve ambiguities in criminal investigations and differentiate between victims and suspects by linking biological evidence to its donor. The use of DNA in criminal investigations is thought to be the single most significant innovation in forensic medicine in the late 20th century [2]. Today, DNA testing has become an extremely reliable tool serving justice and is considered by many to be the most critical field in forensic science. This, in turn, has allowed forensic genetics to contribute to many biological applications such as paternity testing [3], identification of missing persons and victims of mass disasters [4], analysis of human migration patterns [5] and geographical population structure [6], ancestry studies [7], analysis of difficult forensic samples and ancient biological specimens [8, 9], and prediction of phenotypic traits [10]. The significant advancements made in forensic genetics during the last few decades has made it a highly recognised and well-accepted field.

1.2 The *Homo sapiens* genome

The first completed sequence of the human genome, containing more than 3 billion base pairs, was released by the Human Genome Project in 2003 [11]. Since then, the human genome sequence has been updated and refined by the Genome Reference Consortium. Aiming to provide the most accurate and complete human genome reference assembly, they have released several enhanced versions, of which the most recent and accurate is GRCh38, released in 2014 and updated four times a year since its release [12].

The human genome is made up of nucleic acid bases in the DNA, the bulk of which are packaged into chromosomes within the nucleus and a small proportion within the mitochondria. A nucleated cell in the human body is diploid, with 23 pairs of chromosomes, one chromosome in each pair inherited from each of the two parents. Mitochondrial DNA is inherited entirely through the maternal line. Nuclear chromosomes include 22 pairs of autosomal chromosomes and one pair of sex chromosomes (X and Y chromosomes). The autosomes (chromosomes 1–22) are inherited from both parents, whereas the Y chromosome is inherited through the paternal line, to be carried further by the male offspring. The X chromosome has a unique inheritance pattern in which females can pass either X chromosomes to the next generation while males pass only the one. These differences in inheritance patterns give individuals their unique genetic characteristics and lead to genetic differentiation (variation) through generations. Detailed patterns of inheritance throughout two generations are shown in **Figure 1.1**.

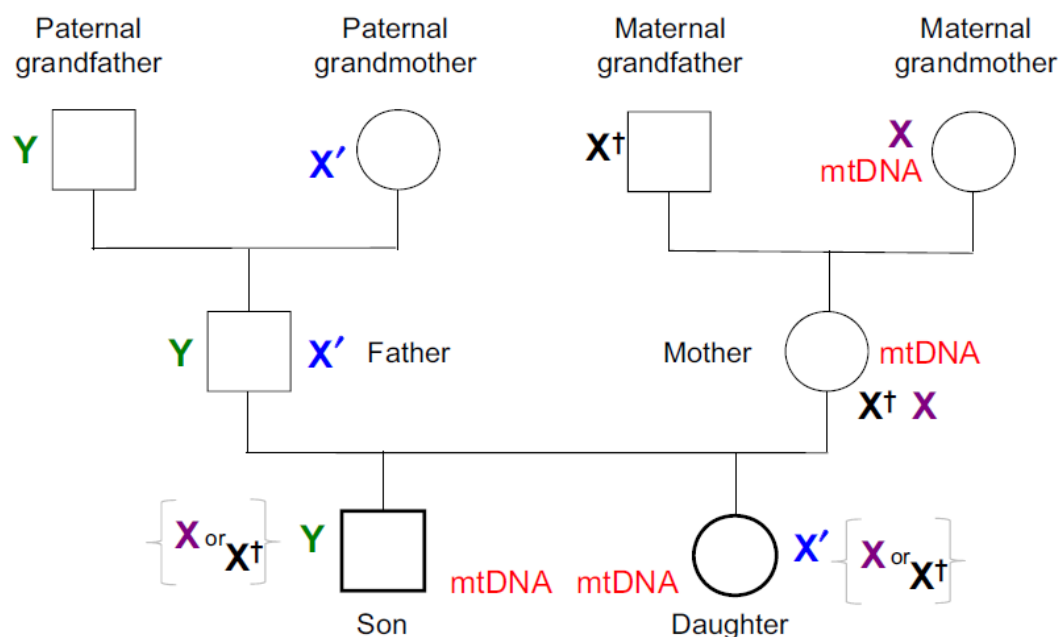


Figure 1.1. Inheritance of mitochondrial DNA (mtDNA) and sex chromosomes.

The Y chromosome is inherited patrilineally, from the grandfather to the grandson through the father, while mtDNA is inherited matrilineally, from the grandmother to the granddaughter and grandson through their mother, while the mtDNA that is passed to the father from the grandmother is not being passed any further. A female has two X chromosomes while a male has only one, and in every generation, a female can only pass one of its either X chromosomes to each of her offspring, while a male passes one X chromosome. The 22 pairs of autosomes are passed on by both parents to offspring so everyone receives 50% of their DNA from each parent and 25% from each grandparent (source [13]).

1.3 Genetic variation

Genetic variation is found within populations as differences in genome sequences between individuals, and the varying impact of forces such as natural selection, random genetic drift, and migration mean that the frequencies of different genetic variants can differ between populations. The main causes of genetic variation are recombination and mutation. Recombination takes place in germ cells, giving rise to a new arrangement of nucleotides from crossing over of the genetic material between maternal and paternal chromosomes, whereas mutation occurs in both germ cells and somatic cells. Variation is a dynamic process in which new variants can be formed, or existing variants can be altered, causing the formation of new alleles across generations, which eventually leads to evolution and diversity.

Several large-scale human sequencing projects have been completed to construct a catalogue of genetic variation (genomic sequence differences) between individuals, which has enriched the knowledge of human diversity. Notably, the 1000 Genomes Project, which is an extension of the HapMap Project [14], has sequenced thousands of individuals to provide a comprehensive description of the genetic variation observed among worldwide populations [15, 16]. In the project's last phase, genetic variation between more than 2,500 individuals was estimated as 88 million variant sites, with low nucleotide diversity between populations and high diversity within populations. A typical human genome has 4-5 million genomic sites that differ from the reference genome, accounting for ~0.07% of the three billion base pairs in the human genome, of which the majority exist in the form of single nucleotide polymorphisms (SNPs) and short insertion-deletion polymorphisms (indels) [17]. The availability of a large number of genetic variants allows for a deeper comprehension of gene-based studies in different disciplines to understand diseases, to investigate human history, and to differentiate between individuals, which is the most critical aspect in forensic genetics. DNA sequence variants, known as DNA polymorphisms, can serve as genetic markers and be targeted by various sequence-capturing approaches.

DNA polymorphisms can be classified into different categories based on how DNA sequences vary between different alleles. DNA polymorphisms can be single nucleotide polymorphisms (SNPs), variable number tandem repeats (VNTRs; also referred to as minisatellites) including short tandem repeats (STRs; also referred to as microsatellites), insertion-deletion polymorphisms (indels, also termed dips), and structural variations (involving genomic

regions > ~1KB [18]) including copy number variation (CNVs; a DNA segment that is larger than 1KB which includes insertions, deletions and duplications), segmental duplications, inversions and translocations. The variants commonly used as genetic markers for DNA profiling, which will be a focus of this thesis, are STRs, SNPs and indels.

1.4 A brief history of DNA fingerprinting

The concept of DNA fingerprinting emerged in 1985, when Professor Sir Alec Jeffreys used VNTRs to generate individual-specific DNA patterns (DNA fingerprints) that can be applied to both human identification and paternity testing [19]. The VNTR technique was first described as an effective tool when it was used to help in an immigration case [20], and to solve a murder case [21]. However, the VNTR technique requires large samples with high molecular weight. Therefore, it is not suitable for samples with low levels of genetic material or with degraded DNA, which are common in forensic cases. To overcome these limitations, polymerase chain reaction (PCR) has been incorporated into DNA profiling tests. The ability of PCR to amplify billions of copies of DNA in a matter of a few hours has revolutionised forensic biology [22].

Since the early 1990s, and with the invention of PCR, DNA testing has taken an enormous step forward in both its discriminatory power and ability to retrieve information from low quantities of DNA, especially after STR profiling was introduced. STRs have supplanted VNTR methods and have been used successfully in forensic laboratories for decades [23, 24]. Even with the recent development of next-generation sequencing, which has the ability to sequence STR loci, forensic geneticists still rely on traditional capillary electrophoresis (CE) as the primary technique to genotype STR markers. Many STR kits have therefore been validated and are being used widely in forensic laboratories around the world. An overview of the main developments in DNA testing events is illustrated in **Figure 1.2**.

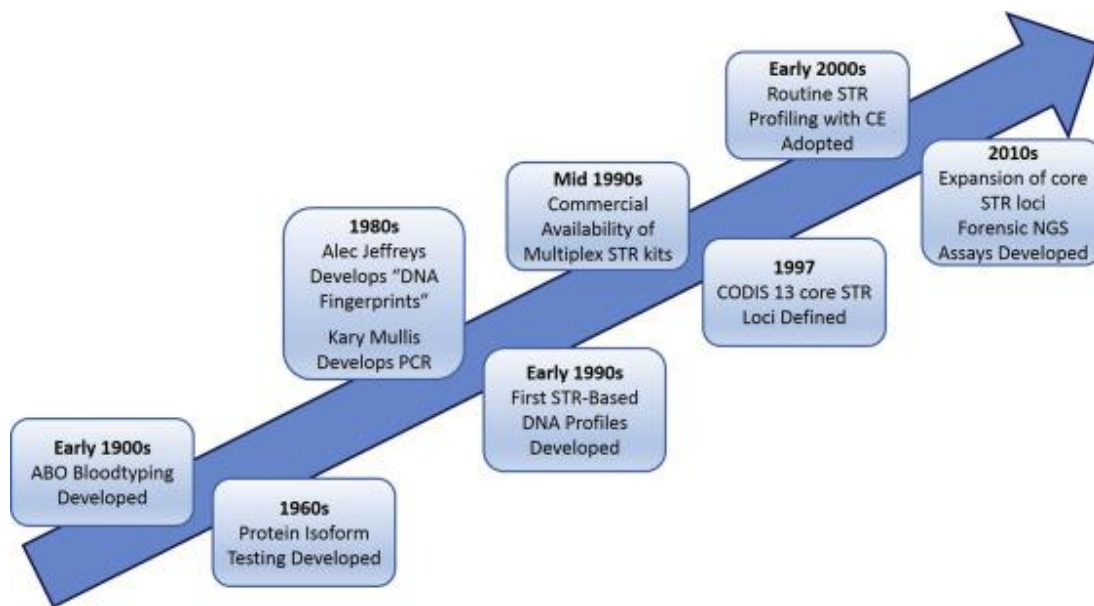


Figure 1.2. An outline of the main history of DNA fingerprinting development in the 20th–21st century.

Before DNA fingerprinting was established, forensic testing was focused in the late 1900s on protein biomarkers. Since Alec Jeffreys developed DNA profiling in the 1980s, forensic testing has significantly advanced in both the forensic genetic marker system and sample throughput (source from [25]).

1.5 Short tandem repeat (STR) markers

STRs are repetitive stretches of nucleotide units accounting for ~3% of the human genome [26], of which 92% are in non-coding regions [27]. The human genome has an estimated 700,000 STRs, with 97% having a core repeat motif 2 to 6 base pairs (bp) in length [28]. Due to high mutation rates at these sites, their variability is the highest among the different types of polymorphisms in the genome. These mutations cause STR polymorphisms between individuals by changing the number of times the core 2–6 bp unit at an STR locus is repeated, resulting in STR loci possessing alleles of different lengths [29] (**Figure 1.3**). Multiple STRs are used for DNA testing given that these loci and their alleles are inherited independently. Therefore, the frequencies of the alleles detected can be multiplied together to calculate a statistic known as the random match probability (RMP), i.e. the probability that two unrelated individuals randomly selected from the study population would possess identical genotypes. The diversity and high discriminatory power of STRs allow them to be used not only for human identification purposes, but also for studying population genetics and in paternity cases [3].

STRs can be classified into three major groups based on their repeat region pattern: ‘simple STRs’ with identical core repeats in terms of the repeat length and sequence of nucleotides, e.g. the TH01 locus containing [AATG]_n; ‘compound STRs’ with two or more adjacent simple repeats, e.g. the vWA locus with [TCTA] [TCTG]_n [TCTA]_n; and ‘complex STRs’ with multiple repeat units with non-repeated intervening sequences, such as the FGA locus, which comprises [TTTC]₃ TTTTCT [CTTT]_n CTCC [TTCC]₂ [30] (**Figure 1.3**).

(a) Simple STR	TH01	
	6	[AATG] ₆
	7	[AATG] ₇
	8	[AATG] ₈
	9	[AATG] ₉
	9.3	[AATG] ₆ ATG [AATG] ₃
(b) Compound STR	vWA	
	16	TCTA [TCTG] ₄ [TCTA] ₁₁
	17	[TCTG] ₄ [TCTA] ₁₃
	18	TCTA [TCTG] ₄ [TCTA] ₁₃
	19	TCTA [TCTG] ₆ [TCTA] ₁₂
	19	TCTA [TCTG] ₄ [TCTA] ₁₄
(c) Complex STR	FGA	
	18	[TTTC] ₃ TTTT TTCT [CTTT] ₁₀ CTCC [TTCC] ₂
	19	[TTTC] ₃ TTTT TTCT [CTTT] ₁₁ CTCC [TTCC] ₂
	19.2	[TTTC] ₃ TTTT TT [CTTT] ₁₂ CTCC [TTCC] ₂
	20	[TTTC] ₃ TTTT TTCT [CTTT] ₁₂ CTCC [TTCC] ₂
	21	[TTTC] ₃ TTTT TTCT [CTTT] ₁₃ CTCC [TTCC] ₂

Figure 1.3. Classification and variability of short tandem repeats (STRs).

(a) Simple STR loci such as TH01. (b) Compound STR loci such as vWA. (c) Complex STR loci such as FGA. The numbers in the middle column refer to the allele designation based on the number of copies of the core repeat present. Note that STR loci can also possess microvariant alleles that contain partial repeat/s such as 9.3 at TH01 (where the sequence of the tenth repeat is ATG instead of AATG) and 19.2 at FGA (TT instead of TTT).

1.5.1 Development of STR multiplexes

STR multiplexes, which combine several STR markers for amplification in one PCR reaction, have undergone several stages of development. In 1994, the first forensic STR multiplex for forensic human identification was developed by the Forensic Science Service (FSS) of the United Kingdom. The quadruplex contained four STR loci: vWA, TH01, FES/FPS and F13A1 [31]. A second generation of the multiplex (SGM) was further developed by the FSS in 1995, consisting of six STRs: TH01, vWA, FGA, D8S1179, D18S51 and D21S11, as well as a sex

chromosome marker (amelogenin) [32]. In 1999, these six markers, in addition to D3S1358, were chosen as a European Standard Set (ESS) of STR loci [33].

In 2000, SGM was replaced by a commercial kit known as SGM Plus, developed by Applied Biosystems. The kit incorporated four more loci (D3S1358, D19S433, D16S539 and D2S1338), and became a standard tool in forensic laboratories in the United Kingdom until 2014, when the DNA-17 system was announced as the standard DNA profiling system in England and Wales [34, 35]. The system comprises the 11 SGM Plus markers (including amelogenin), with an additional six STR markers. In 2015, Scotland adopted a DNA-24 typing system (including 21 STRs, a Y-STR, Y-indel and amelogenin) using the commercial GlobalFiler™ PCR amplification kit produced by Thermo Fisher, which became the largest and most advanced DNA testing assay used in European forensic laboratories [36]. Both systems are highly sensitive and can generate profiles from samples containing small amounts of DNA. The enhanced assays can also bring greater discriminatory power to the evidence, i.e. a high probability that two DNA profiles would differ (would not match).

The US began their STR project in late 1997 when 13 CODIS (Combined DNA Index System) loci were selected for use in the FBI's national DNA database [37]. Subsequently, two commercial kits including the 13 CODIS loci were introduced by Promega Corporation (PowerPlex® 16) [38] and Applied Biosystems™ (AmpFLSTR® Identifiler®) [39]. In 2017, the 13 CODIS STR loci were expanded to 20 loci (**Figure 1.4**) [40]. The incorporation of these extra loci has facilitated the sharing of DNA profiles globally because of greater overlap in the loci surveyed in different countries.

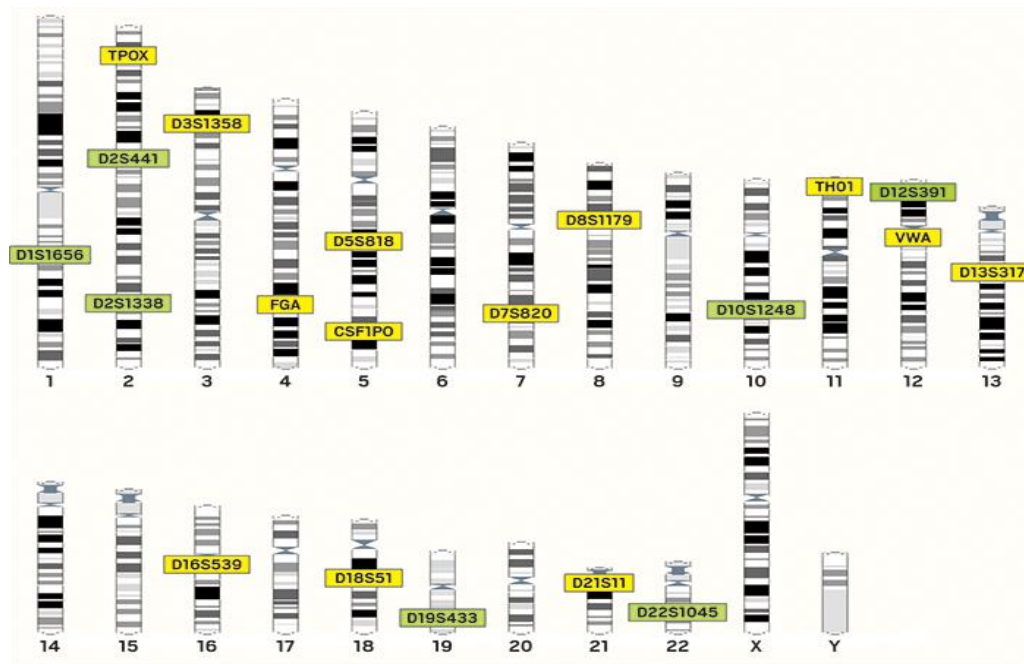


Figure 1.4. Expanded CODIS core short tandem repeat (STR) loci.

The STR loci used in the U.S. Combined DNA Index System (CODIS) database are distributed across all chromosomes. The original 13 are highlighted in yellow, and the seven additional loci are highlighted in green (source [41]).

The major advantage of expanding the number of core STR loci is that it allows greater discriminatory power. In addition, international databases are growing, and the use of these extra loci may help to decrease the possibility of adventitious matches, especially when comparing profiles at the international level [25]. For this reason, and with the development of six-dye chemistry, which allows multiplexing of more loci in a single PCR reaction, several commercial kits have been launched incorporating the seven additional CODIS loci with the original 13 loci, including the Investigator 24plex QS Kit (abbreviated as Investigator 24plex) by QIAGEN (Qiagen, Hilden, Germany), the PowerPlex® Fusion 6C System (abbreviated as Fusion 6C) from Promega (Promega, Madison, WI, USA), and GlobalFiler™ PCR Amplification Kit (abbreviated as GlobalFiler) from Applied Biosystems™ (Life Technologies™, Carlsbad, CA, USA). Each kit has features that allow them to perform differently; therefore, it is recommended that the most suitable kit is implemented in laboratories, based on their demands and operational needs (**Figure 1.5**) [40].

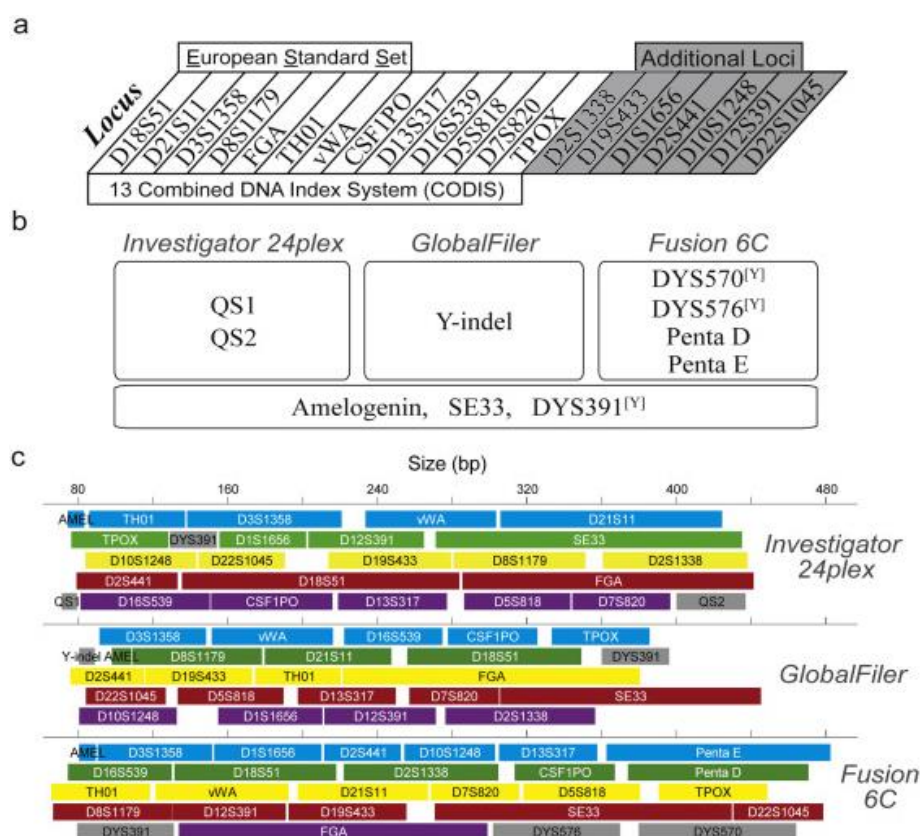


Figure 1.5. Recently developed commercial short tandem repeat (STR) kits containing the expanded CODIS STR loci.

(A) shows the expanded 20 CODIS core loci which consist of the 13 CODIS STRs and the seven additional STRs which are shaded in grey. (B) shows the extra featured markers in each of the three kits, and the loci that are shared between them. The Investigator 24plex kit contains two quality sensors (QS1 and QS2) for performance and quality checks. GlobalFiler has a Y-chromosome insertion and deletion (Y-indel) marker, which has a small size of less than 90 bp, for identification of biological males, especially useful for identifying male minor contribution in sexual assault cases which involve degraded or low-quality samples. Fusion 6C contains two rapidly mutating Y-STR loci [Y] as well as Penta D and Penta E to enhance discriminatory power. (C) The arrangement of loci in each polymerase chain reaction (PCR) kit, showing the colour channel and size distribution for the loci (modified from [42]).

1.5.2 Mini-STRs

Although STR profiling systems are widely used in almost every forensic laboratory, and are considered the gold standard by the community, forensic geneticists often report partial or incomplete genetic profiles with allelic or locus dropouts. To overcome this limitation, especially when typing challenging samples using existing STRs, an approach that has been demonstrated as useful is to reduce STR amplicon size to generate the shortest amplicon possible, which can be achieved by moving the primer binding sites closer to the STR repeat region [43-45]. For example, Butler et al. redesigned the primers of the 13 CODIS STRs by

reducing their size and combined these with other commercial loci commonly used in routine STR analysis (D2S133, Penta D and Penta E) to produce the first multiplex of mini-STRs [46].

Shorter STR amplicons created by primer redesign were successfully used to generate DNA profile information from human remains recovered from the World Trade Centre [47]. The method proved to be successful in producing full profiles from samples, many of which contained degraded or low quantities of DNA, that were partially typed or gave negative results with conventional STR typing [48-52]. This approach was further developed in 2007, when Applied Biosystems released the AmpFISTR® MiniFiler™ kit, the first commercial mini-STR amplification kit. MiniFiler amplifies eight loci, adapted from the existing AmpFISTR Identifier kit, but with the primers redesigned to produce the shortest amplicons possible (Figure 1.6).

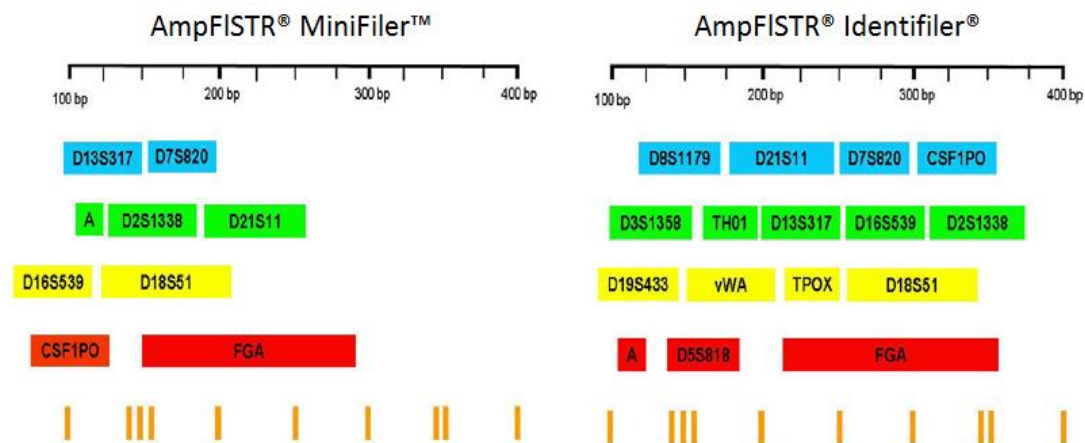


Figure 1.6. AmpFISTR® MiniFiler™ and AmpFISTR® Identifier™ loci.

Loci present in the Identifier kit were re-engineered in the MiniFiler kit to provide a set of amplicons less than 300 bp in length (modified from [33]).

However, the number of stutter peaks (explained in section 1.5.4) have been found to be higher in the MiniFiler kit compared to that in Identifier. In addition, the eight STRs contained in MiniFiler are not as informative as the 15 markers in Identifier [53]. Moreover, one of the biggest disadvantages of this type of system is the small number of loci that can be incorporated into one multiplex design. This is seen in the miniplex created by Butler et al. [46]. In miniplexes, only one locus is labelled with each dye to ensure the shortest amplicon possible and to avoid interference with other markers. In contrast, Applied Biosystems has adopted the mechanism of mobility modifiers, i.e. non-nucleotide linkers that permit spacing

between two loci and avoidance of allele overlap, thereby allowing MiniFiler to amplify two loci per dye [26, 53].

1.5.3 Basic steps of the polymerase chain reaction-capillary electrophoresis STR genotyping process

1.5.3.1 DNA Extraction

The extraction procedure is used to purify the DNA molecule by eliminating other cellular material that might interfere with downstream analysis. Such a procedure is needed to ensure no inhibitory material is present in the sample. The ultimate goal of the extraction procedure is to produce an adequate quantity of high-quality DNA from forensic casework samples. A number of DNA extraction protocols are available; however, most of them involve the same basic steps, including digestion of tissue and cellular proteins by using enzymes and detergents and then separating DNA and eliminating cellular debris [54]. Additional factors for consideration when choosing the most suitable extraction method include cost-effectiveness and the simplicity of throughput and workflow, which will subsequently save crucial laboratory processing time. The most common DNA extraction protocols are organic extraction using phenol–chloroform or chelating resins such as Chelex®-100 resin, solid-phase extraction such using silica or paramagnetic resin, and DNA extraction from FTA blood collection cards (**Figure 1.7**) [33, 54].

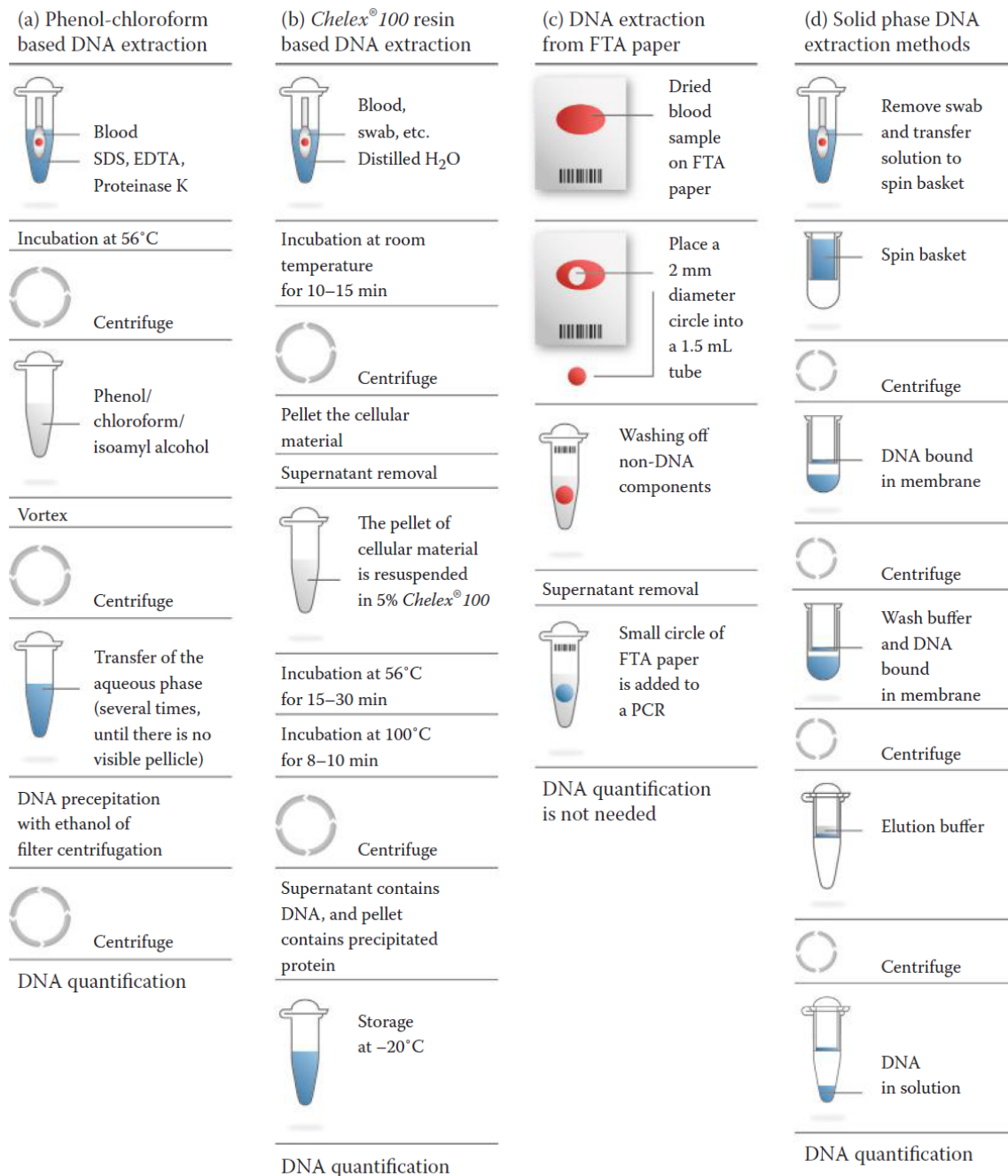


Figure 1.7. Four commonly used extraction protocols.

(a) Organic extraction using phenol–chloroform. **(b)** DNA extraction using Chelex® 100 resins. **(c)** DNA extraction from FTA paper. **(d)** Solid-phase extraction using, e.g. silica-coated magnetic particles (source [54]).

1.5.3.2 DNA quantification

After extraction, the concentration of the isolated DNA is measured using quantification techniques, to obtain the optimum DNA quantity for successful amplification. The most commonly used technology for DNA quantification is quantitative polymerase chain reaction (qPCR), which generally uses a unique TaqMan® oligonucleotide probe that targets the region

of the genome targeted by the qPCR assay [25]. The probe contains a fluorescent reporter dye and a quencher dye on the 5' and 3' end respectively, and whilst there is no amplification of the target, the probe remains intact preventing fluorescence emission from the reporter dye. When the probe anneals to its target and PCR is initiated, *Taq* DNA polymerase extends and cleaves the reporter dye, allowing emission of fluorescence signals (**Figure 1.8**) [33]. The quantity of DNA in a sample is estimated by measuring the intensity of these signals during each PCR cycle, which are then compared to a standard curve generated using known DNA concentrations.

Commercial qPCR kits can accurately determine the quantity of male and female DNA in a sample, as well as the quality of the DNA, i.e. whether the sample is degraded or contains inhibitors. Such a quality-control feature can act as a guide to choosing a suitable class of marker for DNA profiling, or for determining whether an extra purification step is required [25].

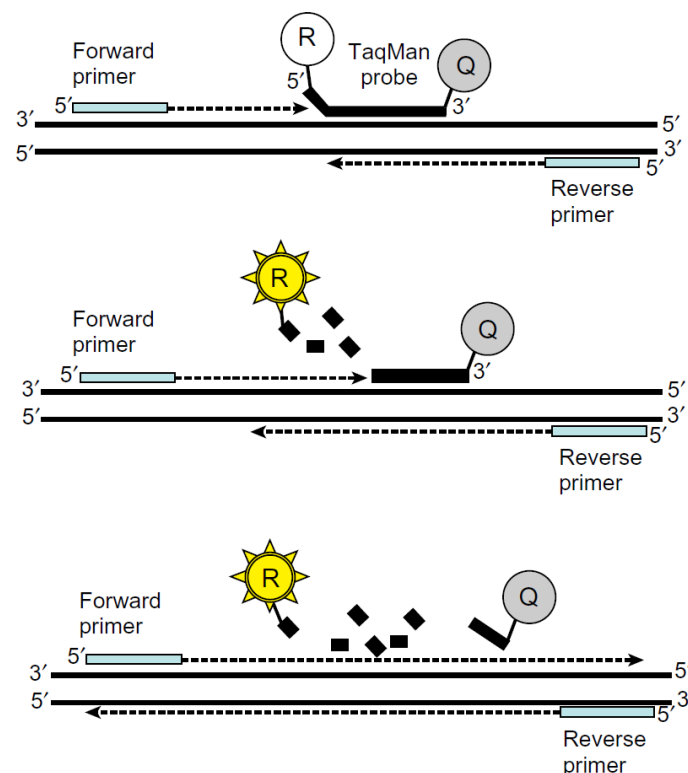


Figure 1.8. Quantitative polymerase chain reaction (qPCR) using a TaqMan assay.

Top: the probe containing a reporter dye (R) and quencher dye (Q) attaches to the target region located between the forward and reverse PCR primers. **Middle:** during PCR, the 5' exonuclease activity of *Taq* DNA polymerase releases the reporter dye which subsequently emits fluorescence. **Bottom:** the cycle is completed when the *Taq* DNA continues elongation of the target region (modified from [33]).

1.5.3.3 DNA amplification

Once quantified, the DNA is amplified using PCR, an enzymatic process in which a specific genomic region is replicated repeatedly to yield many copies (**Figure 1.9**). Fluorescently labelled primers bind to the targeted region of interest and *Taq* DNA polymerase enzyme synthesises the new DNA strand. PCR is performed over several cycles (usually 28 to 35 cycles) of a heating and cooling thermal pattern [55]. In each cycle, a copy of the target STR region is generated, and the PCR products accumulate exponentially with each round of amplification.

Commercial PCR kits usually require 0.5-1 ng of DNA to produce a high-quality and interpretable DNA profile [25]. When too much DNA is used, off-scale data will result, and alleles may be shown as split peaks. In contrast, insufficient template DNA leads to decreased allele peak height and an increase in the likelihood of dropout alleles resulting in a partial DNA profile [55]. Some commercial kits have been engineered to allow amplification directly from DNA-collection cards such as FTA cards (blood collection cards impregnated with chemicals that lyse cells, denature proteins and protect nucleic acids) without the need for extraction or quantification. This approach is suitable for reference and database samples, such as blood and saliva/buccal samples, in which the quality and quantity of DNA are generally good.

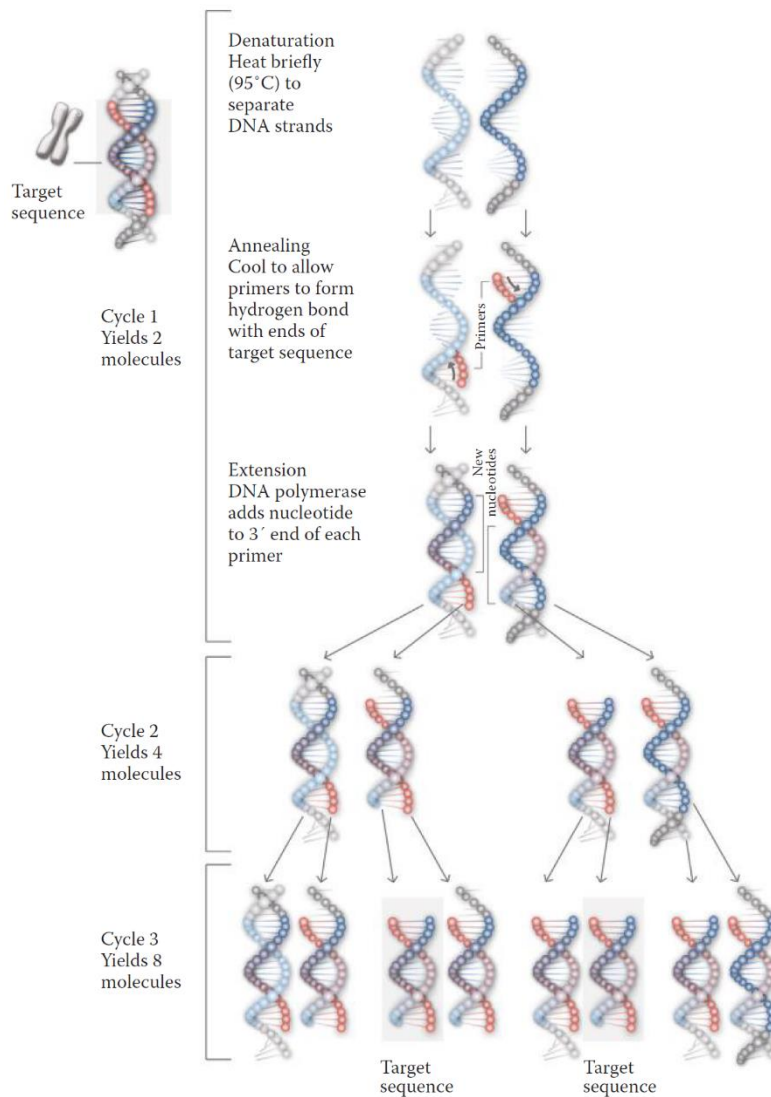


Figure 1.9. DNA amplification with polymerase chain reaction (PCR).

Each cycle involves three steps, denaturing, annealing and elongation, to double the number of DNA molecules (source [54]).

1.5.3.4 Capillary electrophoresis and data interpretation

The amplified PCR products are separated and detected according to their size by CE, using a fluorescence detection method (**Figure 1.10**). Compared to the conventional slab gel method, CE is an automated process allowing multiple samples to be analysed in one run, only requiring a small quantity of DNA material. These advantages have made CE the preferred technique for forensic testing. Several series of CE instruments have been used over the last few decades, namely the Genetic Analyzer 310, 3100, 3130 and more recently the 3500, all of which are manufactured by Applied Biosystems [56]. The CE method uses capillaries which are narrow glass tubes pumped full of polymer. When an electric current is applied, the negatively charged DNA molecules will move through the capillaries towards the positive

electrode, at a speed that is dependent on their size. Amplicons are detected with a charge-coupled device (CCD) camera which, collects the fluorescence emitted as a result of excitation of the dye labels with a laser [33].

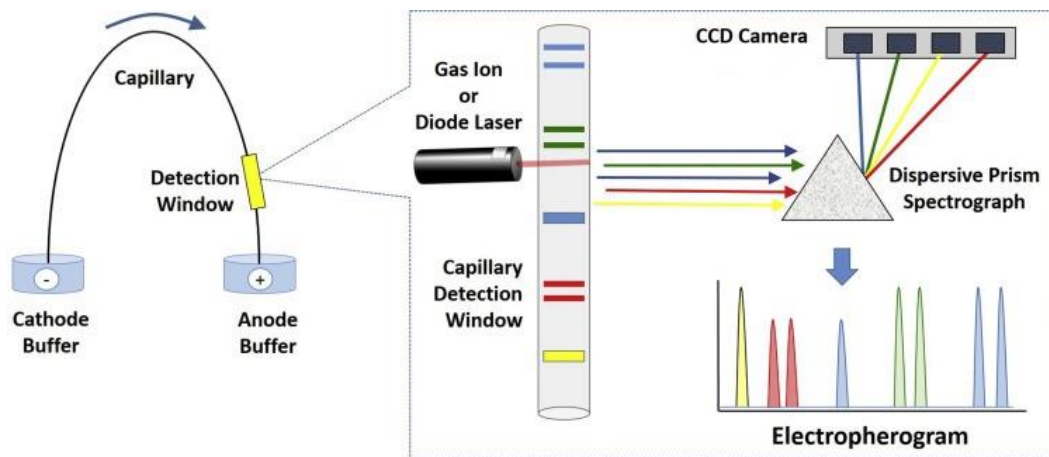


Figure 1.10. The process of separation and detection of short tandem repeat (STR) fragments by capillary electrophoresis.

The fluorescently labelled amplicons travel through narrow glass tubes filled with polymer (capillary arrays) when a high voltage is applied. When amplicons pass the detection window, a laser beam is induced to excite the attached dye which releases a fluorescent signal. Signal colours are separated in a prism spectrograph and signal intensity is quantified by a charge-coupled device (CCD) camera. Dye analysis software is used to identify each locus and the alleles within them by a combination of colour and fragment length, which are shown in an electropherogram (modified from [25]).

Dye analysis and data interpretation are mainly performed using two software packages: GeneScan® or GeneMapper™ ID software. The latter is often used in forensic laboratories. The programme calculates the length of the fragments using internal size standard fragments of known size. The result is a DNA profile represented by an electropherogram, with a range of peaks that represent the alleles in each locus contained in the specific PCR assay. Each peak is identified with a number that corresponds to the number of repeat units. The peaks are identified by comparing them to an allelic ladder that contains all the possible alleles for each locus. Peak height is measured as relative fluorescent units (rfu), which is equivalent to the detected amount of PCR product [33]. An example of a part of an electropherogram (DNA profile) generated by CE is illustrated in **Figure 1.11**.

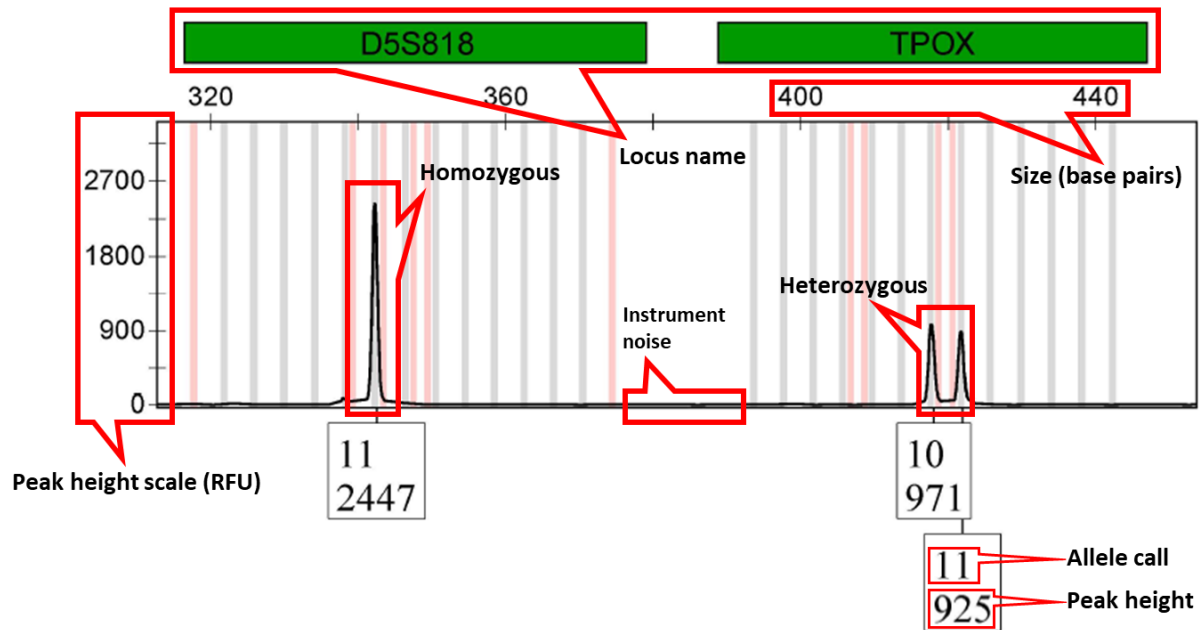


Figure 1.11. A segment of a DNA profile represented by an electropherogram generated by capillary electrophoresis.

Two STR loci are shown: D5S818 with a homozygous allele 11 (one peak) and TPOX which is heterozygous possessing two alleles, 10 and 11 (two peaks). Peak height is measured as relative fluorescent units (rfu). The signal is instantly detectable from instrument noise when it reaches twice the intensity of the highest instrument noise.

The profiles generated can be easily interpreted in cases involving good quality crime scene samples with intact DNA, or reference samples such as blood or saliva. However, interpretation of difficult forensic samples, for example those with a low amount of DNA, degraded DNA or mixed samples of more than one contributor, can be complicated and need a high level of expertise. For these types of samples, the presence of certain artefacts associated with the amplification or CE steps often complicate the interpretation of the DNA profile.

1.5.4 STR artefacts and limitations

Although STR markers are robust, they can display artefacts that may interfere with the interpretation of data and particularly the analysis of mixed samples. Stutter peaks are artefacts that occur during PCR amplification through a mechanism known as slipped-strand mispairing [57]. A stutter is produced during amplification by the slippage of a repeat unit when the template strand is copied; this results in a repeat unit either being missed or copied twice in the new strand produced. Therefore, a shorter or longer PCR product than expected

will be expressed as an additional peak seen at a position one repeat unit shorter or longer than the true STR peaks (**Figure 1.12**) [58]. STRs that contain short di-/trinucleotide core repeats are more vulnerable to stutter production than the longer tetra-/pentanucleotide repeats [58]. Although stutter peaks can usually be eliminated by the analyst in a single-source DNA donor profile, distinguishing between stutter peaks and minor alleles is one of the most difficult tasks in data interpretation, especially in cases where two or more individuals contributed to the DNA profile (mixed profile), or in crime scene samples, which contain low amounts of DNA.

Another artefact seen in some DNA profiles is null (silent) alleles which occur due to allele dropout. Null alleles are observed when sequence variation occurs in the STR repeat region, the flanking region or the primer binding region [13]. These variants prevent the primer from hybridising to its complementary DNA template, leading to failure of amplification of that specific region [13]. A null allele could also occur due to low amounts of DNA input, preferential amplification of one allele in a heterozygote [59] or the presence of PCR inhibitors such as textile dyes (cloth) or haematin (blood), which interfere with amplification and lead to the loss of alleles from a profile [26].

In a high-quality DNA sample, the two alleles in a heterozygous locus should be balanced in terms of peak height; the commonly accepted balance is usually more than 60%, indicating the peak height of the smaller allele relative to that of the larger allele [55]. Genotyping of samples with degraded or a limited amount of DNA can result in peak imbalance, dropping to less than 60% (**Figure 1.12**). In rare circumstances, mutations in one of the primer binding sites can decrease the PCR efficiency of one allele relative to the other, which may cause a severe imbalance or a dropout [55]. A second cause of allele imbalance arises when each amplicon has a different GC content, which causes some alleles to be amplified more easily than others [59]. A third cause is preferential amplification, which is particularly common in large loci possessing many alleles, and arises because the smaller (in terms of repeats) allele will amplify more efficiently than the larger allele, e.g. in the PentaE locus [60].

Full profiles are usually generated from intact, non-degraded DNA extracted from various biological samples such as blood, semen, or saliva. However, common forensic samples and mass disaster remains are exposed to harsh environmental conditions, which cause degradation of the DNA molecule. STR profiling shows some limitations when dealing with degraded samples with a low quantity of DNA or damaged DNA. When the DNA fragments in degraded DNA fall below the length of the longest STR loci in a profiling, some alleles will not amplify. Failure to amplify large fragments will generate a partial DNA profile in which alleles at smaller STR marker loci exhibit higher peaks than those at larger loci (which may not be amplified; **Figure 1.12**).

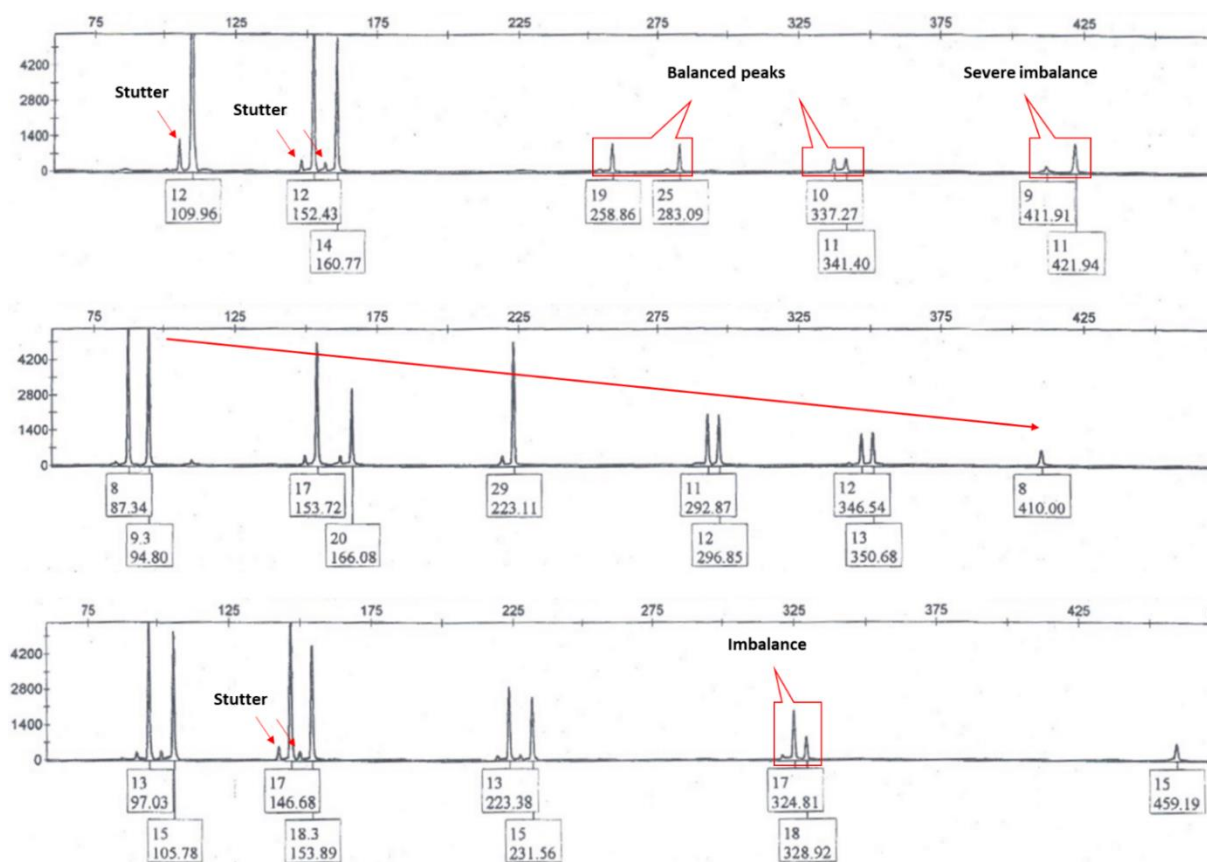


Figure 1.12. A segment of a DNA profile generated from typing old bone sample.

In degraded samples, DNA is often fragmented and therefore falls below the length of larger STR loci. This is shown by a declining 'ski slope' pattern in the profile (as indicated in the middle panel above), where shorter STRs generate higher peaks, while larger STRs generate shorter height peaks. In addition, an artefact such as (severe) imbalance between two alleles in a heterozygous locus can also be observed in such a poor-quality DNA sample. Arrows indicate stutters produced by polymerase chain reaction (PCR), which are usually less than 15% of the associated allele height.

1.6 Single nucleotide polymorphism (SNP) markers

Single nucleotide polymorphisms (SNPs) are single base substitutions which are the most abundant polymorphism in the human genome, being found within genes in exons and introns or in other non-coding regions such as intergenic regions [61]. The abundance of SNPs across the genome enables a deep understanding of the genetics of diseases and makes them promising material for use in human identification. Most SNPs exist in a diallelic (or biallelic) state represented by two alleles, which occur in six substitution alternatives ($A \leftrightarrow G$, $C \leftrightarrow T$, $A \leftrightarrow T$, $C \leftrightarrow G$, $T \leftrightarrow G$ and $A \leftrightarrow C$) [62] (**Figure 1.13**). In the genotype of a diploid organism, an SNP at a single locus can have only one of three possible genotypes: AA, BB (homozygous) and AB (heterozygous) where A and B could be A, G, C or T.

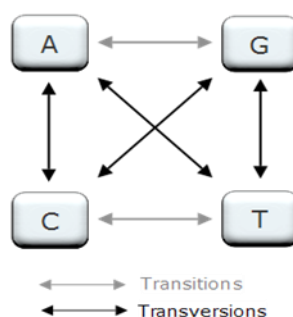


Figure 1.13. Potential nucleotide substitutions that can cause single nucleotide polymorphisms (SNPs).

Diallelic SNPs with six possible mutational changes are represented by arrows. Dark arrows show transversions which are pyrimidine $\leftrightarrow$ purine nucleotide changes resulting in four types of SNP: ($G \leftrightarrow T$), ($A \leftrightarrow C$), ($G \leftrightarrow C$) and ($A \leftrightarrow T$). Transitions shown by grey arrows represent changes between purines ($A \leftrightarrow G$) and between pyrimidines ($C \leftrightarrow T$).

One of the many advantages of SNP markers is their low mutation rate, which has been found to be lower than that of STR polymorphisms [63, 64]. Therefore, SNPs have stable inheritance patterns across generations, making them suitable markers for studying ancestral relationships and in phenotypic prediction studies [65]. Given that SNPs are single-base changes, they can be detected as part of small amplicons with sizes of less than 80 bp [66]. This overcomes the major limitation of STRs when typing degraded DNA samples [67] and avoids interpretation complexity by generating profiles without stutter peaks. These features allow SNPs to be used as forensic markers and victim identification, particularly in missing persons and mass disaster cases, such as the September 11 incident [68, 69].

SNP genotyping has a major limitation in that the discriminatory power exhibited by SNPs is lower than that of multiallelic STRs. An estimated 50 SNPs may be as informative as only 13 STRs, i.e. a single STR can match the identification power of 4–5 SNPs [70, 71]. In addition, deciphering DNA mixtures in crime scene samples is extremely challenging when using SNP markers compared to STRs. It is difficult to decide whether any heterozygosity observed in a sample profile represents a mixture (from two homozygotes or homo/heterozygous donors) or a true single heterozygous contributor, although it is not impossible to interpret two samples mixed in different ratios by calculating differences in allelic peak heights [72, 73].

1.6.1 Types of forensic SNP markers

The International Society for Forensic Genetics (ISFG) categorises SNPs into four main groups according to their potential uses in forensic genetics: (i) individual identity SNPs; (ii) ancestry informative SNPs; (iii) phenotype informative SNPs; and (iv) lineage informative SNPs [65, 74].

1.6.1.1 Individual identification SNPs

Identity testing SNPs (ID-SNPs) provide genetic information about individuals and are characterised by low probabilities of two individuals sharing the same alleles across multiple loci, making them useful for differentiating between people [33]. Kidd et al. reported the first panel of ID-SNPs containing 19 markers [75], which they selected from more than 90,000 SNPs. The reduction of markers was based on finding the best SNPs that met their selection criteria of high heterozygosity and small variation in allele frequency among populations, starting from testing the markers across three populations, and ending with highly polymorphic markers across 40 populations [75].

A large panel containing 52 ID-SNPs (amplified in two multiplexes) was reported by the SNPforID Consortium, a European project aimed at developing forensic SNP assays [67], and later developed by Applied Biosystems as the GenPlex commercial kit [76]. The average match probability of the 52-plex gives similar values to those of the 15-locus commercial STR kits (e.g. Identifiler and PowerPlex 16) [67, 77]. Unfortunately, more SNPs than described here are needed to generate a multiplex that could reach or exceed the discriminatory power of recently introduced STR typing systems, given that some commercial STR kits now contain more than 20 markers (e.g. Powerplex Fusion and GlobalFiler).

1.6.1.2 Ancestry informative SNPs

Ancestry informative markers (AIMs) can provide information about the ancestral background of unknown samples or individuals. Autosomal SNPs (i.e. inherited down both parental lineages) that exhibit large allele frequency differences across a wide range of populations are considered the best candidates to investigate ancestry [78]. AIMs are selected to have low heterozygosity and high genetic differentiation (high F_{ST}) among populations. Kidd's laboratory has introduced a panel of 55 ancestry informative SNPs that are highly differentiated among 73 populations [79]. Other multiplexes have also been generated, including a 34-plex created by *SNPforID* which was used to predict the origin of suspects from crime scene samples in the Madrid train bombings [78, 80].

Although AIMs have many advantages in forensic science, some limitations also exist. One of the major limitations is the lack of database information for some populations. As a consequence, populations that have not been previously studied cannot be defined by ancestry markers [79]. Another limitation is the inability of ancestry markers to reveal significant association with specific populations when multiple populations share the same area/continent and a smooth gradient of allele frequencies exists between them [81]. Lastly, there are difficulties associated with identifying admixture of different biogeographical ancestries unless the ancestral populations are represented in the reference populations [79, 82].

1.6.1.3 Phenotype informative SNPs

Phenotype informative SNPs are genetic markers that have the potential to provide information about the physical appearance of individuals, and can be used as 'biological eyewitnesses' to predict certain phenotypes of crime suspects/perpetrators with greater reliability than human witnesses [10, 83]. A part of this process is building databases containing genetic polymorphisms that are responsible for changes in human physical appearance [84]. Genome-wide association studies (GWAS) have been used as a tool to determine the markers underlying several phenotypic traits by using common informative SNPs. DNA markers can be selected from GWAS by identifying polymorphisms associated with a specific trait.

Many SNPs that affect externally visible characteristics (EVCs) have been studied, including height [85-88], male hair loss/baldness [89, 90], hair morphology [91-94], and facial

morphology [95-99]. Due to inadequate genetic information, the predictive accuracy of these SNPs do not conform to the high predictive accuracy required for forensic phenotyping; thus, they remain under development [10]. On the other hand, the genetics underlying pigmentation are well understood through animal and human studies. Therefore, prediction of traits associated with colour, such as skin, hair and eye colour, is relatively accurate [65]. Of these, eye pigmentation is the most precise predictive trait [100].

The first forensic DNA phenotyping system, IrisPlex, was reported by Walsh and colleagues [101]. The multiplex was tested on several populations and contained six phenotype SNPs associated with iris pigmentation [101-104]. A new forensic phenotyping assay, HIrisPlex, was developed with further enhancements, as an updated version of IrisPlex [105]. Notably, HIrisPlex comprises 24 forensically validated phenotypic markers, including 18 hair-colour polymorphisms along with the six IrisPlex SNPs in a single multiplex [105]. These markers have been successfully used on several occasions as a predictive tool to analyse ancient skeletal remains [106-109].

1.6.1.4 Lineage informative SNPs

Lineage informative SNPs (LI-SNPs) are tightly linked neighbouring SNPs which are inherited together as blocks known as haplotype blocks. These haplotypes are found in mitochondrial DNA (mtDNA) and on the Y chromosome where they are inherited in a stable pattern through generations [33]. Therefore, LI-SNPs have been used in studies of ancestry [7], migration patterns [5] and geographical population structure [6].

1.6.2 SNP genotyping

There are several approaches to genotyping SNPs. Forensic geneticists look for techniques that are validated, capable of delivering reliable results, use conventional instruments and are cost-effective. The recommended best techniques for SNP genotyping are those capable of multiplexing more SNPs in a single reaction, rather than assays with limited multiplexing capacity [75, 110]. To date, the most commonly used technique is minisequencing by SNaPshot, based on single base extension (SBE) [111]. The SNaPshot kit (Applied Biosystems) is a generic commercial kit which contains the reagents required for PCR (*Taq* DNA polymerase, buffers, fluorescent-labelled dideoxynucleotide triphosphates [ddNTPs]), to which specialists add their own designed primers. An additional advantage of the SNaPshot

assay is that it uses CE instruments, such as the ABI 3130 and 3500 Genetic Analyzers, which are used in almost all forensic laboratories for STR analysis.

1.7 Insertion and deletion (indel) markers

Indels, also termed dips (deletion and insertion polymorphisms), are diallelic small length variations found in the genome in the form of the absence or presence of specific nucleotide sequences. These types of polymorphisms are useful because they combine the same forensic characteristics of SNPs but use simple STR typing techniques (**Figure 1.14**) whose implementation is feasible in forensic laboratories. Other advantages include their abundance throughout the human genome, low mutation rate, ability to be multiplexed in large numbers with short amplicon size, lack of production of stutter artefacts, compatibility with high-throughput sequencing technologies, and allele frequency variation between populations [112-115].

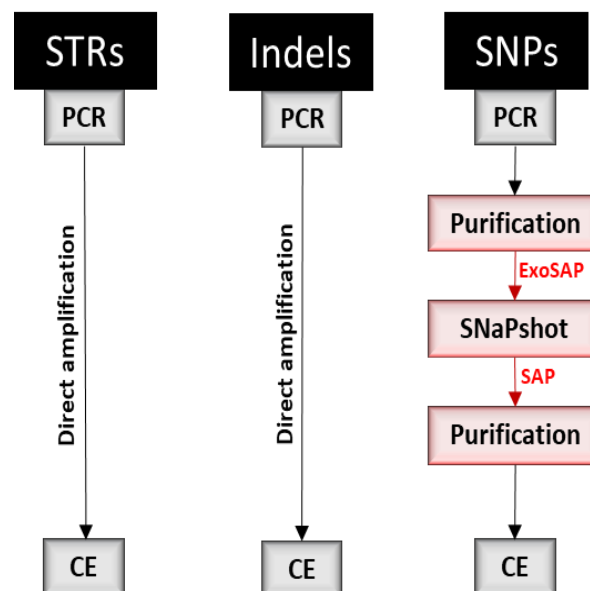


Figure 1.14. Comparison of typing methodologies of short tandem repeats (STRs), indels and single nucleotide polymorphisms (SNPs).

Direct typing of STRs and indels using polymerase chain reaction (PCR) for capillary electrophoresis (CE) compared to SNP typing which requires three additional complex steps.

1.7.1 A brief history of the discovery of indels

The introduction of high-throughput sequencing, i.e. next-generation sequencing (NGS) or massively parallel sequencing (MPS) has made a significant impact on the study of human genetic variants (SNPs and indels). The first studies to identify these variants focused on chromosome 22 [116]. As a consequence, 18% of the overall variants were found to be

insertions and deletions, located in both coding and non-coding regions, which suggests that indels could be useful in studying human diseases and diversity [117]. Furthermore, through a genome-wide study, Weber and his colleagues discovered 2,000 small indels distributed throughout the human genome of which the majority (~96%) had a length of 2–15 bp [114]. They also found that indel polymorphisms comprise ~21% of variants in the human genome and are the second most abundant after SNPs.

A revolutionary large-scale study of indel polymorphisms was reported by Mills et al. in 2006 [112]. Through a novel computational pipeline, they identified more than 415,000 indels, located on both autosomes and sex chromosomes, in samples collected from 36 individuals. They characterised indel variants into four major groups ranging from 1 to 10,000 bp in length (**Figure 1.15**). Another study reported by Kidd et al. identified approximately 796,000 indels ranging from 1–100 bp generated from samples taken from eight unrelated individuals [118]. Despite the number of indels discovered, the proportion found to be in common between the two previous studies was as low as 10% [119].

Moreover, the largest study, which was published in 2015 by the 1000 Human Genomes Project, reported 84.7 million SNPs and 3.6 million indels [120]. These polymorphisms were obtained by whole-genome sequencing of 2,504 individuals from 26 different populations.

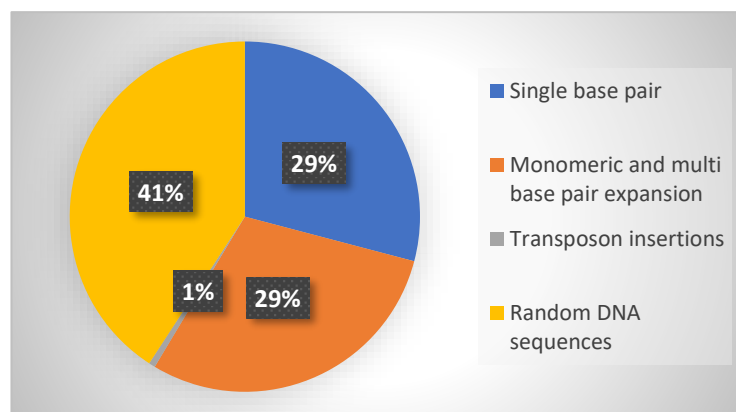


Figure 1.15. Indel polymorphism classes.

Indels were classified into four main categories. The first group consisted of single bp indels of which 84% were insertion and deletion of A and T bases. The second group represented repeat expansions comprising monomeric and multi-base pair expansions (2–15 bp). The third group was transposon insertions. The last group consisted of indels with a wide range of random DNA sequences ranging from 2 bp to 10,000 bp in length, of which 99% were indels with a length below 100 bp.

1.7.2 Forensic indel multiplexes

The first non-commercial multiplex of insertion-deletion polymorphisms for forensic human identification was published in 2009, and contained a set of 38 autosomal diallelic indel markers [115]. These markers were obtained using the Marshfield Clinic's database, which contains indels identified in the study discussed above by the Weber group [114]. Certain criteria were chosen for markers to be included in the multiplex, including (i) a length of 2-5 bp for the indel sequence, (ii) location in noncoding autosomal regions, (iii) heterozygosity of >0.40 , and (iv) a minimum allele frequency of >0.25 in European, African, and Asian populations.

Consequently, a multiplex of indels located in amplicons ranging from 50 to 155 bp was reported and validated using individuals from Africa (Angola and Mozambique), Europe (Portugal) and Asia (Macau and Taiwan). The multiplex was shown to perform well when a complete profile was generated from a bone sample for which a partial STR profile had been produced (**Figure 1.16**) [115]. However, the multiplex lacked a sex marker. Therefore, in late 2015, Braganholi et al. were able to incorporate the amelogenin marker into the multiplex [121]. Several studies have considered the usefulness of this 38-marker multiplex, especially in cases of degraded samples for which partial STR profiles were produced, or in paternity/kinship testing when STRs gave low paternity indices [121-123].

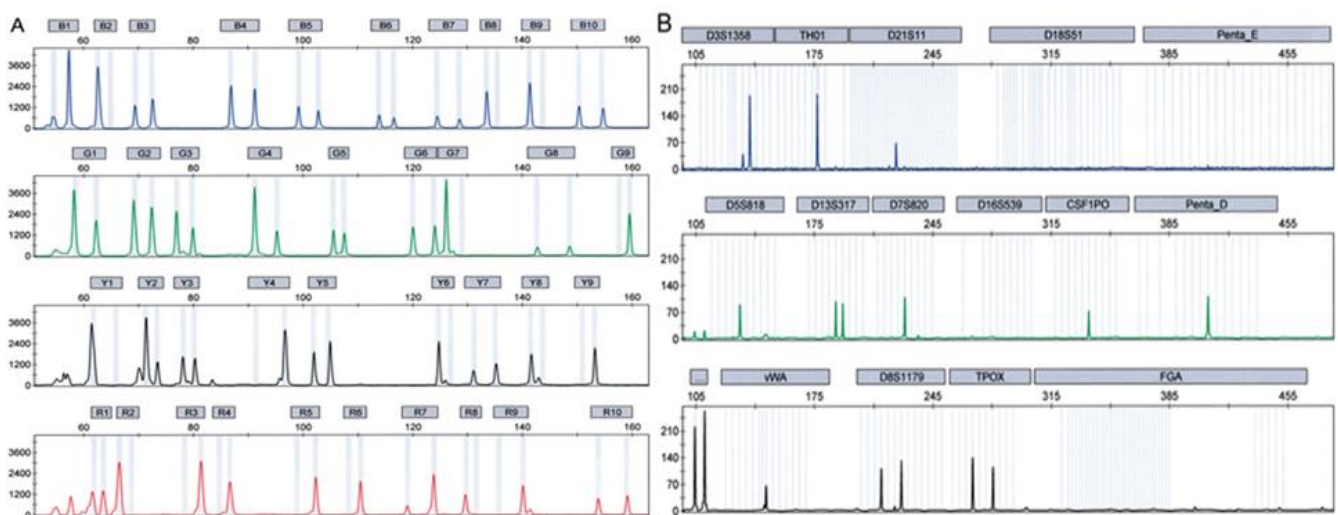


Figure 1.16. Electropherogram of indel typing of a degraded femur bone sample.

(A) shows a full profile generated using the indel marker 38-plex, and (B) represents a partial profile obtained from PowerPlex 16 typing of the same sample (modified from [115]).

1.7.3 Commercial indel markers

The only commercial indel kit available is the Investigator® DIPplex (previously named Mentype [124]) produced by Qiagen. It contains 30 indel markers as well as the amelogenin sex marker. The marker amplicon sizes are in the range of 75 bp to 160 bp. The length of the inserted/deleted sequences differ between alleles and range from 4 bp to 22 bp. The DIPplex kit has been tested on several populations, and studies have shown its usefulness for human identification [125-132]. Some of the DIPplex markers can also be used as AIMs due to the differences in their allele frequencies between populations [133-136]. For example, markers were used to differentiate between native populations and immigrants in Brazil [131]. Moreover, the markers were used to cluster populations and subpopulations at the level of continent and geographical location respectively [137, 138]. The DIPplex markers were also found to be capable of generating full profiles from degraded samples that failed to be typed by conventional STR markers [126, 139]. Similarly, the small amplicon markers proved to be successful in typing ancient skeletal remains (~3000 years old), therefore, they can be potentially used in forensic anthropology [140].

Artefacts such as allele dropout and off-ladder and imbalanced alleles have been reported for some DIPplex markers in some populations; after further investigations, variants within the primer binding sites have been determined to be the cause of these artefacts [124, 137, 141-144]. Similar artefacts have also been observed for other DIPplex markers, but the reasons behind these are undetermined [137, 145]. To overcome some of these artefacts, it has been suggested that DIPplex buffers and primers should be optimised within populations [146].

Both the DIPplex kit and the 38-indel non-commercial multiplex use four fluorescent dyes to label their amplicon products; however, they differ in the number and size of markers. In the DIPplex kit, the alleles of many of the longer amplicons (having larger indel sequences) are located separately from each other and are interspersed with alleles from other loci, whereas the alleles of the shorter amplicons are located adjacent to each other. On the other hand, all the loci in the 38-indel multiplex produce amplicons with alleles that are located next to each other, due to the small length (2–5 bp) of the indel part of the amplicon. The choice of small

indel sizes was an advantage that allowed the incorporation of 38 indels in one multiplex compared to the 30 DIPplex markers which have larger indel sizes (**Figure 1.17**).

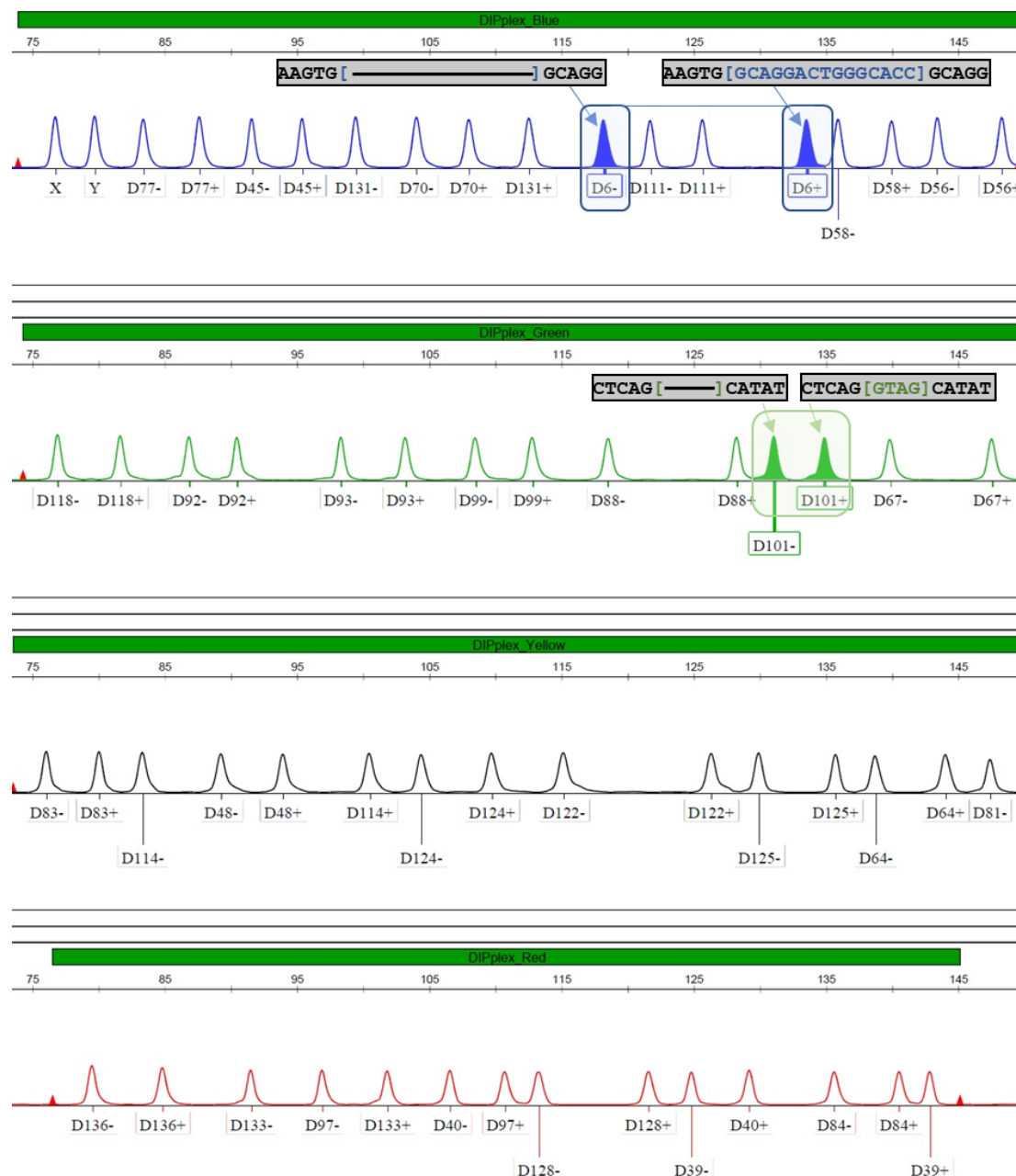


Figure 1.17. Electropherogram of indel typing using the DIPplex kit.

The outlined alleles for the HLD6 marker in the DIPplex blue panel shows alleles of deletion (D6-) and insertion (D6+) of 15 bases interspersed with HLD111 alleles (D111- and D111+). The marker HLD101 in the green panel shows four-base deletion (D101-) and insertion (D101+) alleles located next to each other.

1.8 The Y chromosome

The Y chromosome, the smallest chromosome in the human genome, is a distinctive genetic characteristic of males. It contains about 60 million base pairs but only a few genes [25]. Architecturally, the nonrecombining region (NYR), also referred to as the male-specific (MSY) region, is the largest portion of the chromosome and is 95% of its total length. The whole MSY region is in linkage disequilibrium and is inherited along the paternal lineage from father to son with no recombination [147]. The rest of the Y chromosome includes two pseudoautosomal regions (PAR1 and PAR2) located at the distal ends and encompassing at least 29 genes within a length of less than 3 Mb of the total 60 Mb length of the chromosome [148] (**Figure 1.18**). These regions are homologous to the X chromosome pseudoautosomal regions and are crucial for pairing of the X and Y chromosomes during meiosis. These regions therefore undergo recombination during meiosis, and thus are inherited in an autosomal rather than sex-linked manner. Reports also suggest the existence of a third pseudoautosomal region (PAR3) located on the short arm of the chromosomes (at Yp11.2), which may be found in 2% of the general population [149]; however, its functional role is yet unknown [148].

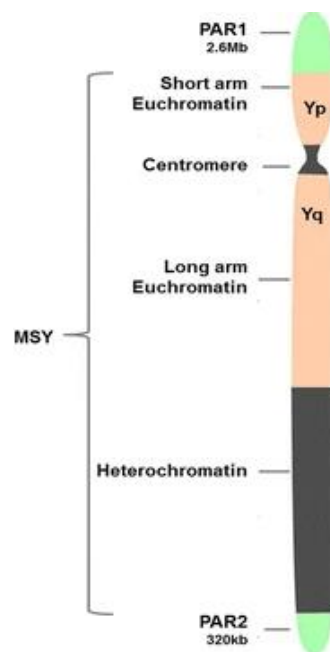


Figure 1.18. The structure of the Y chromosome.

The largest portion of the Y chromosome is the male-specific region (MSY) which does not undergo recombination. Cytogenetically, the region is separated into heterochromatin located in the long arm (Yq) and a euchromatic region. The latter is located on both the long (Yq) and short (Yp) arms, separated by the paracentromeric region around the centromere. At the terminal tips are the short pseudoautosomal regions (PAR1 and PAR2), which are homologous to the X chromosome and undergo recombination during meiosis (modified from [148]).

The unique properties of the Y chromosome – its inheritance in a non-recombining pattern and its variation caused by mutations through generations along the paternal line – have generated interest from scientists in studying polymorphisms in the MSY region, especially with increased amounts of genomic data from next-generation sequencing studies. Subsequently, these polymorphisms have been used in studies of human history [150], evolutionary genetics [151], medical genetics [148], genealogical reconstruction based on surnames [152, 153], and forensic genetics [154].

1.8.1 Y-chromosome markers

Even with the great improvements made to recently released STR kits, such as AmpFISTR® GlobalFiler® and PowerPlex® Fusion, in terms of the large number of markers and their sensitivity in retrieving profiles from challenging samples, it is still impossible to selectively amplify male DNA from samples that contain a mixture of female and male DNA. Forensic kits designed specifically for the Y chromosome are therefore valuable tools for analysing these types of samples.

The Y-chromosome-specific polymorphism typing approach is frequently used as a significant tool in forensic investigations to identify the male contributor DNA in sexual assault cases. In these types of cases, mixed stains are the common biological material available, with the female DNA proportion often being higher than the male, or the quantity of male DNA being low. This often results in profiles representing only the female DNA or mixed profiles having a high contribution of female DNA and low levels of male DNA that are difficult to interpret when genotyping using conventional autosomal STRs. In these cases, the analysis of the Y chromosome allows direct analysis of the male. In addition, paternal transmission of the Y chromosome allows identification of the patrilineality of the family lineage, which can be applied to specific scenarios in paternity investigations such as cases where the father/mother is deceased/not available for testing. It is also applicable in missing person identification, the study of human migration patterns, and historical and genealogical research [26].

The two widely used Y-chromosome genetic markers are multiallelic STRs (Y-STRs), which produces a single haplotype, and biallelic SNPs (Y-SNPs) which define haplogroups (haplotype

groups that share a common ancestor); the former displays a much higher mutation rate than the latter. Forensic samples are mostly analysed using Y-STRs due to their allelic diversity, small size (~100–400 bp), and ability to be multiplexed in a single PCR assay, whereas Y-SNPs can be useful for inference of paternal biogeographic ancestry [154].

1.8.1.1 Y-chromosomal short tandem repeat (Y-STR) markers

The discovery of the first polymorphic STR on the MSY region of the Y chromosome in 1992 [155] and its utility in analysing casework samples [156] subsequently triggered the development of more Y-STRs for forensic applications. Their valuable property of tracking the genetic relationship between two or more males by the distinctiveness of their haplotypes allows Y-STRs to be applied in a range of forensic analyses such as human identification, paternity testing and familial searching [157]. In addition, their haplotype diversity, which generates different patterns of genetic variation between populations, has been used to study human anthropology and for deconvoluting human population history by estimating the relationships between different populations/individuals [158, 159]. It is estimated that there are 4,500 STRs and millions of SNPs on the Y chromosome, all of which are variable due to mutations that occur independently on the non-recombining region; however, the mutation events of Y-STRs are more rapid than those of SNPs [160].

In a large mutation rate study conducted to genotype 186 Y-STR loci in 2000 father–son pairs, The loci showed mutation rates ranging from $\sim 10^{-2}$ to 10^{-4} per locus per generation [161]. The different mutation rates make it interesting to study both Y-STRs with slow mutation rates to increase the likelihood of finding relatives and Y-STRs with fast mutation rates to differentiate male relatives. The variability in the mutation rate within Y-STRs has allowed the identification of 13 highly mutating Y-STRs, named Rapidly Mutating (RM) [161]. Due to their high mutation rate, RM Y-STR markers possess high haplotype diversity (high allelic range) and provide great discriminatory power, which can increase the likelihood of distinguishing distant relatives from closely related ones, and can help in cases such as those that involve unrelated persons sharing similar haplotypes (homoplasmy cases) [162]. Therefore, kits such as PowerPlex® Y23 and AmpFISTR® Yfiler® Plus have been developed recently to incorporate a greater number of slowly mutating Y-STRs to increase the likelihood of finding male relatives, with a small number of RM Y-STRs to distinguish between closely related individuals.

1.8.1.2 Y-STRs kits

The development of Y-STR kits has gone through several stages starting in 1997, when an international group of laboratories that had been working with Y-STRs established a set of nine markers (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a and DYS385b) as the European minimal core haplotype of Y-STRs for forensic use [163]. In 2003, the Scientific Working Group on DNA Analysis Methods (SWGDAM) announced a panel containing 11 Y-STRs, including the European set of nine with an additional two loci, DYS438 and DYS439 [164]. Subsequently, Promega Corporation and Applied Biosystems developed two kits, the PowerPlex® Y system (12 loci) and AmpFISTR® Yfiler® (17 loci) respectively, which both incorporated the 11 Y-STRs with one additional locus in PowerPlex Y® and six additional loci in Yfiler®. Further development has brought a newer generation of kits: the PowerPlex® Y23 and AmpFISTR® Yfiler® Plus which have expanded multiplexing capacity to 23 and 27 Y-STR loci, respectively. To augment the power of identification between unrelated and especially related individuals, several RM Y-STR markers have also been included in the PowerPlex® Y23 (six RM Y-STRs) and Yfiler® Plus (three RM Y-STRs). **Figure 1.19** shows the commonly analysed forensic Y-STRs and their approximate location on the Y chromosome.

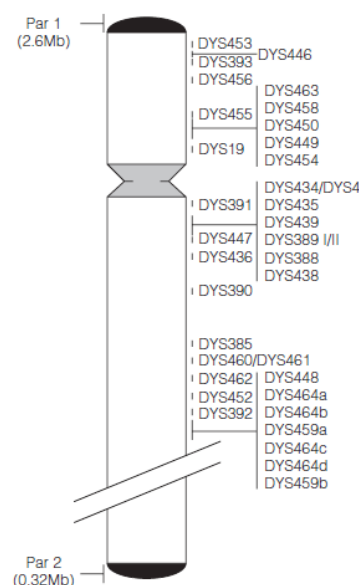


Figure 1.19. The locations of commonly used Y chromosome short tandem repeat (Y-STR) markers.

The approximate position of Y-STR markers, which are often used in forensic applications, on the NYR region of the Y chromosome (source [165]).

1.8.2 Y-chromosomal single nucleotide polymorphism (Y-SNP) haplogroups

Compared to Y-STRs, Y-SNPs show a lower mutation rate of $\sim 3 \times 10^{-8}$ mutations per base per generation [166], allowing ancestral signatures to be retained much longer in Y-SNPs than in Y-STRs, via population-specific paternal lineages (generally termed haplogroups) in a stable pattern through human history [154]. According to De Knijff, a Y-chromosome haplogroup refers to a lineage defined by binary markers, mostly SNPs, whereas the term haplotype is reserved for all sublineages of haplogroups that are defined by the variation in Y-STRs [167, 168]. Haplogroups have a high level of specificity between geographic regions and their distributional pattern can be used to study human history, colonisation, and migration [154, 169, 170]. In addition, the study of slowly evolving Y-SNPs with low mutation rates has led to the construction of robust and detailed Y-chromosomal phylogenies which have become a major part of the study of demographic history [171].

In forensic genetics, Y-chromosome haplogroups are used to define paternal lineages shared between closely and distantly related males. Thus, the numerous Y-SNP haplogroups can be used to uncover information about paternal geographic ancestry [154]. The analysis of Y-SNPs to track the ancestry of an unknown DNA donor is mainly performed by genotyping a small number of SNPs in a hierarchical manner using SNaPshot® minisequencing technology [172]. Consequently, many multiplexes have been developed to detect the major Y haplogroups, and the more precise sub-haplogroups (more geographically restricted) within these [172-175]. Since the hierarchical approach requires the use of several multiplexes to obtain the final haplogroup and SNP multiplexing is limited to a few dozen markers, it is generally not feasible in forensic practice. Next generation sequencing (NGS) technology has overcome these limitations by increasing multiplexing capability to target thousands of Y-SNPs in a single assay, allowing detailed and high-resolution haplogroup detection [176, 177].

1.8.2.1 Y-SNP haplogroup inference and prediction

The use of Y-SNPs in forensic casework and research for paternal lineage classification and geographic ancestry inference relies on universal phylogenetic trees of Y chromosomes and the characterisation of their evolutionary lineages [178-180]. Before 2008 when the last publication of the Y Chromosome Consortium (YCC) was issued [181], scientific questions about Y-chromosome variations were answered using one reference phylogenetic tree, the

YCC tree, constructed using the binary Y-chromosomal markers validated at that time and updated regularly [169, 181, 182].

The wave of genomic data from massively parallel sequencing studies has led to several phylogenetic trees with newly discovered Y-SNPs. For instance, a large detailed phylogenetic tree which represents a strong evolutionary framework was constructed with more than 13,000 high-confidence SNPs by sequencing 3.7 Mb of the MSY region in 448 human males [183]. However, many of the Y-SNPs used to construct such phylogenies are not conditioned for forensic use because they are rare and estimating their allele frequencies across human populations could be challenging, if practical at all [184].

It has been suggested that the most pragmatic approach for forensic geneticists is to use a reduced consensus tree such as the YHRD (Y-STR Haplotype Reference Database [185]) comprising a limited number of Y-SNPs with a clear phylogenetic structure and male lineage nomenclatures, and which captures the more frequent Y haplogroups [171, 184].

While Y-SNPs can define haplogroups with high geographical specificity, Y-STR haplotypes can also be used to predict haplogroups. Software tools for haplogroup prediction are widely used to determine the most likely haplogroups using a set of Y-STR haplotypes [186-189]. Examples of such tools are Whit Athey's Haplogroup Predictor (<http://www.hprg.com/hapest5/>) [190, 191], Jim Cullen's World Haplogroup & Haplogroup-I Subclade Predictor (<http://members.bex.net/jtcullen515/haplotest.htm>) and NevGen Y DNA Predictor (<https://www.nevgen.org/>). Factors thought to influence the accuracy of the haplogroup prediction software include the Y-STR profiles associated with determined haplogroups, the similarities of haplotypes from different haplogroups, and the size and composition of the databases consulted by the software [189, 192]. Athey proposed that the accuracy of predicting haplogroups could exceed 99% when using a large number of Y-STRs [190, 193], and small numbers of Y-STRs (e.g. 12) are not recommended for use in haplogroup prediction as they lack accuracy [187].

1.9 The X chromosome

The unique characteristics of the X chromosome generally arise from how it is inherited. Females have two homologous X chromosomes, whereas males are hemizygous for the X chromosome. In every generation, the female can pass either X chromosome to the next

generation while the male passes only the one (two thirds of the total). Therefore, X chromosome polymorphisms in a population are mostly reflected in female histories [194]. The X chromosome has intrinsic features, including lower recombination and mutational rates and smaller effective population size, which results in more rapid genetic drift compared to autosomal chromosomes. These factors contribute to higher linkage disequilibrium (LD) and more pronounced population structure than the autosomes [194]. X-chromosome genetic markers have been used in studies of genetic diversity and population differentiation [195] and in medical genetics for the diagnosis of X-linked diseases [196].

1.9.1 X-chromosome markers

In forensic science, analysis of the X chromosome can be beneficial in complex paternity cases such as parent/child relationship investigations that involve at least one female relative such as mother/son, mother/daughter, and father/daughter [197]. The benefits can be even greater in investigations of paternal grandmother/granddaughter relationships, particularly in deficiency cases where the father is deceased/not available and the paternal grandmother's DNA can be used to reassociate the granddaughter since one of the X chromosomes is transmitted from the grandmother to the granddaughter through the father [198, 199]. In addition, the X chromosome can be more informative than other marker systems in cases of alleged incest [200] and criminal rape [201] and can provide additional information in missing person investigations [202].

Short tandem repeat markers on the X chromosome (X-STRs) have been increasingly used as a supplementary typing system to traditional STR markers, whereas other X chromosome markers such as SNPs (X-SNPs) have been used to a lesser extent [203-205]. Due to the low recombination rate of the X-chromosome as well as the physically close location of some X-STRs, it has been suggested that the closely located X-STRs can be treated as a Linkage Group (LG) instead of as independent loci [206]. Alleles at linked loci that are physically close combine to form haplotypes that may be inherited as blocks during meiosis. To calculate match probabilities, the frequencies of different linkage groups, rather than alleles, are determined and then multiplied together across linkage groups [198]. Four linkage groups consisting of 12 X-STRs in total are widely used in forensic kinship investigations, namely DXS10148, DXS10135 and DXS8378 (linkage group 1), DXS7132, DXS10079 and DXS10074 (linkage group 2), DXS10103, HPRTB and DXS10101 (linkage group 3), and DXS10146,

DXS10134 and DXS7423 (linkage group 4) [207]. These markers are incorporated into the Investigator® Argus X-12 QS kit, which is commercially available from Qiagen. **Figure 1.20** shows the markers and their relative position on the X chromosome.

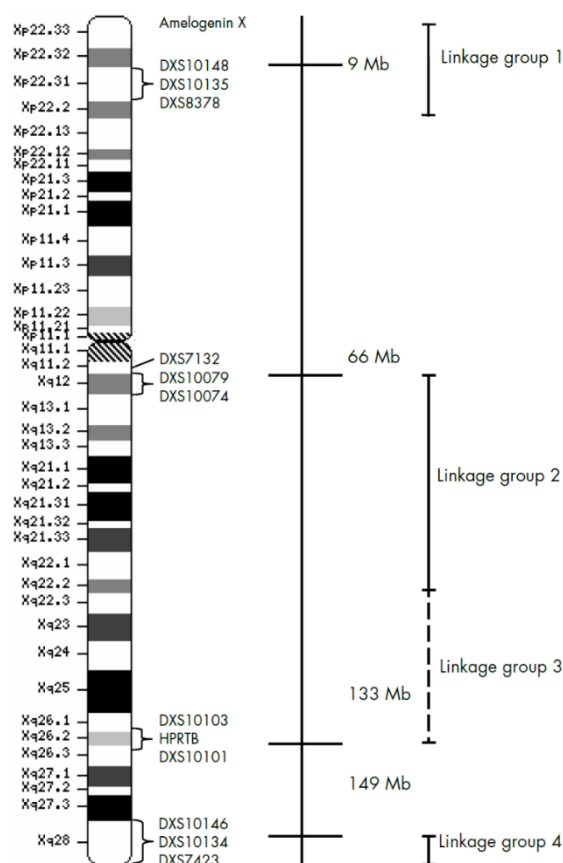


Figure 1.20. X chromosome short tandem repeat (X-STR) markers in the Qiagen Investigator® Argus X-12 QS kit.

Ideogram of the X chromosome and relative cytogenetic positions of the X-STRs in the Argus kit. Each linkage group is composed of tightly localised loci that are genetically linked (source: Investigator® Argus X-12 QS Handbook).

1.10 Mitochondrial DNA (mtDNA) markers

The majority of the DNA in cells exists within the nucleus. However, the mitochondrion, an organelle in the cell cytoplasm responsible for energy production, contains its own small genome in the form of mitochondrial DNA (mtDNA). The complete nucleotide sequence of human mtDNA was sequenced by Anderson et al. in 1981; and is referred to as the Cambridge Reference Sequence (CRS) [208]. In 1999, the CRS was upgraded to a newer version; the revised Cambridge Reference Sequence (rCRS), which is the one accepted and used today [209]. The human mtDNA genome is a double-stranded DNA molecule of approximately 16,569 bp in length, which is histone-free and has a closed circular form. The two strands of

the mtDNA duplex can be differentiated according to their sequence composition. The heavy strand is rich in purines (guanine and adenine) whereas the light strand is rich in pyrimidines (cytosine and thymine). The mtDNA coding region accounts for about 93% of the genome and contains 37 genes that encode 22 tRNAs, two RNAs (12S and 16S subunits), and 13 polypeptides. The rest of the genome (~7%) is free of coding sequences and is known as the control region, or is sometimes referred to as the displacement loop or the D-loop [208] (**Figure 1.21**).

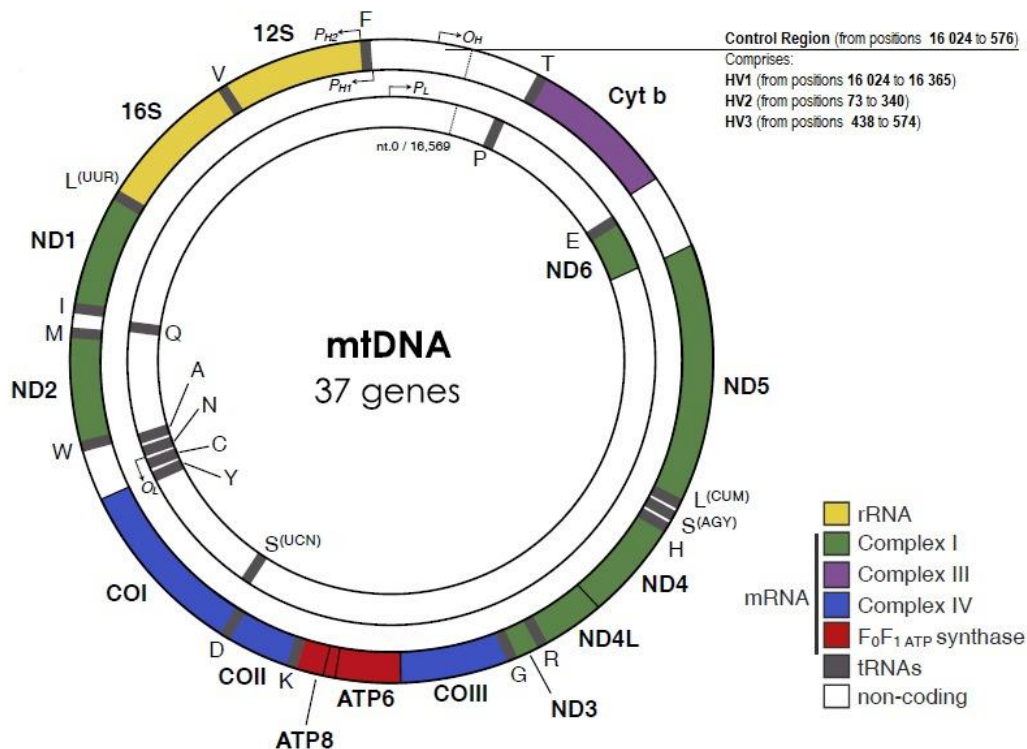


Figure 1.21. The human mtDNA genome with genes and the control region.

The outside circle represents the heavy strand (H) where most of the coding genes are found while the inner circle represents the light strand (L) with fewer genes and a larger non-coding segment. The 37 mtDNA genes are colour coded. Forensic mtDNA analysis targets HV1, HV2, and sometimes HV3 in the control region, which are shown with their positions (source [210]).

The control region contains specific segments that are highly variable between individuals, namely hypervariable region 1 (HV1; position 16,024-16,365) and hypervariable region 2 (HV2; position 73-340) [211] (**Figure 1.21**). In general, the mutation rate of mtDNA is ten-fold higher than that of nuclear DNA due to its rudimentary repair system and the damage caused by reactive oxygen species (ROS) during oxidative phosphorylation [212, 213], which leads to the accumulation of mutations within cells and the germ line [214]. In addition to the two hypervariable segments, a third region (HV3; position 438-574) with slightly lower variability

has been found to be useful for differentiating between samples that have similar sequences in HV1/HV2 [215]. Its small size and the relatively high variability of its sequences make the control region an attractive target for forensic applications.

A mitochondrion contains an average of five copies of the mtDNA genome, and depending on the cell type, each cell can have hundreds to thousands of mitochondria [216, 217]. These characteristics enable mtDNA to be used in typing forensic samples with degraded or a limited amounts of DNA, such as poorly preserved skeletal remains [218] and hair shafts [219]. Another important feature of mtDNA is its mode of inheritance; mtDNA is exclusively inherited maternally and all male and female relatives within the same maternal lineage will share an identical mtDNA sequence [220-222]. Although the male sperm carries a small number of copies of mtDNA (about 2-100), the abundance of mtDNA in the female oocyte (> 100,000) makes the few copies of male mtDNA negligible after fertilisation [213, 223, 224]. Like the Y chromosome, the mtDNA sequence is treated as a haplotype which is defined by the sequence differences relative to the Revised Cambridge Reference Sequence [209].

The unique transmission of mtDNA can be useful in forensic investigations in cases in which maternal relatives, even several generations back, can be investigated to identify missing persons. This was demonstrated in the recent case of the skeletal remains of Richard III. The king's maternal relatives' mtDNA was sequenced and then compared with the sequence from the skeleton. A perfect match was found between the mtDNA sequences indicating that the remains belonged to Richard III [225]. For human identification, mtDNA is analysed mainly by Sanger sequencing targeting HV1 and HV2 in the control region [226].

1.11 DNA sequencing

In the 1970s, two technologies were introduced to 'read' or sequence DNA. Both methods – the dideoxy chain termination method engineered by Sanger known as Sanger sequencing and the chemical modification method developed by Maxam and Gilbert called Maxam–Gilbert sequencing – used radioactive labels to infer the order of nucleotides. Sanger sequencing involves the target DNA being denatured and annealed to a tagged primer and then extended by DNA polymerase incorporating chain-terminating nucleotides to generate different-sized fragments [227]. Maxam–Gilbert sequencing involves digestion by a restriction enzyme of a terminally-tagged double-stranded DNA fragment and using different

combinations of chemicals to create base-specific segments of DNA [228]. In both methods, base order is determined by separation of DNA fragments on polyacrylamide slab gels using electrophoresis.

By 1987, several enhancements had been made to the Sanger method which mainly involved using a fluorescent dye instead of a radioactive label in a single reaction and developing fragment detection using CE [229, 230]. Subsequently, the first generation of automated fluorescence-based sequencing was developed by Applied Biosystems (ABI) using the Sanger technique [231, 232], which was the quantum leap that forever changed the progress of DNA sequencing technology. These improvements led to substantial growth in sequencing data from half a million bases in 1982 to around ten million bases by 1986 [233]. Sequencing instruments were also improved in the following years. The first capillary sequencer, the ABI 310, was limited to one capillary, whereas the improved model, the AB 3730xl, has 96 capillaries, can deal with larger fragment sizes and has a shorter run time [234]. Such improvements had a significant impact by speeding up the sequencing process of the Human Genome Project, the largest and most valuable project based on the Sanger sequencing technique.

However, the Sanger method suffers from some serious drawbacks, such as its high cost and low sequencing throughput, laborious operating procedures, and its limitations in analysing complex genomes [235, 236]. Fortunately, the past few decades have seen the introduction of massively parallel sequencing (MPS), also termed next-generation sequencing (NGS), with an improved DNA sequencing technique that outperforms the traditional Sanger method in terms of cost, throughput capacity and sequencing efficiency. This has allowed rapid progress in many biological fields such as clinical genetics [237, 238], human history [239], ancient DNA analysis [240], microbiology [241], genetics [242] and forensic science [243, 244].

1.11.1 Next-generation sequencing (NGS) in forensics

Application of Sanger sequencing in forensic genetics has been limited to the analysis of mtDNA [245] and SNPs [111]. The technique is not used as a standard DNA typing system due to its complex methodology, low throughput and high cost. On the other hand, NGS is a high-throughput sequencing technology that sequences DNA templates along the human genome in a massively parallel manner to generate millions of small DNA fragments (reads) [246]. NGS

can be used to sequence the whole genome as well as targeted regions in the genome. It is capable of detecting large numbers of SNPs, indels, structural variations (SVs) and copy number variations (CNVs) [236].

Currently, polymerase chain reaction-capillary electrophoresis (PCR-CE) is the dominant system used to type forensic markers such as STRs, SNPs and indels. STR markers, including autosomal STRs, Y-STRs and X-STRs, are commonly used in almost every forensic laboratory, and are considered the gold standard for DNA investigation due to their informativeness (high discriminatory power and diversity) and the simple workflow (PCR-CE) involved, as well as the existence of well-established global databases. Autosomal STRs are routinely used for both identification of individuals and paternity testing, whereas Y-STRs and X-STRs are less frequently used. However, they are applied in specific scenarios; for instance, Y-STRs are regularly used in sexual assault cases to detect the male contributor of DNA and in paternity cases to trace paternal relationships, whereas X-STRs are used in complex paternity/kinship cases. In addition, SNP/indel markers are used for degraded samples that cannot be typed by conventional STRs, or as additional markers in challenging kinship cases in which mutations exist in familial STR profiles.

It should be noted that each class of genetic marker requires an individual assay; thus, additional PCRs together with more of the often-limited DNA sample are needed for PCR input, which limits multiplexing capacity when different markers cannot be typed together. The use of NGS can overcome this limitation by allowing hundreds of different classes of forensic markers to be combined into a single assay. This in turn saves time, especially when additional DNA investigations are needed.

When using PCR-CE to genotype STR markers, partial profiles are often observed, especially in degraded or old samples, due to the failure to amplify large STR fragments. NGS offers an improved system in which the amplicons for different classes of markers are designed to be as short as possible to improve DNA recovery from challenging samples. In addition, NGS can detect the actual nucleotide sequences of STR alleles, rather than just their length. This substantially increases the information obtained and, consequently, the discriminatory power of forensic testing, in addition to improving mixture interpretation and assisting in paternity

testing. For example, NGS can distinguish between two persons with the same-length allele(s) at an STR locus if the nucleotide sequence of the allele(s) differs between them.

1.11.2 Forensic NGS platforms and commercial kits

Thermo Fisher Scientific (MA, USA) and Illumina (CA, USA) are leading the way in developing benchtop NGS instruments suited for forensic applications. Thermo Fisher Scientific has already launched several NGS panels designed to sequence forensic markers including mtDNA, STRs and SNPs. The panels are sequenced using Ion Torrent platforms, including the Ion Torrent PGM (Personal Genome Machine) and the Ion S5. Each forensic panel is designed to sequence one class of markers in an individual assay; for example, the Precision ID GlobalFiler NGS STR Panel includes only autosomal STRs and the Precision ID Identity Panel includes only SNPs. Thermo Fisher has not released any forensic NGS assays that incorporate different classes of markers in a single test.

Verogen, Inc. (CA, USA) with a focus on forensic genomics recently launched the MiSeq FGx[®] system, a massively parallel sequencing system created by Illumina for forensic applications [247]. The FGx[®] system comes with the ForenSeq[™] DNA Signature Prep Kit, which contains two primer sets: primer mix (A) to amplify 153 markers, including amelogenin (the sex chromosome marker), 27 autosomal STRs, 24 Y-STRs, 7 X-STRs and 94 identity SNPs; and primer mix (B) containing all the markers in primer mix (A), in addition to 22 phenotypic SNPs and 56 biogeographical ancestry SNPs (a total of 231 markers). Recently, Promega Corporation (Madison, USA) also developed a beta version of PowerSeq Auto/Y, an NGS kit compatible with the MiSeq System, which can amplify 22 autosomal STRs, 23 Y-STRs and amelogenin [248].

1.11.3 Overview of the MiSeq FGx[®] workflow

The MiSeq FGx[®] Forensic Genomics System consists of five components: (1) MiSeq FGx[®], a benchtop sequencing instrument; (2) a reagent cartridge containing the required reagents for cluster generation and sequencing; (3) a flow cell, which is a glass slide containing fluidic channels through which the sequencing reactions occur; (4) ForenSeq primer mixes; and (5) ForenSeq Universal Analysis Software. The ForenSeq DNA Signature Prep Kit is optimised to analyse 0.2-1 ng of extracted DNA. Moreover, a small 1.2-mm punch from an FTA card can be directly added to the reaction without the need for DNA extraction or quantification [249].

The FGx® instrument with the ForenSeq kit can sequence up to 96 samples (libraries) using 200 amplicons ranging in size from 63 bp to 180 bp in a single run [250]. The amplified libraries are sequenced using the Illumina sequencing workflow, which involves four steps: template and library preparation; cluster generation; sequencing; and data analysis.

1.11.3.1 Template and library preparation

The ForenSeq DNA Signature Prep Kit primer mixes have a specific set of tagged forward and reverse primers designed to amplify the targeted STRs and SNP regions in the DNA sample. Through the first set of PCR cycles (PCR1), these regions are amplified and tags are added, and then the amplicons are used as templates for the second set of PCR cycles (PCR2) in which specific indexed adapters (barcodes) are attached to the amplified regions (**Figure 1.22**). Because each sample is assigned a unique barcode (i5 and i7 indices), the instrument software can identify and separate each sample according to their associated barcodes.

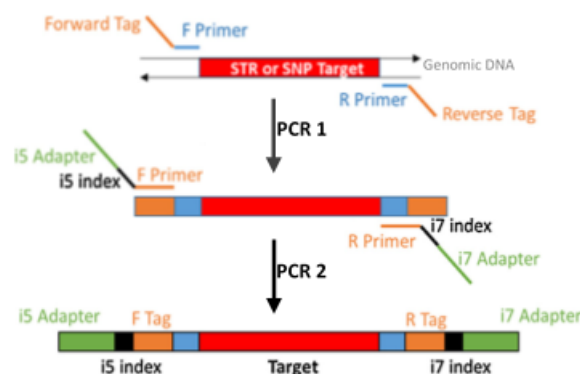


Figure 1.22. MiSeq FGx® target amplification and library preparation.

Two PCR rounds are required to produce the amplicon libraries. The first is to amplify the STR and SNP targets using tagged primer pairs (PCR 1). The second is to attach a sample-specific index and adapter 'barcodes'. Barcodes are used to identify and sort the reads after sequencing (modified from [247]).

1.11.3.2 Cluster generation by bridge amplification

When libraries are introduced to the flow cell, they hybridise to oligonucleotides (primers), complementary to their adapter sequences, which cover the MiSeq FGx® flow cell surface to form new strands. The new strands are then subjected to bridge amplification, a process in which the strands are bent over to hybridise to nearby complementary oligonucleotides and replicated (copied), before being denatured (**Figure 1.23**). As the process is repeated, millions of clusters (dense groups of clones of all DNA fragments) are generated on the flow cell surface. All the reverse strands are then washed away leaving only the forward strands ready to be sequenced.

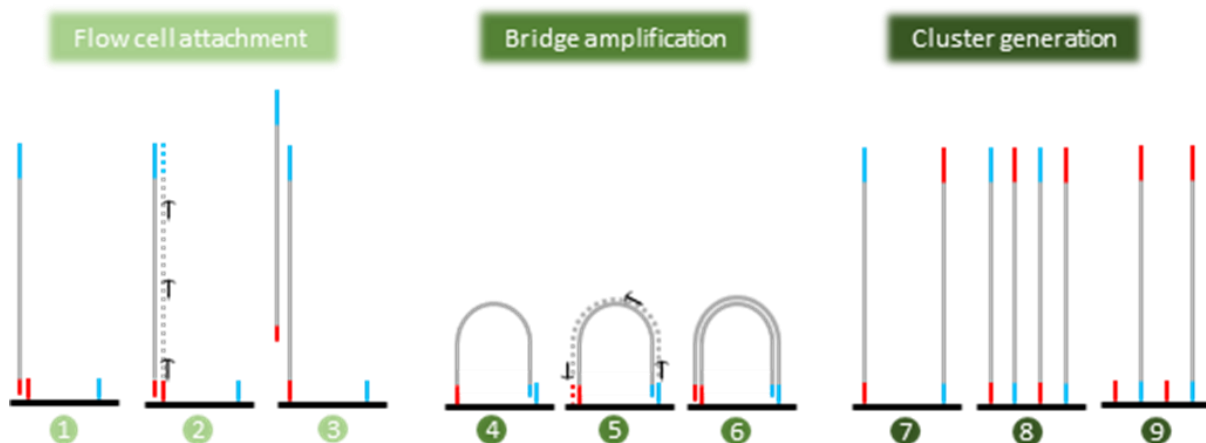


Figure 1.23. Cluster generation by bridge amplification.

(1) Target-specific oligonucleotide primers on the flow cell surface hybridise to complementary adapters on the DNA fragments. (2) Copies of the hybridised fragments are created by PCR polymerases forming double-stranded DNA which is then denatured. (3) The original strands are washed away. (4) The new strands fold over and their adapters hybridise to complementary primers on the flow cell surface. (5) Copies of the folded strands are made by polymerases, (6) forming a double-stranded DNA bridge. (7) Each bridge is denatured leaving both strands anchored to the flow cell surface. (8) The process is repeated, forming millions of clusters of DNA fragments on the flow cell surface. (9) All the reverse strands are cleaved by specific base cleavage and washed away, leaving only the forward strands (drawn using information on www.illumina.com).

1.11.3.3 Sequencing by synthesis

The MiSeq FGx[®] instrument sequences by synthesis, performed using a reversible termination process and four fluorescently labelled deoxynucleotides (dNTPs), each attached to chemical terminators (blocking groups). Sequencing is initiated when sequencing primers anneal to their complementary sequences on the forward strands on the flow cell. The primers add a single complementary fluorescently labelled dNTP to each DNA strand while the blocking groups halt further extension of the DNA. This allows incorporation of only one dNTP to each strand per cycle. When a nucleotide is attached, a fluorescent signal is emitted and then signals are captured in an image simultaneously for all the clusters. The next cycle begins after the cleavage of the blocking groups and the incorporation of new nucleotides (**Figure 1.24**). The resulting images are used to identify base calls of each DNA cluster sequence depending on their wavelengths and signal intensity.

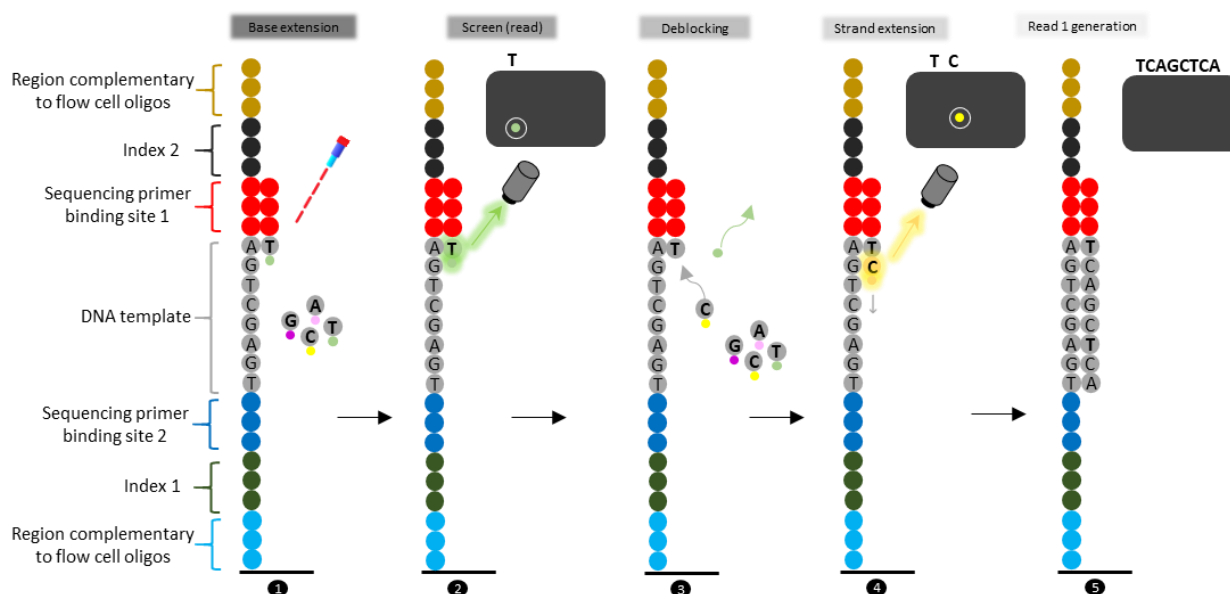


Figure 1.24. Sequencing by synthesis using reversible terminators.

(1) Sequencing primers attach to the complementary binding site of the DNA strand. This initiates the first cycle of base extension by adding the correct deoxynucleotide (fluorescently labelled with attached blocking group). When a light beam is introduced, (2) a fluorescent signal is emitted and screened. The captured images are used to identify the bases that were incorporated. (3) The fluorescent dyes, blocking groups and unincorporated dNTPs are washed away. The next cycle begins by adding another complementary base. (5) DNA elongation continues until completion of the first read. The resulting images are used to generate base calls of each DNA cluster sequence (drawn using information on www.illumina.com).

1.11.3.4 Indices and paired-end sequencing

The index sequences attached to the DNA samples during library preparation are used to identify each DNA sample (library). The ForenSeq DNA Signature Prep Kit uses 12 unique 8-bp Index 1 (i7) sequences and eight unique 8-bp Index 2 (i5) sequences to generate up to 96 dual-indexed libraries. Index 1 and Index 2 are sequenced separately to produce the Index 1 Read and then the Index 2 Read (**Figure 1.25**).

The MiSeq FGx® uses paired-end sequencing technology which allows both strands (forward and reverse) to be sequenced. After completion of forward strand sequencing, reverse strands which were washed away during cluster generation, are replicated and sequenced to generate Read 2 (**Figure 1.25**).

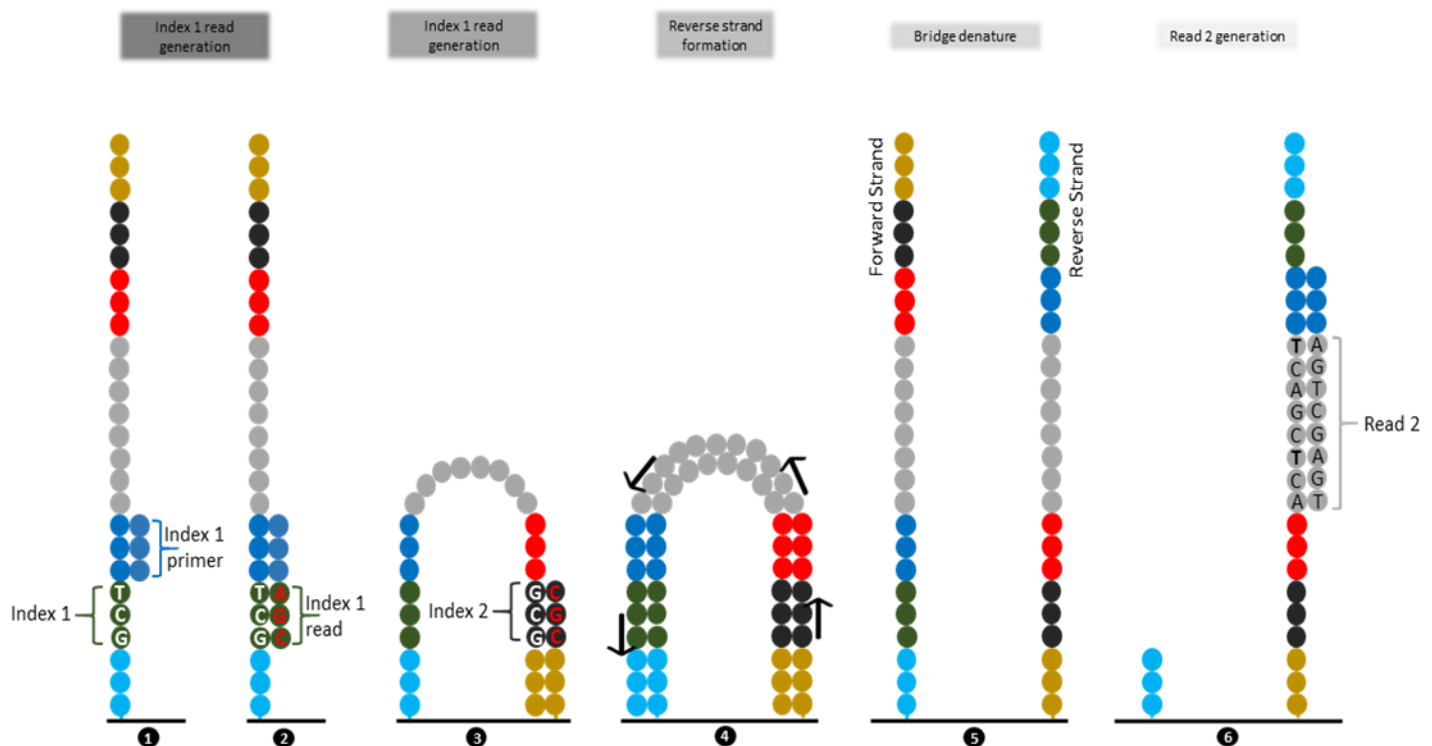


Figure 1.25. Indices and paired-end sequencing.

(1) The Index 1 primer attaches to the template DNA, and (2) begins adding complementary bases to produce the Index 1 Read. Once completed, the Index 1 Read is washed away. (3) The sequenced forward strand folds over to hybridise to the complementary primer on the flow cell, and the Index 2 Read is generated in the same manner as Index 1. (4) Reverse strand formation begins when primers replicate the forward strand to generate double-stranded DNA bridges. (5) The bridge is linearised, and (6) the forward strand is washed away leaving only the reverse strand. The sequencing primer hybridises to generate Read 2 by synthesis (like Read 1). After completion, Read 2 is washed away. The process of sequencing both strands is called paired-end sequencing and generates millions of high quality and accurate reads (drawn using information on www.illumina.com).

1.11.3.5 Data analysis

The massive amount of raw data produced by this sequencing-by-synthesis method contains every DNA sequence read for all the samples sequenced. It is not practically feasible to review raw sequence data manually and, thus, bioinformatics tools are commonly used to analyse and convert raw data to a manageable form. Two of the bioinformatics tools typically used for raw data analysis are a mapping tool and a variant-calling tool. The mapping programmes align (match) all sequenced reads to a known human reference genome so that the exact location of each sequence read can be identified. The variant-calling programmes then compare the aligned reads with the reference genome and determine where they differ. The raw data produced by MiSeq FGx® using the ForenSeq DNA Signature Prep Kit are automatically analysed, and the results are visualised using the ForenSeq Universal Analysis

software (Verogen, CA, USA), a programme with a graphical interface that shows allele/genotype calls.

1.11.4 ForenSeq DNA Signature Prep Kit studies

Recently, a considerable amount of literature has been published assessing the forensic utility of the ForenSeq DNA Signature Prep Kit with the MiSeq FGx® sequencing instrument. Much previous research has focused on efficiency assessments in relation to sensitivity by studying different DNA concentrations, reproducibility by sequencing replicated samples, concordance by comparing data to that produced with CE, and mixture analysis by using different ratios of contributors' DNA [251-257]. These studies indicate that the MiSeq FGx® system with the ForenSeq DNA Signature Prep Kit can be successfully used in applications for forensic DNA analysis.

In one of the earliest studies in this field, Churchill et al. validated the beta version of the ForenSeq DNA Signature Prep Kit by performing different experiments to test for reproducibility, sensitivity, mixed samples, concordance and the ability to analyse challenging samples [258]. They obtained reliable and full profiles from 1 ng of input DNA, identified minor and major contributors' alleles in mixed samples at different DNA ratios, and generated more alleles from challenging samples compared to conventional CE. The authors confirmed the validity of the beta version of the ForenSeq DNA Signature Prep Kit for use in forensic DNA investigations. Furthermore, Caratti et al. carried out another study with the beta version of the kit using samples that had already been typed with CE, and showed full genotyping concordance of all STR markers between both technologies [259].

In another significant study, Jager et al. reported developmental validation of the MiSeq FGx®/ForenSeq DNA Signature Prep Kit system based on SWGDAM (www.swgdam.org) guidelines [247]. They performed several experiments to validate the kit, including assessment of species specificity, DNA sensitivity, degraded samples, the impact of PCR inhibitors, DNA mixtures and concordance. As a result, sensitivity studies showed that accurate genotypes were obtained from samples with DNA input ranging between 62.5 pg (0.0625 ng) and 1 ng. In the mixture studies, the system was able to detect the alleles of the minor contributor when present at less than 5% of the major contributor. The results also indicated 100% concordance of STR genotypes with kits using CE in more than 1000 samples

tested. These findings confirm that the MiSeq FGx® sequencing platform with the ForenSeq DNA Signature Prep Kit meets forensic DNA validation and quality assurance guidelines, given that the validated NGS system is capable of providing accurate and reliable results and has great potential in forensic DNA analysis. Numerous studies have reported robust performance of the NGS kit for forensic applications.

In studies of degraded DNA, the ForenSeq DNA Signature Prep Kit was used to identify individuals from mass graves by obtaining profiles with a success rate of more than 90% [260]. It was also able to detect more STR loci than CE-based methods when old skeletal remains were sequenced [261]. In another study, STRs and SNPs were successfully detected by sequencing highly degraded DNA from poorly preserved ancient skeletal remains that failed to be typed by conventional STR markers [255]. Studies on paternity testing have shown the efficiency of the kit to solve relationship cases in which traditional STR markers assayed using CE were inconclusive due to mutations in family profiles, and in cases in which additional markers were needed to provide more supporting evidence [262, 263].

In a study investigating ancestry, Ramani et al. used the biogeographical ancestry SNP markers contained in the kit to differentiate between ethnic groups in the Singapore population [264]. Ancestry prediction was found to be improved by the inclusion of flanking region variants (surrounding the ancestry SNPs) in the genotypic data, when compared to using the ancestry SNPs alone [265].

1.11.5 Forensic NGS improvements

While there is great interest from forensic laboratories and agencies in adopting this technology, the data produced by NGS must be universally accepted by the forensic community. Therefore, the DNA Commission of the ISFG has made recommendations for standardising the description (nomenclature) of complex STR sequences produced by NGS [266]. This will allow laboratories to store NGS data in a consistent format and facilitate data interpretation.

For the sequencing system to be implemented in laboratories, population studies must be carried out to produce allele frequencies that enable calculation of statistics to estimate the strength of DNA profile evidence. Population studies are also necessary to assess the performance, in that specific population, of each marker contained in the NGS kit to provide

guidelines and information for data interpretation. To date, several studies have tested the performance of the ForenSeq DNA Signature Prep Kit markers in different populations [267-275]. They demonstrated that sequence variants within both STRs and flanking regions (indels or SNP variants) provide additional allele diversity compared to CE-based methods. The increased number of genetic variants detected within forensic marker regions using NGS increases the power of this kit as a forensic DNA test.

Currently, one of the main challenges facing forensic laboratories wishing to implement NGS is the laborious workflow involved, which is not feasible for many. It is anticipated that this methodology would become increasingly straightforward in the coming years, as the technology is rapidly improving; this would have a positive impact on forensic genomics applications. Moreover, the high cost of sequencing (instrument and kits) hinders its implementation in many forensic laboratories. Again, it is anticipated that the system would become increasingly affordable as the cost of sequencing has dramatically dropped over the last two decades. It should be noted that NGS technologies/platforms are not expected to replace the traditional PCR-CE-based typing system, as much still needs to be accomplished, but will solve its limitations and deliver high-resolution genotyping in instances when standard testing has not been successful. Guidelines or recommendations to address technical, interpretative and reporting issues still need improvement. Moreover, other issues must be explained such as adapting the different national databases to include STR alleles detected by sequencing as well as their genetic variation. As there are many MPS kits already developed, as well as many in the pipeline, with numerous software packages to analyse and produce sequencing results, developing recommendations for similar issues will be complex.

1.12 Statistical analysis

Interpretation and reporting of DNA genotyping results is a complex procedure requiring professional judgment and expertise. Prior to comparing the DNA profile evidence recovered from a casework sample to a reference sample taken from suspects/victims, profiles must be initially evaluated in terms of amplification/instrumental artefacts and the number of contributors. Such determination is important for accurate interpretation and the choice of statistical method used to evaluate the evidential weight. Once the profile has been assessed, comparison between profiles results in one of three conclusions: inclusion, exclusion, and inconclusive [276].

The worst possible outcome of the comparison is 'inconclusive', which indicates that the profile evidence did not meet interpretation requirements, i.e., it did not give any decisive conclusion about the source of DNA in the samples. For example, an 'inconclusive' result may be obtained for a partial profile generated from low-quality or degraded DNA or due to an inability to deconvolute a mixture profile. When a DNA profile from a victim/suspect meets the quality requirement but remains inconsistent with the profile evidence, the individual is 'excluded' (or is a 'no match') as the potential donor of the evidence. If the victim/suspect DNA profile is consistent with the profile evidence, the individual is 'included' (or is a 'match') as a potential source of that evidence. To report the individual as included, DNA profiles must be statistically evaluated. An estimation of how often the DNA profile occurs in a population by chance (the DNA profile frequency), i.e. how rare a profile is among unrelated individuals in the general population, is significant in strengthening the power of the DNA evidence [276].

1.12.1 Random match probability

The random match probability (RMP) indicates the frequency of a DNA profile that would be expected to be found in the general population, calculated from the frequencies of the alleles in the profile within a group from that population [13]. It is calculated by multiplying the allele frequencies to give genotype frequencies, and then multiplying the genotype frequencies of all the loci included in a particular DNA profile, referred to in the forensic community as the product rule. The probability of seeing a genotype is based on how frequently the genotype occurs. For instance, the probability of seeing a genotype of a heterozygous locus is $2pq$, where p and q are the allele frequencies of the genotype, while the probability of seeing a genotype of a homozygote locus is p^2 , where p is the frequency of the homozygote allele.

Several assumptions are made when calculating the RMP. First, the alleles at each locus must be inherited independently, i.e. they must conform to the expectations of the Hardy–Weinberg equilibrium (HWE) which allows the use of allele frequencies to calculate expected genotype frequencies as p^2 and $2pq$. Second, each locus must be inherited independently, i.e. loci are in linkage equilibrium (LE) which allows multiplication of genotype frequencies and the use of the product rule. Third, the population data used to calculate genotype frequencies are not structured into subpopulations. In a structured population, a correction factor (θ) is used with the RMP calculation. A conservative $\theta = 0.01$ is recommended for most general populations and $\theta = 0.03$ for smaller and more reproductively isolated populations [277].

1.12.2 Likelihood ratio (LR)

The LR statistic assesses the relative probability of two competing hypotheses under a specific set of prior assumptions. In forensic genetics, it can be used to measure the evidential weight or the probative value of a match, when matching a suspect's profile with an evidence profile. LR shows how many times more likely it is that the match would be obtained if the evidence originated from the suspect than from a random individual from the population. Mathematically, the numerator is the first hypothesis which is the probability of finding the match given that the profile comes from the suspect (representing the prosecution position), and the denominator is the alternative hypothesis which is the probability of finding the match given that the profile comes from a random individual (representing the defence position) [13]. An LR value > 1 provides support for the first hypothesis (prosecution), whereas a value < 1 supports the alternative hypothesis (defence). Simply put, the LR is the inverse of the genotype frequency at a locus, i.e. $1/\text{RMP}$. For example, when the RMP is 0.005, the LR shows that the match would be 5000 times more likely if the individual is the contributor of a sample compared to a random individual being the contributor.

In paternity testing, the LR can be used to estimate the probability of a relationship by the most used statistical calculation: the typical paternity index (TPI). The TPI is calculated by a likelihood ratio of two probabilities of the same events under different hypotheses using specific information given in a paternity case. For example, the TPI is the likelihood ratio of the probability of the alleged man being the biological father compared to a random individual in the population. The larger the value is, the greater the chance that the alleged father is the real father compared to a random male. For example, if the combined paternity index (CPI) for all the loci in a profile is 100 (100 is the minimum accepted value for paternity inclusion), the alleged father is 100 times more likely to be the real father, with a probability of paternity of $> 99\%$, than a random individual from the population [26].

1.12.3 The counting method

Since Y-STRs are inherited paternally and treated as a haplotype, the product rule cannot be used to estimate the rarity of the Y-STR profile, as it can be used with autosomal STRs. Instead, the counting method is used to estimate the frequency of a profile by searching for the same haplotype in a large Y-STR reference database which contains Y-STR profiles of unrelated males from the relevant population [278]. In this case, profile frequency differs from match

probability as the former reflects the number of times the profile is seen, reported against the total number of profiles in the database. Therefore, the Y-STR profile has lower discriminatory power compared to autosomal STR profiles, and estimation of its discriminatory power is limited by the size of the population database.

Y-STRs have been argued as more useful for excluding suspects in crime scene investigations and for clearly indicating non-relatedness in paternity cases. However, inclusion (match) is more difficult due to several factors that could influence the discriminatory power of Y-STR markers, such as the existence of related individuals in the reference database due to non-random sampling, the number of Y-STRs and the genetic structure of the reference population [279]. In cases where there is evidence that an individual closely related to the known suspect is the donor of the evidence (inclusion), rapidly mutating (RM) Y-STR markers can be informative for differentiating between them. However, the current 13 RM Y-STRs have limitations in discriminating between very close relatives (brothers) [280], which creates a statistical challenge for reporting inclusion and exclusion. Several approaches have been proposed by forensic statisticians to evaluate the evidential weight of Y-STRs [281-284], but no agreement has been reached yet on the most suitable method [154].

The use of the counting method with a confidence interval (generally 95% or more) for correction of sample variation and database size has been suggested by SWGDAM for statistical interpretation and reporting of Y-STRs [278]. The larger the database, the less sampling variance there will be for the given profile probability.

1.13 The Kuwaiti population

The State of Kuwait is located on the Arabian Gulf (also called the Persian Gulf) in the northwest of the Asian continent and in the heart of the Middle East. Kuwait is bordered by the Kingdom of Saudi Arabia in the south, the Republic of Iraq in the north and west, and Iran in the east, across the Arabian Gulf (**Figure 1.26**). Kuwait's population is about 4.8 million, which includes 1.4 million Kuwaiti nationals and 3.4 million foreign nationals, according to the 2019 census [285].

Located in the Arabian Gulf, the nation state of Kuwait is a relatively small country. It has been ruled by the Al Sabah family for more than two centuries. Kuwait is about 14 times geographically smaller than the United Kingdom; however, this size comparison does not

adequately represent modern Kuwait's geostrategic weight, regionally or globally. Less than century ago, Kuwait was a poor country, by almost any measure, occupied by desert nomads. However, it is today one of the largest producers of oil, making it one of the richest countries in the world. Regrettably, oil wealth has made Kuwait the financial envy of much of the world, particularly its larger neighbours, Saudi Arabia, Iraq, and Iran. For this reason, Kuwait was invaded by Saddam Hussain's Iraqi army in 1990. A year later, a US-led military coalition liberated Kuwait.

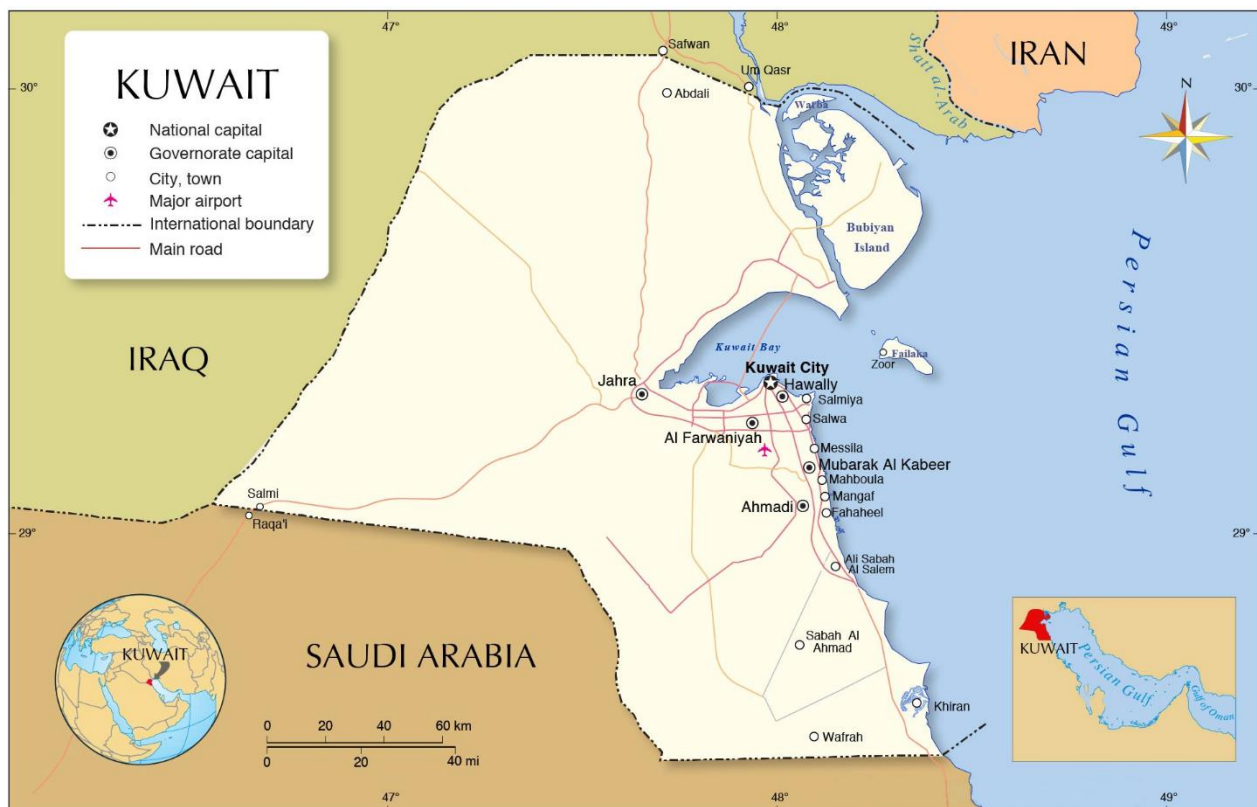


Figure 1.26. The map of Kuwait.

The map shows the State of Kuwait with its neighbours and international borders, major cities, and roads. Regionally, Kuwait is located in the northwest of the Persian Gulf and globally, in the middle of the Middle East (modified from [286]).

1.13.1 Kuwaiti history and population genetic structure

The present-day country of Kuwait was founded in about 1613 as a small village populated mainly by fishermen ruled by the Bani Khalid tribe [287]. At the start of the 17th century, tribal groups began to migrate from Najd, which is situated largely in Saudi Arabia at the centre the Arabian Peninsula and is populated by the Bedouin, the nomadic people of Arabic origin [287].

The migration was due to a lack of rainfall and extreme drought which made it impossible for the tribes and their herds to survive. When the situation became critical, the Bani Utub clans, which included the Al-Sabah family (the rulers of Kuwait), began their migration to Kuwait, and were joined by other families along their route [287]. The Utub families later took over the control of the settlement from the Bani Khalid and began to change tribal life; they soon developed a new settlement which lost its nomadic character over time.

As a result of the establishment of a new system of government in Kuwait, a large number of Bedouins, preferring city life to the desert, migrated to Kuwait under the new ruler whose authority was based on justice [287]. Other tribes seeking pasturage settled near the homeland of Kuwait and never travelled far. With these migrations from the desert, Kuwait grew in the 18th century and its population increased. During that time, trading networks were established with nearby countries such as Iran and Iraq, and several merchants moved to Kuwait and established businesses and settled. In addition, Iranian families and groups migrated from areas along the Iranian littoral to Kuwait, where they worked as merchants or labourers and eventually settled [288]. The modern genetic structure of the Kuwaiti population has been influenced significantly by these migrations.

Several studies have investigated the genetic structure of the Kuwaiti population using different genetic markers. By far the largest study was conducted by Alsmadi et al. who investigated the ancestry of 273 individuals using 572,366 SNP markers. Their results revealed three groups of ancestry: Persian, city-dwelling Saudi Arabians, and tent-dwelling Bedouin [289]. Other studies using mitochondrial DNA and Y-STR markers have shown that the Kuwaiti population is genetically close to the populations in neighbouring countries, especially Saudi Arabia and Iran, which is expected since gene flow between neighbouring countries is generally high [290-292].

In a rare study focusing on Bedouin tribes, Mohammed et al. investigated the genetic structure of six tribes residing in Kuwait using a combination of Y-chromosome and autosomal STR markers [293]. They found that these Kuwaiti tribes were clustered with neighbouring populations and were genetically similar. In addition, the Bedouin displayed strong evidence of genetic isolation and drift. This is expected with a high level of intertribal marriage between men and women who prefer unions with those with whom they share tribal, clan, or lineage affiliations [294]. Consequently, the prevalence of consanguineous marriage between those

with Bedouin tribal roots is high [295, 296]. Kuwaiti society in general practises consanguineous marriage, with a reported rate of 54.3% [297], and a higher rate is reported within and between tribal groups, which represent the majority of citizens in Kuwait [296].

The consequence of this practice is to increase homozygosity and reduce genetic variation, leading to population substructuring. In forensic science, the non-random mating effect in a population must be understood in order to estimate accurate match probabilities. Population substructure can be adjusted by using a correction factor (θ) as discussed above (in section **1.12.1**).

1.13.2 Forensic studies in Kuwait

Currently, forensic DNA analysis in Kuwait is carried out by the Kuwaiti Identification DNA Laboratory (KIDL) using only STR markers, including autosomal STRs, Y-STRs and X-STRs. Autosomal STRs are routinely used for both identification of individuals and paternity testing, whereas Y-STRs and X-STRs are used less frequently, and only in specific scenarios.

To date, few papers have been published investigating the forensic utility and genetic diversity of autosomal STR markers in the Kuwaiti population. In 2008, Alenizi and colleagues reported the allele frequencies of 15 STR loci included in the AmpFISTR® Identifiler® kit (Thermo Fisher Scientific, MA, USA) [298]. Based on these STRs, the F_{ST} distances between Kuwaiti nationals and foreign nationals from seven other populations residing in Kuwait were found to be consistent with their geographical distances [299]. Another recent study investigated the forensic utility of 25 autosomal STRs included in two separate kits: the PowerPlex® CS7 system and the PowerPlex® 21 system (Promega Corporation, WI, USA) [300]. Although the existing STRs are efficient for analysing cases of simple relationships, more STRs are increasingly required, particularly for complex paternity cases or to increase the discriminatory power in cases of partial DNA profiles and DNA mixtures.

1.14 Aims and objectives of this study

The overall aim of this thesis is to investigate the genetic diversity of different classes of forensic markers, using both traditional CE and the recently developed NGS techniques, in the Kuwaiti population. The outcome of this study makes a major contribution to research in forensic genetics in Kuwait by demonstrating the forensic utility of such systems, which is

necessary before implementing in the Kuwaiti Identification Laboratory (KIDL). The detailed aims and objectives for each chapter of this thesis are described below.

Chapter 3 aim: To evaluate the forensic utility of and generates genetic data for the indel markers included in the Investigator® DIPplex kit, which have not previously been reported for the Kuwaiti population.

Objectives:

- Carrying out CE-indel genotyping of 150 unrelated Kuwaiti individuals.
- Generating allele frequencies for the indel makers.
- Calculating population and forensic parameters to assess the forensic utility of the indels.
- Examining the utility of the indel loci to assign individuals to their origins at global and regional levels.
- Examining their ability to detect population differentiation by using genetic data from global populations.

Chapter 4 aim: To evaluate the forensic utility of 23 autosomal STR markers in the PowerPlex® Fusion 6C kit, which uses the PCR-CE method, for 400 samples from the Kuwaiti population. Four loci (D10S1248, D22S1045, D2S441 and SE33) will be reported for the first time in the Kuwaiti population. A second aim is to generate data comparable to those generated by NGS technology (reported in chapter 5) to examine the concordance using different methods/kits.

Objectives:

- Generating allele frequencies for the STRs to increase the amount of genetic data available for the Kuwaiti population.
- Calculating population and forensic parameters to assess the forensic utility of the STR loci.
- Examining the genetic relationship between the Kuwaiti population and other populations at a global and regional level.

Chapter 5 aim: to evaluate the forensic performance of 27 STR markers contained within the Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics NGS system, in the Kuwaiti population.

Objectives:

- Sequencing 376 unrelated Kuwaiti individuals using the ForenSeq™ DNA Signature Prep Kit (primer set A).
- Comparing the results to a CE-based STR genotyping kit, the Promega PowerPlex® Fusion 6C kit (reported in chapter 4) to examine concordance between the two systems.
- Generating allele frequencies, and assessing the forensic parameters for the sequenced STR makers.
- Exploiting and cataloguing the underlying genetic variation in the sequenced markers in the Kuwaiti population.
- Identifying and evaluating the performance of the SE33 locus, sequences, which were obtained from raw sequencing files using a bioinformatics approach.

Chapter 6 aim: to evaluate the forensic performance of the 94 iiSNP markers included in the ForenSeq™ DNA Signature Prep Kit in the Kuwaiti and to contribute to the population genetic data available for this under-represented region.

Objectives:

- Calculating the allele frequencies for, and assess the forensic parameters calculated using the sequenced iiSNP makers.
- Identifying the genetic variants within the entire sequence of 94 iiSNP-containing-amplicons and assess the level of diversity associated with these variants among the Kuwaiti population.
- Examining the influence of including these additional variants on the performance of forensic parameters.
- Examining the genetic distance between the Kuwaiti population and other populations, at the global and regional levels.

Chapter 7 aim: to evaluate the forensic performance of the 7 X-STRs contained within the Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics system, in the Kuwaiti population. The X-STR markers, along with their sequence variation, will be reported for the first time in the Kuwaiti population. A second aim is examining the

concordance of the Y-STR markers collected using different technologies (MPS and PCR-CE), and to predict the haplogroups from the Y-STR haplotypes.

Objectives:

- Generating allele frequency data for the X-STR makers and their haplotypes to increase the amount of genetic data available for the Kuwaiti population.
- Cataloguing the underlying genetic variation in the sequenced X-STR markers in the Kuwaiti population.
- Generating Y-STR profiles using the Promega PowerPlex® Y23 System for 248 Kuwaiti males.
- Comparing the Y-STR profiles generated by CE-based genotyping (Promega PowerPlex® Y23) with those generated by sequencing (ForenSeq™ DNA Signature Prep Kit), to examine concordance between the two systems.
- Identifying the main Y-chromosome haplogroups in the Kuwaiti population, using the Y-STR haplotypes.

Chapter 2 Materials and methods

2.1 Chapter overview

This chapter details the methodology involved in generating DNA profiles using different marker types, starting from sample collection through experimental methods to statistical analysis. Prior to carrying out the lab work, personal protective equipment (PPE) including safety glasses, laboratory coat and face mask were put on. All benches were cleaned using bleach before and after use, to remove debris and contaminations. In addition, where possible, equipment used were wiped using alcohol (75% ethanol) then subjected to Ultraviolet light using a CL-508 CROSSLINKER® (UVItect Ltd, Cambridge, UK) to eliminate any exogenous DNA and to prevent cross-contamination.

2.2 Ethical approval, sample collection and preparation

Ethical approval for all parts of the project was acquired from the Department of Pure and Applied Chemistry Departmental Ethics Committee (DEC). Sample collection was carried out in the Kuwait Forensic DNA Centre, which is part of the General Department of Criminal Evidence. Volunteers were recruited to donate blood sample by emails which included the basic objectives of the research, and potential participants were asked to respond to the email if they were willing to take part. Volunteers were also asked verbally if they were willing to participate in the project.

A room designated for blood donation in the Kuwait Forensic DNA Centre was used to collect samples. Prior to donation, the nature of the research and the procedure of taking the blood sample were fully explained to participants verbally. They were asked to read an information sheet to ensure they understood what was involved and to sign a consent form to agree to participate in the research.

Three to five blood drops were taken from each volunteer by an experienced investigator by making a finger prick using a Bullseye Safety Lancet™ (Phoenix Healthcare Solutions, LLC, USA). Alcohol wipes were used to clean the finger before and after a puncture to reduce any infection risk, and an adhesive dressing was given afterwards. Blood was dropped onto blood-collection cards within the printed circular area and left at room temperature for three hours to dry.

Sample preparation was carried out by punching discs out of the blood spot using a Harris Micro-Punch® (GE Healthcare Life Sciences, UK). This was done on an alcohol-cleaned cutting mat. The cutting edge of the puncher and the mat were cleaned using 75% alcohol after each use. Discs were carried on for downstream processing.

2.3 DNA extraction

For parts of the project where DNA extraction from blood samples was carried out, it was performed using the AutoMate Express™ DNA Extraction system using the PrepFiler Express™ Forensic DNA Extraction Kit (Thermo Fisher Scientific, WA, USA). The chemistry behind the PrepFiler extraction system is based on using polymer-embedded (coated) magnetic particles to capture DNA molecules to form a magnetic particle/DNA compound. Inhibitors and other biological remnants are eliminated through several automated washing steps, followed by the elution stage, where DNA is released from the magnetic particles.

The extraction procedure was done following the manufacturer's protocol [301]. The procedure starts with punching two 2 mm discs out of the blood spot followed by a cell lysis stage in which samples were incubated in a Thermomixer comfort (Eppendorf, HH, Germany) for 30 min with 500 µL lysis buffer and 5 µL of freshly prepared 1 M dithiothreitol (DTT), incubated at 70 °C and shaken at 900 rpm. The lysate was transferred directly to the automated extraction run. After several automated washing steps, DNA was eluted in 50 µL of elution buffer and extracted DNA samples were stored at 4 °C.

2.4 DNA quantification

DNA concentration was measured using the Promega PowerQuant™ System (Madison, WI, USA) performed on a 7500 Real-Time PCR instrument (Thermo Fisher Scientific). PowerQuant™ is a quantitative PCR (qPCR) multiplex containing five dye probes to detect the amount of human autosomal DNA, human male DNA (Y chromosome), level of DNA degradation and the presence of inhibitors. The procedure was performed according to the manufacturer's protocol [302]. The degradation is detected using Quasar® 670 dye-labelled probe. The probe detects the existence of 294 bp amplicon from a different region of a locus. The large degradation amplicon is more vulnerable to degradation and the presence of inhibitors. The TMR dye-labelled probe is used to detect the Internal PCR control (IPC). The

IPC is amplified fully with the absence of inhibitors. However, its amplification will be reduced with the presence of inhibitors.

The procedure starts with preparing serial dilutions of known human male DNA standards (50 ng/ μ L) resulting in four DNA standards with concentrations of 50, 2, 0.08 and 0.0032 ng/ μ L. Quantification of human DNA was performed in a total reaction volume of 20 μ L in optical 96-well plates (Thermo Fisher Scientific). Each reaction was set up with 7 μ L of water, 10 μ L master mix and 1 μ L primer mix, with 2 μ L of DNA sample. The thermal cycling programme was set up in the HID Real-Time PCR Analysis Software (Thermo Fisher Scientific, USA) using the following conditions: 98 °C for 1 min followed by 39 cycles of 98 °C for 15 sec and 62 °C for 35 sec. Each qPCR run was performed with 51 extracted samples, 8 standards (duplicates of each of the four serial dilutions) and a negative control (duplicated). Results were exported using the Promega PowerQuant™ Analysis Tool software.

The amount of DNA in unknown samples was determined by comparing their C_q value (quantification cycle in which the fluorescence signal was detected) to the standard curve C_q values generated from the known DNA standards. Quantification efficiency (PCR efficiency) was assessed by examining the linearity of the standard curve (slope), which is recommended to be in the range of -3.1 to -3.58, which corresponds to 90% to 110% amplification efficiency, with 100% efficiency achieved at a slope of -3.33. The performance of the primers was assessed using the regression coefficient (R^2) of the standard curve, which is recommended to be >0.99.

2.5 Indel marker genotyping using the Investigator® DIPplex kit

The Investigator® DIPplex kit (Qiagen, Hilden, Germany) was used to amplify 30 insertion/deletion polymorphisms as well as the amelogenin sex marker. Primers in the kit are labelled with five fluorescent dyes named 6-FAM™ (blue), BTG (green), BTY (yellow), BTR (red) and BTO (orange), which therefore subsequently label the resulting PCR product.

2.5.1 Amplification

PCR reactions were set up with a final reaction volume of 25 μ L by combining 5 μ L each of reaction mix and primer mix, 0.6 μ L *Taq* DNA polymerase and 14.4 μ L of quantified template DNA diluted in nuclease-free water to give a final input of 0.5 ng. The thermal cycling programme was set up in the GeneAmp™ PCR System 9700 (Thermo Fisher Scientific) as

follows: initial hot-start step at 95 °C for 4 min to activate the polymerase enzyme, followed by 30 cycles each of denaturing at 94 °C for 30 sec, annealing at 61 °C for 2 min and extension at 72 °C for 75 sec. This was followed by a final extension for 60 min at 68 °C then a final hold at 10 °C. Once the PCR was completed, samples proceeded directly to electrophoresis.

2.5.2 Fragment analysis

Detection and separation of the DNA fragments were carried out using an Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific). A loading mixture of 0.5 µL DNA Size Standard 550 (BTO) and 12 µL Hi-Di™ formamide was prepared for the required number of samples. An amount of 12 µL of the mixture was added to the reaction wells of a MicroAmp optical 96-well plate (Thermo Fisher Scientific). To this 1 µL of each sample or the allelic ladder was added in the appropriate well. Samples were denatured for 3 min at 95 °C and then placed on ice (snap-frozen) for 3 min. The electrophoresis was run under the following conditions: 3 kV injection voltage, 8 sec injection time, 13 kV run voltage, 15 min run time and BT5 dye set.

2.6 Autosomal STR genotyping using the PowerPlex® Fusion 6C kit

The PowerPlex® Fusion 6C kit (Promega Corporation) was utilised to amplify 27 STR loci and amelogenin. Six fluorescent dyes are used to label the primers namely: FL (blue), JOE (green), TMR (yellow) and CXR (red), TOM (purple) and WEN (orange). The procedure used followed the manufacturer's protocol [303].

2.6.1 Amplification

Amplification of human DNA was carried out directly from 1.2 mm blood card punches in a final volume of 12.5 µL combining 5 µL nuclease-free water and 2.5 µL each of master mix, primer mix and AmpSolution. The PCR was performed in a SureCycler 8800 thermal cycler (Agilent Technologies, Santa Clara, USA) using the following settings: initial hot-start step at 96 °C for 4 min to activate the polymerase enzyme, followed by 25 cycles of denaturing at 96 °C for 5 sec, annealing at 60 °C for 1 min and extension at 60 °C for 10 min, and a final hold at 4 °C.

2.6.2 Fragment analysis

DNA fragments analysis was performed in the Applied Biosystems 3500 Genetic Analyzer. A loading mix of 0.5 µL WEN ILS 500 size standard and 9.5 µL Hi-Di™ formamide was prepared

for the required number of samples. An amount of 10 µL of the mix was added into the reaction wells of a MicroAmp optical 96-well plate, followed by the addition of 1 µL of each PCR product or allelic ladder into the appropriate well. Samples were denatured at 95 °C for 3 min and placed on ice for 3 min. The electrophoresis was run under the following conditions: 1.2 kV injection voltage, 15 sec injection time, 13 kV run voltage, 25 min run time and J6 dye set.

2.7 Y-chromosomal STR genotyping using the PowerPlex® Y23 System

The PowerPlex® Fusion 6C kit (Promega Corporation) was utilised to amplify 23 Y-STR loci. Six fluorescent dyes are used to label the primers namely: FL (blue), JOE (green), TMR (yellow) and CXR (red), TOM (purple) and WEN (orange). The procedure used followed the manufacturer's protocol [304].

2.7.1 Amplification

Amplification of male DNA was carried out directly from 1.2 mm blood card punches in a final volume of 12.5 µL combining 6.25 µL nuclease-free water and 2.5 µL each of master mix and AmpSolution and 1.25 µL primer mix. The PCR was performed in a SureCycler 8800 thermal cycler (Agilent Technologies, Santa Clara, USA) using the following settings: initial hot-start step at 96 °C for 2 min to activate the polymerase enzyme, followed by 25 cycles of denaturing at 94 °C for 10 sec, annealing at 61 °C for 1 min and extension at 72 °C for 30 sec. This was followed by a final extension for 20 min at 60 °C then a final hold at 4 °C.

2.7.2 Fragment analysis

DNA fragments analysis was performed in the Applied Biosystems 3500 Genetic Analyzer. A loading mix of 0.5 µL WEN ILS 500 Y23 size standard and 9.5 µL Hi-Di™ formamide was prepared for the required number of samples. An amount of 10 µL of the mix was added into the reaction wells of a MicroAmp optical 96-well plate, followed by the addition of 1 µL of each PCR product or allelic ladder into the appropriate well. Samples were denatured at 95 °C for 3 min and placed on ice for 3 min. The electrophoresis was run under the following conditions: 1.2 kV injection voltage, 15 sec injection time, 15 kV run voltage, 1,210 sec run time and G5 dye set.

2.8 DNA sequencing using the MiSeq FGx® with the ForenSeq™ DNA Signature Prep kit

Massively parallel sequencing was carried out using the MiSeq FGx® instrument with primer mix (A) of the ForenSeq™ DNA Signature Prep Kit (Verogen, CA, USA), which contains 153 markers as follows: 27 autosomal STRs, 24 Y-STRs, 7 X-STRs, 94 SNPs, and amelogenin. The system can sequence up to 96 samples (libraries) in a single run. A 1.2 mm punch from FTA cards was used directly for PCR after washing in 100 µl of freshly prepared 1X TBE (Tris/Borate/EDTA) buffer (Thermo Fisher Scientific). The whole process was performed following the manufacturer's protocols [305].

2.8.1 Template and library preparation using two-step PCR

The first PCR was performed to amplify and tag STR and SNP sequences. A master mix was prepared containing 4.7 µL PCR1 Reaction Mix, 0.3 µL Enzyme Mix, 5 µL DNA Primer Mix A (DPMA) and 5 µL nuclease-free water, for the required number of samples. An amount of 10 µL of master mix was transferred to each well of a 96-well plate. Ninety-four of the wells contained blood discs (samples); one contained diluted positive control 2800M DNA (supplied with the kit) and one nuclease-free water as a negative control. PCR1 was performed using a Veriti thermal cycler (Thermo Fisher Scientific) with the following conditions: hot-start at 96 °C for 3 min; 8 cycles of 96 °C for 45 sec, 80 °C for 30 sec, 54 °C for 2 min and 68 °C for 2 min; then 10 cycles of 96 °C for 30 sec and 68 °C for 3 min; with a final hold at 10 °C.

The second PCR was performed to attach sample-specific indexed adapters (barcodes) to the amplified DNA. Barcodes were added by first transferring 12 unique Index 1 (i7) adapters into each column (1-12) of a 96-well plate and then eight unique Index 2 (i5) adapters to each row (A-H); followed by the addition of 27 µL of PCR2 Reaction Mix into each well. The thermal cycling conditions were as follows: hot--start at 98 °C for 30 sec; 15 cycles of 98 °C for 20 sec, 66 °C for 30 sec and 68 °C for 90 sec; followed by a final hold stage at 10 °C.

2.8.2 Library purification and normalisation

The purification step began with transferring 45 µL of the amplified indexed libraries into 45 µL of Sample Purification Beads (SPB) in a 96-well midi plate. The plate was sealed then shaken for 2 min at 1800 rpm; followed by a 5 min incubation at room temperature. A washing step was carried out by adding 200 µL of freshly prepared 80% EtOH to each sample while the plate

was sitting on a magnetic stand. The supernatant containing inhibitors and other unwanted biological components was removed, and the washing step repeated. Prior to the normalisation step, DNA molecules were released from the magnetic beads by adding 52.5 μ L Resuspension Buffer (RSB) to each well. The plate was shaken for 2 min at 1800 rpm then incubated at room temperature for 2 min. Finally, the plate was rested on the magnetic stand and 50 μ L of the supernatant containing the purified libraries was transferred to a new plate, ready for normalisation.

The purified libraries were normalised by taking a fixed quantity of magnetic beads, which capture the same amount of DNA from each library. Firstly, a master mix of 46.8 μ L Library Normalisation Additives (LNA1) and 8.5 μ L Library Normalisation Beads 1 (LNB1) were prepared for the required number of samples. A total of 45 μ L of the LNA1/LNB1 mix was transferred into a new midi plate, followed by 20 μ L of the purified libraries in the corresponding well. The plate was sealed and shaken for 30 min at 1800 rpm. Two washes were performed, each with 45 μ L Library Normalisation Wash 1 (LNB1). Finally, 32 μ L of freshly prepared 0.1 N HP3 (2N-NaOH) was transferred to each well and the plate shaken for 5 min at 1800 rpm. After placing the plate on a magnetic stand, a volume of 30 μ L of the supernatant was transferred to a new plate, which contained 30 μ L Library Normalization Storage Buffer 2 (LNS2) in each well.

2.8.3 Library denaturing and sequencing

Once normalisation was completed, 5 μ L of each library was transferred to a single tube, then 7 μ L of the pooled libraries was diluted in 591 μ L Hybridization Buffer (HT1). This was followed by the addition of 2 μ L Human Sequencing Control (HSC) into the pooled-libraries tube. HSC is a positive sequencing control for the FGx[®] instrument containing a pool of 23 STRs. Prior to loading its contents into the reagent cartridge, the tube containing the pooled libraries was denatured for 2 minutes at 96 °C and then placed on ice for 5 min, before 600 μ L was loaded into the designated well of the cartridge. Before loading the flow cell into the FGx[®] instrument, it was rinsed with nuclease-free water and thoroughly dried with lint-free lens-cleaning tissue. The required reagents were loaded into the instrument according to the step-by-step instructions in the MiSeq FGx[®] software. The run was initiated using the manufacturer's recommended sequencing parameters.

2.9 CE data analysis

Data generated using capillary electrophoresis were analysed using GeneMapper ID-X software version 1.4 (Thermo Fisher Scientific). The result of the analysis is an electropherogram, which plots the sizes of the amplified DNA fragments estimated by comparison to the Internal Size Standard. Alleles are determined by matching their position relative to the alleles in the allelic ladder (which contains all the possible alleles for each locus) using bins that are specified in panels. Bins and panels were downloaded from the manufacturer's website and set up in the software as recommended by the manufacturer.

2.10 Sequencing data analysis

2.10.1 Universal Analysis Software (UAS)

The sequence data generated using the ForenSeq™ DNA Signature Prep Kit on the MiSeq FGx® instrument were analysed using the ForenSeq™ Universal Analysis Software (UAS) version 1.3. Several analysis procedures were performed during the process of parsing the data through the UAS including: demultiplexing (assembling sequenced reads on the basis of their barcodes), aligning sequences with the reference genome (GRCh38) and allele calling. The UAS generates a detailed genotype report for each sample in a Microsoft Excel workbook file (.xlsx) with allele/genotype calls, the number of reads, nucleotide sequence and the quality control indicators for each marker. The UAS also produces a flanking-region report, which contains the sequences between the primers for each amplicon in the kit.

2.10.2 STRait Razor

STrait Razor version 3.0 [306] is a bioinformatics suite that can identify STR sequences and assign the appropriate nomenclature according to ISFG recommendations [266], characterise indel and SNP variants in the STR flanking region, and identify novel alleles based on its database, which was published by Novroski et al. [267]. The analysis suite is composed of Perl scripts and an Excel workbook. The Perl scripts align the FASTQ files with the reference genome (GRCh38) to identify allele sequences according to the predefined 5' and 3' anchor sequences of each locus. The Excel workbook visualises the resulting data providing allele calls, number of reads and nomenclature for both length and sequence-based alleles. Sequences that are not present in the software database are highlighted with 'NovelSeq'.

2.10.3 Manual bioinformatic analysis

In some cases, manual bioinformatic analysis was required for result confirmation. This was carried out using the Burrows-Wheeler aligner (BWA) version 0.7.15-r1140 to align the sample FASTQ reads to the human genome reference GRCh38 [307]. The resulting bam file was indexed using SAMtools version 1.3.1. Indexing facilitates a fast visualisation of the aligned reads, which is further carried out using Integrative Genomics Viewer (IVG) version 2.8.0 [308].

2.11 Statistical analysis

2.11.1 Genetic and forensic indices

Arlequin software version 3.5 was used to calculate genetic parameters such as allele frequencies, observed and expected heterozygosity, and to carry out the Hardy-Weinberg Equilibrium (HWE) and linkage disequilibrium tests [309]. Forensic parameters such as the random match probability (RMP), discrimination power (DP), power of exclusion (PE), typical paternity index (TPI) and the polymorphic information content (PIC) were calculated using STRAF, an online tool for STR data analysis [310].

2.11.1.1 Random match probability (RMP)

The RMP is the probability that two unrelated individuals randomly selected from the studied population will possess identical genotypes [311]. This is calculated for each locus, as illustrated in **Equation 2.1** [55]. The combined random match probability (CMP) for all markers is calculated by multiplying individual RMPs for all loci across a profile.

Equation 2.1. Random match probability (RPM)

$$RPM = \sum_{a=1}^n P_k^2$$

Where ***n*** is the number of observed genotypes, and ***P_k*** is the frequency of each genotype found in the population.

2.11.1.2 Power of discrimination (PD)

The PD is the probability that two unrelated individuals selected from the studied population will exhibit different genotypes. This is calculated for a single locus using 1-RPM [312]. The combined power of discrimination (CPD) for all markers is calculated by multiplying the discrimination power across all loci or simply by taking 1-CMP.

2.11.1.3 Power of exclusion (PE)

The PE is the efficiency of a locus to exclude a random individual in the population as being the alleged father in paternity cases. The PE calculation for a single locus is described in **Equation 2.2** [313]. PE values for several markers are combined (CPE) to estimate the overall exclusion power (**Equation 2.3**).

Equation 2.2. Power of exclusion (PE)

$$PE = H^2(1 - 2Hh^2)$$

Where **H** is heterozygosity and **h** is homozygosity at the locus.

Equation 2.3. Combined power of exclusion (CPE)

$$CPE = 1 - \prod_{a=1}^n (1 - PE_a)$$

Where **n** is the number of the loci, **PE_a** is the power of exclusion for the **a**th locus, and the **Π** is the sign for multiplication.

2.11.1.4 Typical paternity index (TPI)

The TPI reflects the strength of a locus to be used in paternity testing. TPI is calculated by a likelihood ratio of two probabilities of the same events under different hypotheses, using specific information given in a paternity case. The TPI for each locus is calculated as shown in **Equation 2.4**. The combined paternity index (CPI) is the multiplication of all TPI values for each locus.

Equation 2.4. Combined paternity index (CPI)

$$TPI = \frac{1}{2h}$$

Where **h** is homozygosity at the locus.

2.11.1.5 Polymorphism information content (PIC)

The PIC is used to estimate the polymorphism level for each marker. It is calculated as shown in **Equation 2.5** [314].

Equation 2.5. Polymorphism information content (PIC)

$$PIC = 1 - \sum_{a=1}^n P_a^2 - \sum_{a=1}^{n-1} \sum_{j=a+1}^n 2P_a^2 P_b^2$$

Where n is the number of alleles, P_a is the frequency of the a th locus and P_b is the frequency of b th locus.

2.11.2 Population analysis

STRUCTURE software version 2.3.4 was used to study population structure. STRUCTURE uses a Bayesian-based clustering approach that uses multi-locus genotypes data to infer the population structure and assign individuals to their populations based on their genotypes [315]. This is achieved by taking a range of numbers of populations (K) in order to calculate the proportion of the genome of each individual in the sample originating from each inferred population [315]. The software calculates the likelihood of the data (X) for a range of K values, and the true number of K is determined by the maximal value of $\ln P(X|K)$. The maximal value of the rate of change (ΔK) in the $\ln P(X|K)$ between successive K values should also be considered in order to accurately infer the true number of genetic clusters in the data [316].

2.11.3 Principal component analysis (PCA)

Principal component analysis (PCA) is an unsupervised clustering method used to uncover population structures and to correct for stratification in genome-wide association studies without losing any variation in the data [317]. This method discovers hidden relationships between samples by removing variables (genetic markers) with no variation and keeping those with maximum variation. In the PCA plot, similar samples will be localized near to each other and dissimilar samples will be far apart [318].

Chapter 3 Evaluation of 30 forensic indel markers in the Kuwaiti population

3.1 Introduction

Short tandem repeat (STR) profiling systems are used in almost every forensic laboratory and are considered the gold standard for human identification. Their diversity and high discriminatory power allow them to be used not only for human identification purposes but also in studies of population genetics and in paternity cases. Although they are robust, there are several limitations to STR markers, including large amplicon size, high mutation rates, and artefacts such as stutter peaks. Single nucleotide polymorphism (SNP) typing is an alternative approach to address these shortcomings of STRs. One of the advantages of SNP markers is their low mutation rate, which has been found to be lower than that of STR polymorphisms [63, 64]. Given that SNPs are single base changes, they can be multiplexed using much shorter amplicons than STR amplicons [66], overcoming the major limitation of STRs, which arises when typing degraded DNA samples [67]. Finally, by generating profiles without stutter peaks, SNP markers avoid one of the interpretational complexities of STR typing. However, SNP typing requires complex techniques, such as minisequencing or Sanger sequencing, which are not standard procedures in forensic genetics laboratories.

In contrast, indel markers, relatively recently developed for use as genetic markers, combine many of the benefits of SNP markers but use simple STR typing techniques, making their implementation feasible in forensic genetics laboratories. The advantages of indel markers include their abundance throughout the human genome, low mutation rate, ability to be multiplexed in large numbers with short amplicons, lack of production of stutter peaks, and compatibility with platforms found in any standard forensic genetics laboratory [112-115].

The Investigator® DIPplex (Qiagen) kit is available to genotype 30 indel markers and the amelogenin sex chromosome marker in a single PCR reaction. The markers are distributed across 19 chromosomes with amplicon sizes ranging from 75 bp to 160 bp, and length variation of 4-22 bp between the two alleles (del, ins) at each locus. To date, no study has investigated the utility of the DIPplex markers in the Kuwaiti population. Therefore, this chapter will focus on assessing the population genetics of indel markers in the Kuwaiti population prior to their implementation in the Kuwait Identification DNA Laboratory (KIDL).

3.2 Aims and objectives

The main aim of this study is to evaluate the forensic utility of and generate genetic data for the indel markers included in the Investigator® DIPplex kit, which have not previously been reported for the Kuwaiti population and to increase the amount of population genetic data available for this under-represented region. This study involves:

- Generating DNA profiles for 150 blood samples collected from the Kuwaiti population using the DIPplex kit.
- Calculating allele frequencies and evaluate population genetic and forensic statistics for the 30 indel markers in the Kuwaiti population.
- Examining the utility of the indel markers for assigning individuals to their origins at global and regional levels, using genotypic data available in the literature.
- Determining the utility of these markers for population differentiation using allele frequencies from 27 global populations.

3.3 Materials and methods

Sample collection and DNA profiling were carried out at the Kuwait Identification DNA Laboratory (KIDL), which is part of the General Department of Criminal Evidence. Statistical analyses were carried out at the University of Strathclyde.

3.3.1 Sample collection and preparation

Blood samples were collected from 150 unrelated Kuwaiti males and females and deposited onto Fitzco 705 Classic cards (Fitzco Inc., USA). The number of individuals sampled was chosen according to the recommendations of Chakraborty who suggested that 100–150 individuals would provide adequate data to study the genetics of forensic markers [319]. Ethical approval and sample collection procedures are described in section 2.2. The participant information sheet and consent form for this study are shown in **Appendix 1 (A1.1)**. Samples were divided into three sets, each containing fifty samples. Each set was prepared and analysed in the laboratory independently to reduce the risk of contamination, minimise large samples and mishandling, and increase efficiency.

3.3.2 DNA extraction and quantification

DNA was extracted from blood samples using the AutoMate Express™ DNA Extraction system using the PrepFiler Express™ Forensic DNA Extraction Kit (Thermo Fisher). The extraction

procedure is described in section **2.3**. DNA concentration was measured using the PowerQuant™ System (Promega) on a 7500 Real-Time PCR instrument (Thermo Fisher Scientific) as described in section **2.4**. Autosomal DNA concentration was used to estimate the quantity required for PCR, whereas the existence of male (Y-chromosome) DNA was used only for sex verification.

3.3.3 Indel amplification and detection

The Investigator® DIPplex kit (Qiagen, Hilden, Germany) was used to amplify the DNA extracted from samples. The multiplex amplifies 30 insertion/deletion polymorphisms and amelogenin. Three PCR assays were carried out, as described in section **2.5.1**. For each, the PCR master mix was prepared for 53 samples: 51 test samples, one PCR positive control, and one PCR negative control. Samples with DNA concentrations of > 0.5 ng/μL (more than the optimal amount) were diluted to bring them into the recommended range. For the control DNA 9948 sample, 1 μL of a concentration of 0.5 ng/μL was used.

DNA fragment analysis was carried out using the Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific, USA) with the DIPplex allelic ladder and DNA size standard 550 BTO. New POP-4™ polymer, anode/cathode buffers and a 36 cm capillary array were installed. Before electrophoresis, spectral calibration was carried out according to the DIPplex manual, using the matrix standard BT5. The protocol is described in detail in section **2.5.2**.

3.3.4 Data analysis

3.3.4.1 Population and forensic statistical analysis

The results were analysed using GeneMapper® ID-X version 1.4. Template files required for analysis, including bins, panels, size standard, analysis method, plot, and table settings, were downloaded from the Qiagen website. Genotypes were sorted using Qiagen DIPSorter software. Arlequin version 3.5 was used to calculate allele frequencies and observed and expected genotype heterozygosities and to test for Hardy-Weinberg equilibrium (HWE) [309]. Genepop version 4.7 software was used to test for linkage disequilibrium [320]. Forensic indices were computed using Promega PowerStats version 1.2.

3.3.4.2 Inferring individuals' origin

Snipper App suite version 2.5 a web-based statistical analysis tool [321], was used to assess the utility of the 30 indels for assigning individuals to their population of origin, based on a

Bayesian algorithm. The software calculates allele frequencies, assuming HWE, from the genotypes of each population input, then assigns individuals to their origins using the maximum likelihood estimate for membership of different populations [322].

The analysis used 2079 DIPplex profiles comprising 150 Kuwaitis (this study), 80 Danes [124], 119 Chinese (She) [323], 565 Chinese (Han) [323], 373 South Koreans [141], 500 Emiratis [324], and 292 Bahrainis [325] to assess whether the DIPplex markers could be used to assign individuals to their population of origin. This was done using the 'cross-validation analysis of population data' in the Snipper. Principal component analysis (PCA) was performed using the 'classification of multiple profiles with a custom Excel file of populations' option in Snipper.

3.3.4.3 Population structure analysis

STRUCTURE analysis (described in section **2.11.2**) was carried out on the dataset described above, using the analysis option without population information to check whether the results approximately agreed with the separation of samples into their populations. The predefined populations were only included as a population label rather than information for the analysis. The parameters for the analysis were set as follows: 'admixture' and 'correlated allele frequencies' models using 100,000 Markov chain Monte Carlo (MCMC) steps for each run, with the first 10,000 discarded as a burn-in, and the inferred number of K set from 1 to 10. At each K , the analysis was repeated five times in order to test the results for consistency. The results were visualised using CLUMPAK (Clustering Markov Packager Across K) [326, 327], and the best K was determined using Structure Harvester software [328, 329].

3.3.4.4 Population differentiation

To assess how genetically distant the Kuwaiti population is from other global populations, PCA was conducted based on previously published allele frequencies for 27 different populations: Bahrain ($N = 292$), United Arab Emirates ($N = 500$) [324], Iraq ($N = 101$) [137], Turkey ($N = 93$) [137], Lithuania ($N = 110$) [137], Slovenia ($N = 157$) [137], Iran ($N = 253$) [330], Spain ($N = 71$) [142], Poland ($N = 122$) [133], Italy ($N = 200$) [125], Finland ($N = 151$) [128], Greece ($N = 177$) [331], Portugal ($N = 108$) [127], Denmark ($N = 117$), Japan ($N = 251$) [140], Czech Republic ($N = 55$) [146], Somalia ($N = 175$) [128], Taiwan ($N = 126$) [133], Uruguay ($N = 131$) [332], Korea ($N = 373$) [141], Brazil ($N = 519$) [130], Germany ($N = 80$) [333], Han Chinese ($N = 565$) [323], US Caucasian ($N = 262$) [143], US Asian ($N = 50$) [143], US Hispanic ($N = 140$) [143], and US African ($N = 260$) [143]. This resulted in a final data set of 5,589

individuals typed for the 30 DIPplex loci. The PCA plot was visualised as a three-dimensional plot using the plot3D package in R.

3.4 Results

3.4.1 Extraction and quantification

The AutoMate Express™ DNA Extraction System using PrepFiler Express™ produced extracts with good concentrations of DNA. In addition, the PowerQuant™ System was successfully used to quantify autosomal DNA concentrations in all the samples and the Y-DNA concentrations in male samples only.

Quantification efficiency was assessed by examining the linearity of the standard curve, which is recommended to be in the range of -3.1 to -3.58 , corresponding to PCR amplification efficiencies of 90% to 110%, with 100% efficiency achieved at a slope of -3.33 . The three quantification assays for the three sets of samples were in the recommended range of slope linearity and fitted as standard curves with R^2 values of > 0.99 . The quantification parameters generated by PowerQuant™ for each run are shown in **Appendix 2 (A2.1)**.

The DNA concentrations varied between the Kuwaiti samples; 91 samples yielded 0.1-0.5 ng/μL of autosomal DNA while the remaining 59 contained more than 0.5 ng/μL. The latter samples were diluted to the appropriate concentration to permit the addition of the recommended input of 0.5 ng to the amplification reaction. Concentrations of autosomal and Y chromosome DNA obtained from the samples are shown in **Appendix 2 (A2.2 and A2.3, respectively)**.

3.4.2 Indel genotyping

The DIPplex kit uses primers designed to amplify each indel marker and labelled with one of four fluorescent dyes that are detected by CE. Samples were analysed using the Applied Biosystems 3500 Genetic Analyzer, which separates amplicons according to their size and arranges them based on the dye colour emitted. Peak size is determined by a calibration curve, which is produced from internal size standard fragments of known size; the internal size standard used here was BTO, which contains 13 fragments (**Figure 3.1**). Alleles/genotypes are determined by comparing peaks with an allelic ladder, which contains all the possible alleles for each locus, as shown in **Figure 3.2**. No sign of contamination was detected in any

of the negative controls. The positive control DNA 9948 profile (**Figure 3.3**) was concordant with the positive control profile given in the manufacturer's protocol.

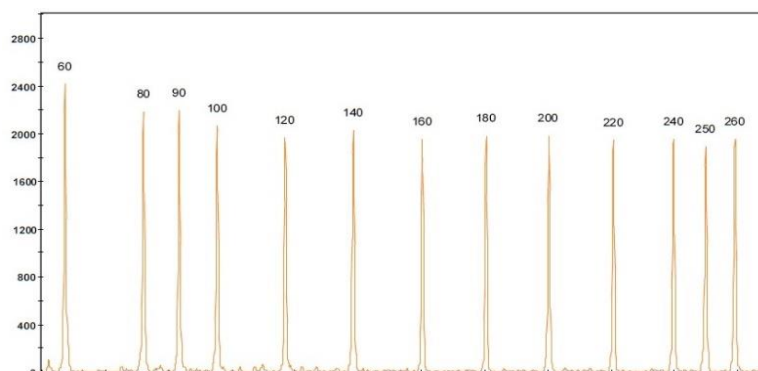


Figure 3.1. Electropherogram of the internal size standard (BTO).
Thirteen fragments ranging from 60 bp to 260 bp are separated according to their length.

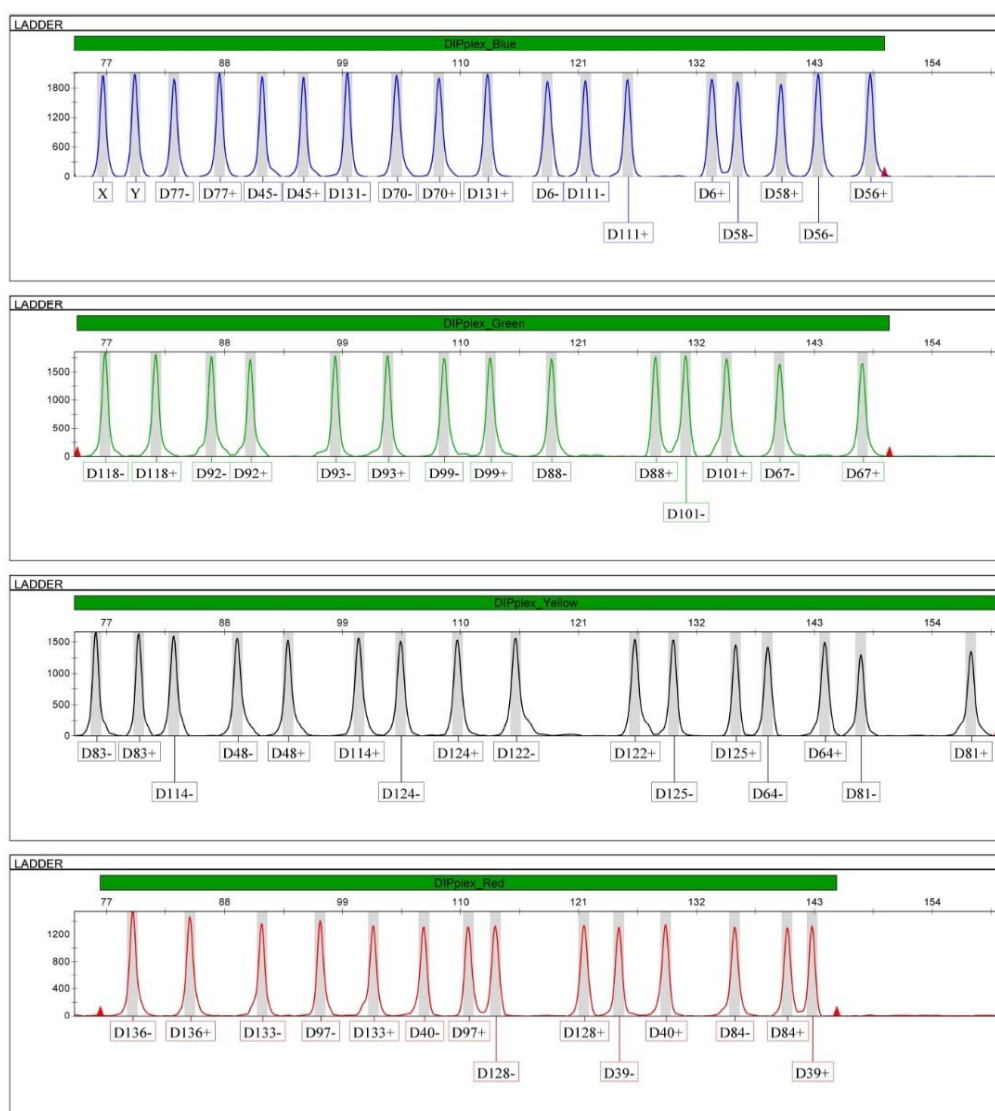


Figure 3.2. Electropherogram of the DIPplex allelic ladder.
The allelic ladder contains all possible alleles for each locus of the DIPplex kit

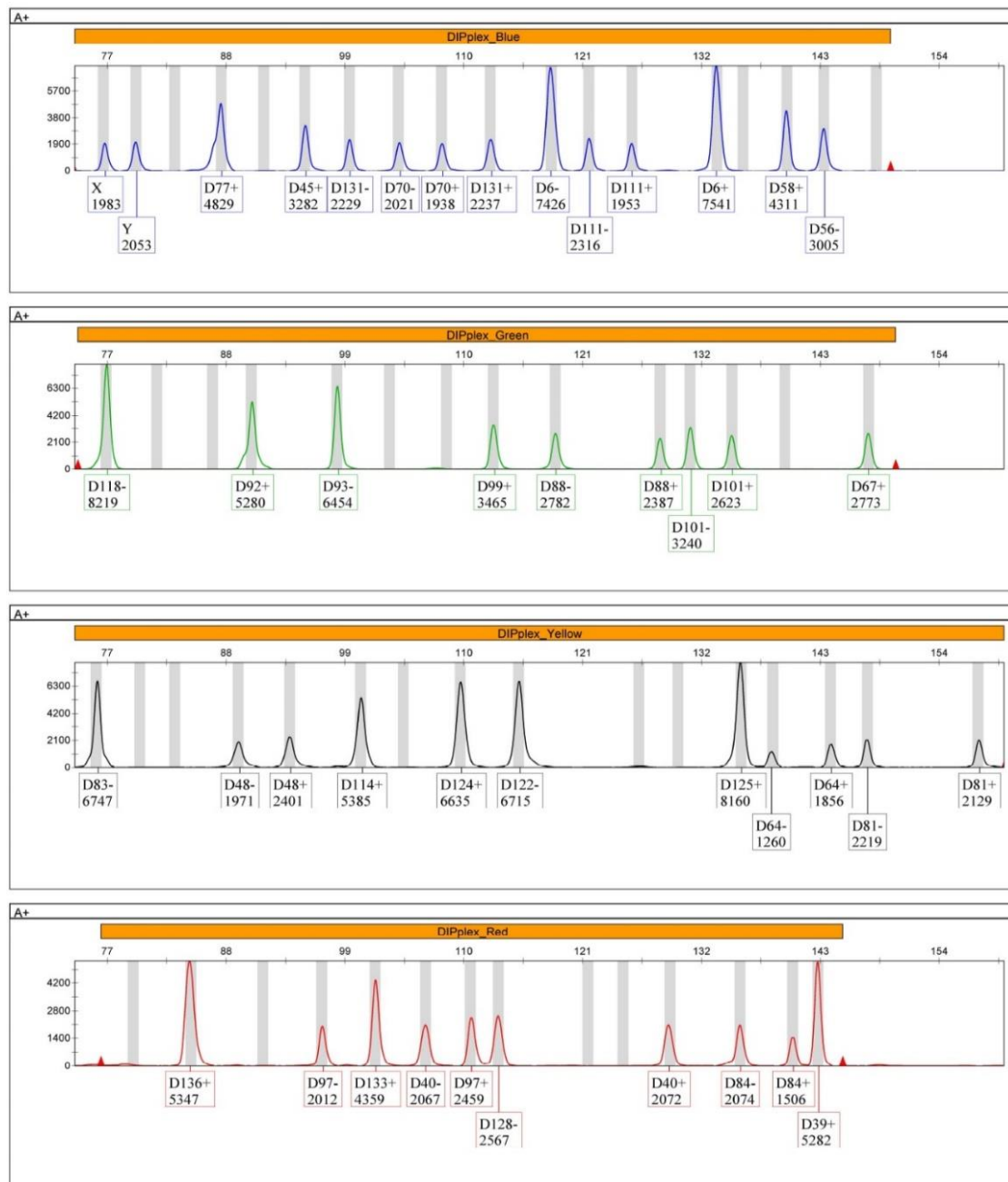


Figure 3.3. Electropherogram of the DIPplex positive control.
The positive control profile generated from 0.5 ng of male DNA 9948.

3.4.3 DIPplex artefacts

3.4.3.1 Allelic imbalance

Allelic imbalance between heterozygous alleles was observed for several DIPplex markers. Among them, HLD128 was the most notable, with 15 individuals (21% of the observed heterozygotes) displaying more than a 50% difference in peak heights. In all cases, the peak for the deletion allele D128- was higher than that for the insertion allele D128+ (**Figure 3.4**).

Moreover, allelic imbalance was also found at HLD101, HLD125, and HLD58 in one, two and four individual(s), respectively.

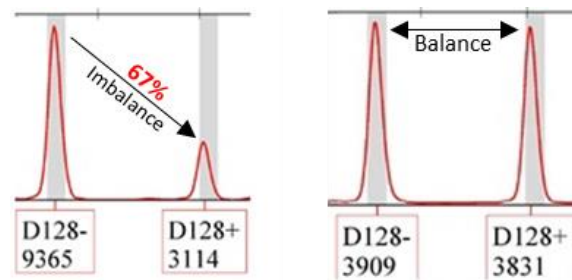


Figure 3.4. Allelic imbalance at the HLD128 locus.

Unbalanced alleles showing a 67% difference in peak height (left), compared to balanced alleles (right), observed in heterozygotes from the Kuwaiti population.

3.4.3.2 Off-ladder allele

The presence of an extra, unclassified allele as a second allele at the HLD58 locus was observed in one of the Kuwaiti samples genotyped. The off-ladder allele appeared as a one base insertion at D84-, as shown in **Figure 3.5**.

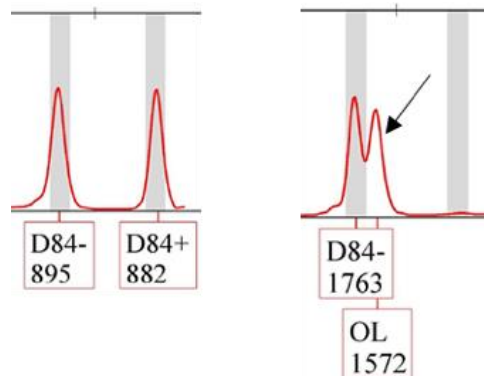


Figure 3.5. Off-ladder allele at the HLD58 locus.

An additional off-ladder allele was observed at the HLD58 locus, indicated by the arrow (right), compared to typical heterozygous peaks (left).

3.4.3.3 Drop-out alleles

A partial drop-out allele (also known as a partial null allele) is an allele that failed to be fully amplified as expected. Mutations in one of the primer binding sites decrease the PCR efficiency of one allele relative to the other, which may cause a severe imbalance or a dropout.

This artefact was seen in 16 of the Kuwaiti sample at the HLD97 locus. Among these individuals, a low-signal insertion allele (D97+) was observed in five heterozygotes and 11 homozygotes, as illustrated in **Figure 3.6**.

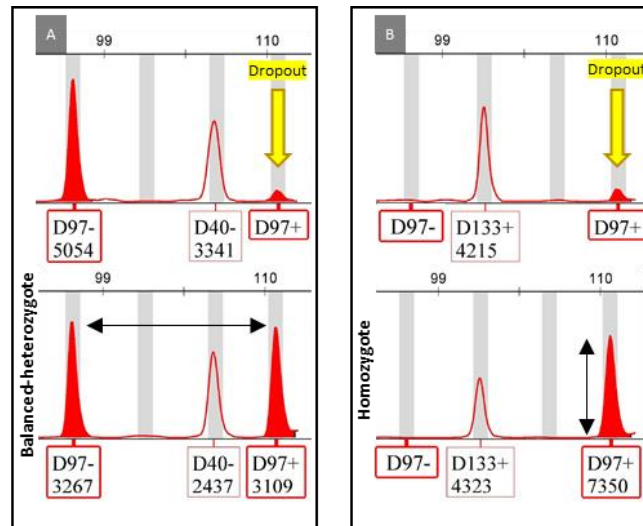


Figure 3.6. The partial drop-out allele of HLD97.

(A) shows the partially dropped-out D97+ allele in a heterozygous individual (top) compared to balanced alleles (bottom). (B) shows the partially dropped-out allele in an individual homozygous for the D97+ allele (top) compared to a normal homozygote (bottom).

3.4.4 Allele and genotype frequencies

Allele frequencies were calculated by genotyping 150 Kuwaiti samples using 30 indel markers. Each marker is represented by two alleles, giving three possible genotypes at each locus: homozygotes indicated by either (del, del) or (ins, ins) and heterozygotes by (del, ins). Therefore, allele frequencies were produced by counting the number of del and ins alleles observed at each marker, then dividing the number by 300, which is the total number of alleles observed in 150 genotypes. Allele frequencies are shown in **Figure 3.7** and genotype frequencies in **Table 3.1**. Of the 30 markers examined, 13 showed higher insertion than deletion frequencies, whereas deletion frequencies were higher for 16 markers, with equal deletion and insertion frequencies at the HLD124 marker.

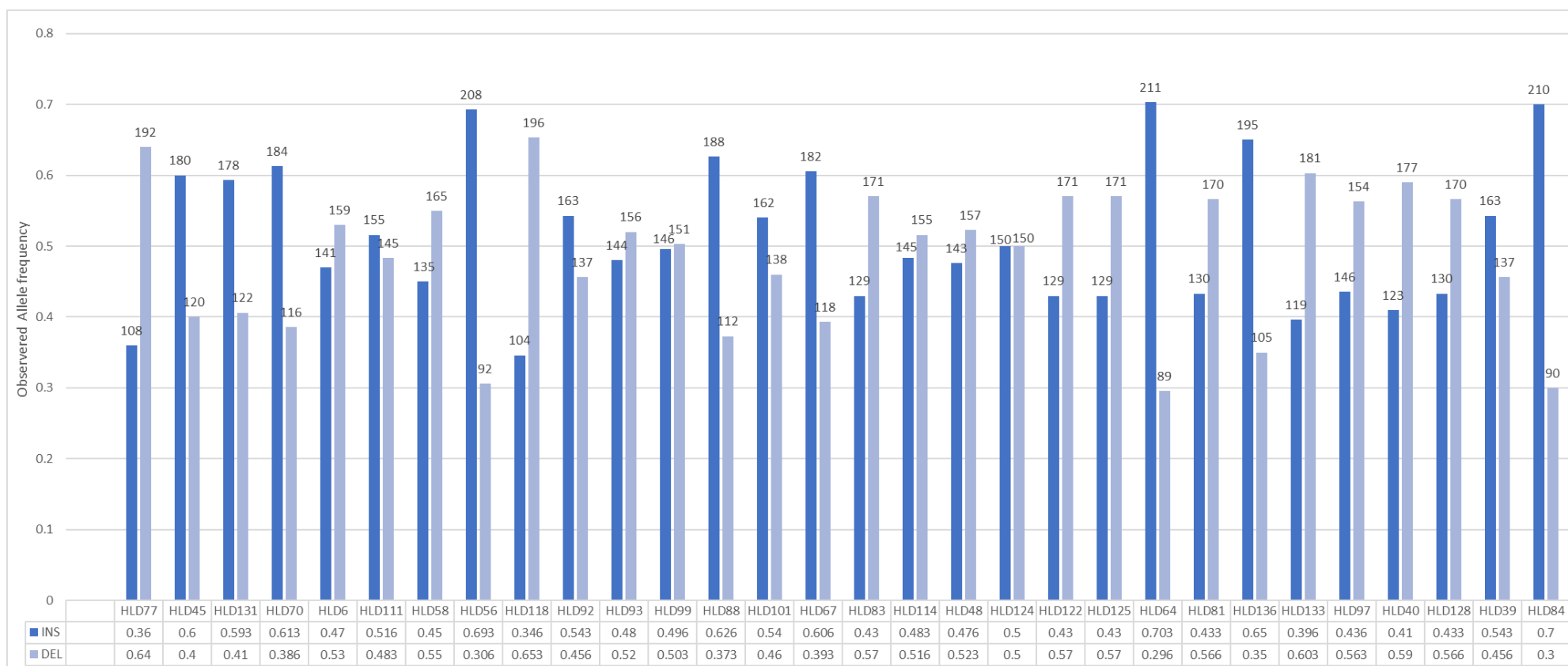


Figure 3.7. DIPplex allele frequencies in the Kuwaiti population.

Allele frequencies for 30 indel markers genotyped in 150 individuals from Kuwait. The number above each bar is the number of deletion (DEL) or insertion (INS) alleles at each marker.

Table 3.1. Genotype frequencies for 30 DIPplex indel loci in the Kuwaiti population.

Locus	Observed genotype number			Observed genotype frequency		
	(DEL, DEL)	(INS, INS)	(DEL, INS)	(DEL, DEL)	(INS, INS)	(DEL, INS)
HLD77	63	21	66	0.420	0.140	0.440
HLD45	19	46	82	0.127	0.307	0.547
HLD131	30	58	62	0.200	0.387	0.413
HLD70	22	56	72	0.147	0.373	0.480
HLD6	43	34	73	0.287	0.227	0.487
HLD111	39	44	67	0.260	0.293	0.447
HLD58	44	29	77	0.293	0.193	0.513
HLD56	21	79	50	0.140	0.527	0.333
HLD118	69	23	58	0.460	0.153	0.387
HLD92	32	45	73	0.213	0.300	0.487
HLD93	36	30	84	0.240	0.200	0.560
HLD99	41	40	69	0.273	0.267	0.460
HLD88	30	68	52	0.200	0.453	0.347
HLD101	35	47	68	0.233	0.313	0.453
HLD67	26	58	66	0.173	0.387	0.440
HLD83	51	69	30	0.340	0.460	0.200
HLD114	42	37	71	0.280	0.247	0.473
HLD48	41	34	75	0.273	0.227	0.500
HLD124	37	37	76	0.247	0.247	0.507
HLD122	51	30	69	0.340	0.200	0.460
HLD125	46	25	79	0.307	0.167	0.527
HLD64	11	72	67	0.073	0.480	0.447
HLD81	52	32	66	0.347	0.213	0.440
HLD136	14	59	77	0.093	0.393	0.513
HLD133	55	24	71	0.367	0.160	0.473
HLD97	55	36	59	0.367	0.240	0.393
HLD40	57	30	63	0.380	0.200	0.420
HLD128	49	29	72	0.327	0.193	0.480
HLD39	29	42	79	0.193	0.280	0.527
HLD84	14	74	62	0.093	0.493	0.413

3.4.5 Tests for Hardy-Weinberg equilibrium

To check whether a given marker conforms to the expectations of HWE, expected genotype frequencies must be obtained for each locus. Expected genotype frequencies are estimated from observed allele frequencies under the Hardy-Weinberg principle, which states that the frequency of deletion and insertion alleles are p and q respectively, and p^2 and q^2 are the probabilities of finding deletion or insertion homozygotes in the next generation, while the probability of finding a heterozygote is $2pq$. Observed and expected values were compared, and p -values were calculated using Fisher's Exact test in Arlequin software, according to the method described by Guo and Thompson [334]. The test resulted in $p > 0.05$ for all markers except for three: HLD56, HLD88 and HLD97 (**Table 3.2**), indicating no statistically significant

deviation from HWE for the majority of markers. The Bonferroni correction was applied to the dataset to decrease the likelihood of finding a false positive (Type I error) in multiple tests [335]. This was performed by dividing the standard significance level of 0.05 by the number of tests being carried out, in this case 30. When the significance value of 0.0016 was used, HLD56 and HLD97 showed no statistically significant deviation from HWE ($p > 0.0016$). In contrast, HLD88 showed statistically significant deviation from HWE ($p = 0.0015$) even after applying the Bonferroni correction. Thus, all but one of the 30 indel loci were shown to conform to the expectations of HWE.

Table 3.2. Results of the Hardy-Weinberg equilibrium (HWE) tests.

Observed and expected heterozygosity and p -values obtained from the Fisher's Exact test. Departure from HWE is highlighted in red.

Locus	Observed heterozygosity	Expected heterozygosity	Exact p -value
HLD77	0.440	0.462	0.5962
HLD45	0.547	0.482	0.1249
HLD131	0.413	0.484	0.0896
HLD70	0.480	0.476	1.0000
HLD6	0.487	0.500	0.8702
HLD111	0.447	0.501	0.1949
HLD58	0.513	0.497	0.7419
HLD56	0.333	0.427	0.0119
HLD118	0.387	0.454	0.0728
HLD92	0.487	0.498	0.8685
HLD93	0.560	0.501	0.1905
HLD99	0.460	0.502	0.3325
HLD88	0.347	0.469	0.0015
HLD101	0.453	0.498	0.3244
HLD67	0.440	0.479	0.3906
HLD83	0.460	0.492	0.5054
HLD114	0.473	0.501	0.5149
HLD48	0.500	0.501	1.0000
HLD124	0.507	0.502	1.0000
HLD122	0.460	0.492	0.5056
HLD125	0.527	0.492	0.4119
HLD64	0.447	0.419	0.4407
HLD81	0.440	0.493	0.2464
HLD136	0.513	0.457	0.1525
HLD133	0.473	0.480	0.8667
HLD97	0.393	0.494	0.0137
HLD40	0.420	0.485	0.1282
HLD128	0.480	0.493	0.8673
HLD39	0.527	0.498	0.5094
HLD84	0.413	0.421	0.8473

3.4.6 Tests for linkage disequilibrium

Analysis of linkage disequilibrium was carried out using the Fisher's Exact test in Genepop software. Linkage disequilibrium tests determine the level of association between all possible pairwise combinations of alleles at different loci. In the 435 pairwise tests of the indel markers, 21 pairs showed significant linkage disequilibrium ($p < 0.05$). However, after Bonferroni correction of the significance level by the number of tests ($0.05/435 = 0.0001$), none of the pairs of markers showed significant linkage disequilibrium.

3.4.7 Analysis of molecular variance

Analysis of molecular variance (AMOVA) was performed in Arlequin to estimate the genetic structure among and within individuals as well as the F -statistics [336]. Wright's F -statistics describe the inbreeding effect, which is defined as the fixation index (F). He developed three fixation indices: the inbreeding coefficient among individuals (f or F_{IS}), total population inbreeding (F or F_{IT}) and inbreeding within subpopulations (θ or F_{ST}) [13, 337]. The AMOVA test showed that 96% of the genetic variation observed in the Kuwaiti sample was found within individuals and only 4% among them. Moreover, the average inbreeding coefficient between individuals (F_{IS}) was 0.0435. F_{IS} values and the respective p -values for the 30 markers are shown in **Table 3.3**.

Table 3.3. Results of the inbreeding coefficient test.

Inbreeding coefficient F_{IS} and the respective p -values for 30 DIPplex markers typed in 150 individuals from the Kuwaiti population.

Marker	FIS	P-value
HLD77	0.04848	0.3303
HLD45	-0.13561	0.96723
HLD131	0.14676	0.05238
HLD70	-0.00865	0.61218
HLD6	0.02649	0.43881
HLD111	0.10898	0.12426
HLD58	-0.0337	0.71208
HLD56	0.21932	0.00535
HLD118	0.14966	0.05248
HLD92	0.02264	0.45931
HLD93	-0.1185	0.9495
HLD99	0.08328	0.20188
HLD88	0.26224	0.00119
HLD101	0.09081	0.17089
HLD67	0.08136	0.20168
HLD83	0.06494	0.26327
HLD114	0.05562	0.30228
HLD48	0.00116	0.5603
HLD124	-0.00999	0.6101
HLD122	0.06494	0.27762
HLD125	-0.07106	0.84584
HLD64	-0.06702	0.85178
HLD81	0.10738	0.1205
HLD136	-0.12491	0.95564
HLD133	0.01444	0.48703
HLD97	0.20371	0.0104
HLD40	0.13516	0.06663
HLD128	0.02597	0.43851
HLD39	-0.05797	0.80723
HLD84	0.01922	0.47861

3.4.8 Forensic parameters

Forensically relevant parameters were calculated using Promega PowerStats for each marker and then combined manually in Microsoft Excel (see **Table 3.4**). The polymorphism information content (PIC) values for the DIPplex markers ranged between 0.25 and 0.50. The combined power of discrimination of the 30 indel markers was greater than 99.999%, and their combined match probability was 2.736×10^{-13} allowing acceptable levels of discrimination (**Table 3.5**) since 15 STR markers in the Kuwaiti population yielded 1.37×10^{-16} [298]. In contrast, the typical paternity index (TPI) ranged between 0.750 and 1.136 and the combined power of exclusion calculated from the 30 markers was 99.46%. These low paternity indices will limit the use of the DIPplex markers in paternity cases.

Table 3.4. Forensic parameters for the DIPplex markers in 150 Kuwaiti individuals.

H_0 = observed heterozygosity; **RMP** = random match probability; **PD** = power of discrimination; **PIC** = polymorphism information content; **PE** = power of exclusion; and **TPI** = typical paternity index.

Marker	H_0	RMP	PD	PIC	PE	TPI
HLD77	0.440	0.390	0.610	0.355	0.140	0.893
HLD45	0.547	0.422	0.578	0.365	0.232	1.103
HLD131	0.413	0.360	0.640	0.366	0.122	0.852
HLD70	0.480	0.391	0.609	0.362	0.171	0.962
HLD6	0.487	0.370	0.630	0.374	0.176	0.974
HLD111	0.447	0.353	0.647	0.375	0.145	0.904
HLD58	0.513	0.387	0.613	0.372	0.199	1.027
HLD56	0.333	0.408	0.592	0.335	0.078	0.750
HLD118	0.387	0.385	0.615	0.350	0.106	0.815
HLD92	0.487	0.372	0.628	0.373	0.176	0.974
HLD93	0.560	0.411	0.589	0.375	0.246	1.136
HLD99	0.460	0.357	0.643	0.375	0.155	0.926
HLD88	0.347	0.366	0.634	0.358	0.085	0.765
HLD101	0.453	0.358	0.642	0.373	0.150	0.915
HLD67	0.440	0.373	0.627	0.363	0.140	0.893
HLD83	0.460	0.367	0.633	0.370	0.155	0.926
HLD114	0.473	0.363	0.637	0.375	0.165	0.949
HLD48	0.500	0.376	0.624	0.374	0.188	1.000
HLD124	0.507	0.378	0.622	0.375	0.193	1.014
HLD122	0.460	0.367	0.633	0.370	0.155	0.926
HLD125	0.527	0.399	0.601	0.370	0.212	1.056
HLD64	0.447	0.435	0.565	0.330	0.145	0.904
HLD81	0.440	0.359	0.641	0.371	0.140	0.893
HLD136	0.513	0.427	0.573	0.351	0.199	1.027
HLD133	0.473	0.384	0.616	0.364	0.165	0.949
HLD97	0.393	0.347	0.653	0.371	0.110	0.824
HLD40	0.420	0.361	0.639	0.367	0.127	0.862
HLD128	0.480	0.374	0.626	0.371	0.171	0.962
HLD39	0.527	0.393	0.607	0.373	0.212	1.056
HLD84	0.413	0.423	0.577	0.332	0.122	0.852

Table 3.5. Summary statistics for population genetic and forensic parameters in the Kuwaiti population.

Forensic parameter	Calculated value
Average heterozygosity	0.461
Average polymorphism information content	0.3645
Combined match probability (CMP)	2.736×10^{-13}
Combined power of discrimination (CPD)	0.9999999999997
Combined power of exclusion (CPE)	0.994
Average paternity index	0.120

3.4.9 Inferring individuals' origins at the global level

The utility of the DIPplex markers for assigning individuals to their population of origin was investigated using the cross-validation test implemented in Snipper App suite version 2.5 [321], using data from five global populations. The results shown in **Table 3.6** indicate that 98.67% of the Kuwaiti individuals were assigned correctly to their origin, with a small percentage being misclassified with the Danes and Chinese (Han). In contrast, 55% of the Danes were misclassified with the Kuwaitis. The probability of correctly classifying each individual in the dataset is visualised in the pentagonal plot in **Figure 3.8**. The distributional pattern shows full overlap between the East Asian populations due to the gene flow between geographically close countries, which consequently causes more similar distributions of allele frequencies among them, whereas the Kuwaitis are the most genetically divergent from the other populations. The misclassified Danish individuals can be seen overlapping with the Kuwaitis in the lower-left part of the plot. Similarly, the PCA analysis of the indel genotypes reflected the same distributional pattern (**Figure 3.9**).

Table 3.6. The success rate of inferring the population of origin of individuals based on 30 indel markers.

	South Korea	China (Han)	China (She)	Denmark	Kuwait
Population of Korean origin	23.59 %	48.26 %	25.47 %	1.34 %	1.34 %
Population of Chinese (Han) origin	0.88 %	29.91 %	64.96 %	3.36 %	0.88 %
Population of Chinese (She) origin	0.84 %	2.52 %	73.95 %	11.76 %	10.92 %
Population of Danish origin	1.25 %	0.00 %	0.00 %	43.75 %	55.00 %
Population of Kuwaiti origin	0.00 %	0.67 %	0.00 %	0.67 %	98.67 %

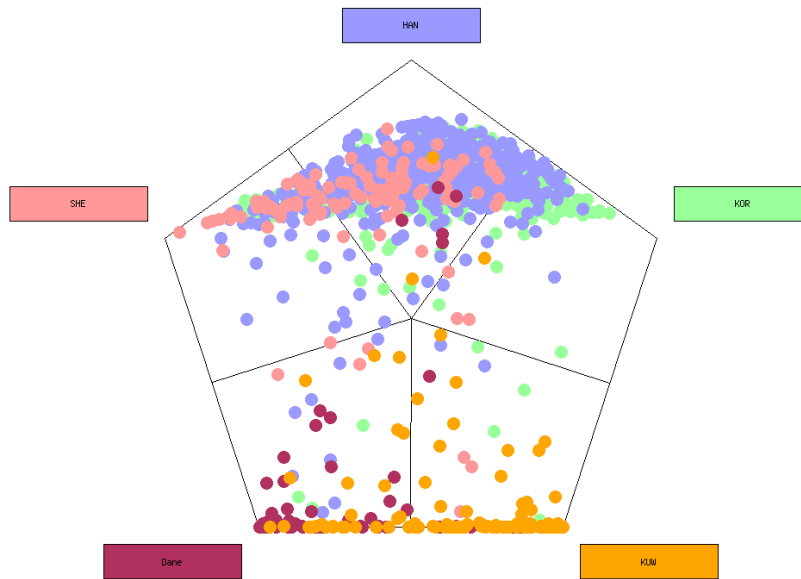


Figure 3.8. Pentagonal plot showing the classification of individuals into their population of origin for 30 indel loci in five global populations.

Pentagon plot visualising the probabilities of assigning 914 genotypes of Kuwaitis (KUW; N = 150), Danes (Dane; N = 80), She Chinese (SHE; N = 119), Han Chinese (HAN; N = 565), and South Koreans (KOR; N = 373) to their original population.

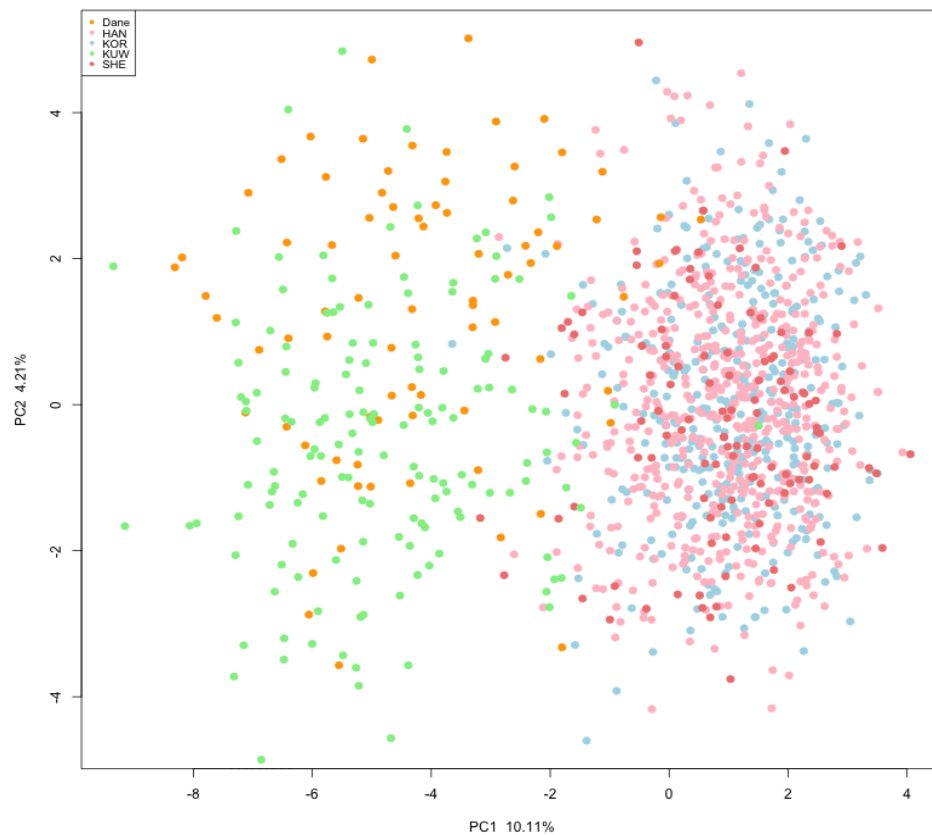


Figure 3.9. Principal component analysis (PCA) of five populations using 30 indel markers.

PCA plot showing the distributional pattern of 914 genotypes of Kuwaitis (KUW; N = 150), Danes (Dane; N = 80), She Chinese (SHE; N = 119), Han Chinese (HAN; N = 565) and South Koreans (KOR; N = 373).

3.4.10 Inferring individuals' origins at the regional level

At the regional level, the cross-validation test showed that accurate individual classification was possible only for Emirati individuals, with a success rate of 96.80% (**Table 3.7**). In contrast, the classification of Kuwaiti and Bahraini individuals had low success at 12% and 22.95%, respectively, and a high rate of misclassification with the Emirati individuals. This is visualised by the triangular plot in **Figure 3.10**, which shows overlap between individuals based on their classification probabilities.

Table 3.7. The success rate of inferring the population of origin, based on 30 indel loci, of individuals from three closely related populations.

	Kuwait	Bahrain	Emirates
Population of Kuwaiti origin	12.00 %	41.33 %	46.67 %
Population of Bahraini origin	1.03 %	22.95 %	76.03 %
Population of Emirati origin	1.60 %	1.60 %	96.80 %

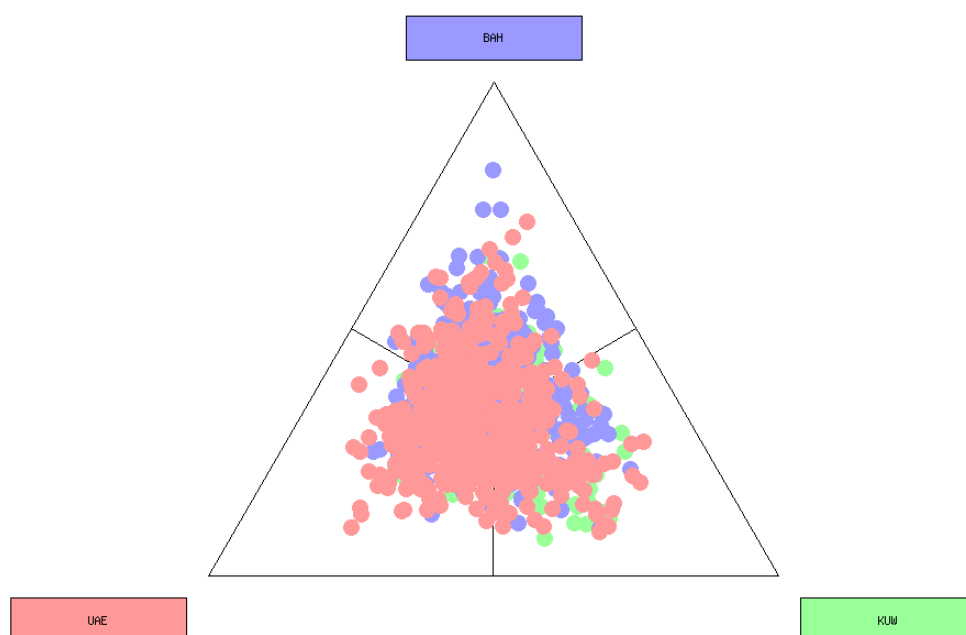


Figure 3.10. Triangular plot showing the classification of individuals into their population of origin for 30 indel loci in three closely related populations.

The plot shows the distributional pattern of the probabilities of assigning 942 indel genotypes of Kuwaiti (KUW; N = 150), Bahraini (BAH; N = 292) and Emirati (UAE; N = 500) individuals to their original population.

3.4.11 Analysis of population structure

Population structure was analysed using STRUCTURE v2.3.4 software to investigate the distribution of genetic variants across populations. Although seven populations were included in the analysis, STRUCTURE results indicated that the most likely number of populations (K) was two, based on the method described in Evanno et al. [316]. As shown in **Figure 3.11**, at $K = 2$, there are two clear clusters comprising East Asians and Middle Easterners with Danes, indicating genetic differentiation between these two groups. A notable pattern is observed at $K = 4$ (**Figure 3.11**), where the Danes are somewhat differentiated in terms of their proportions of genetic ancestry compared to the Middle Eastern individuals, which is in agreement with the PCA distributional pattern in **Figure 3.9**.

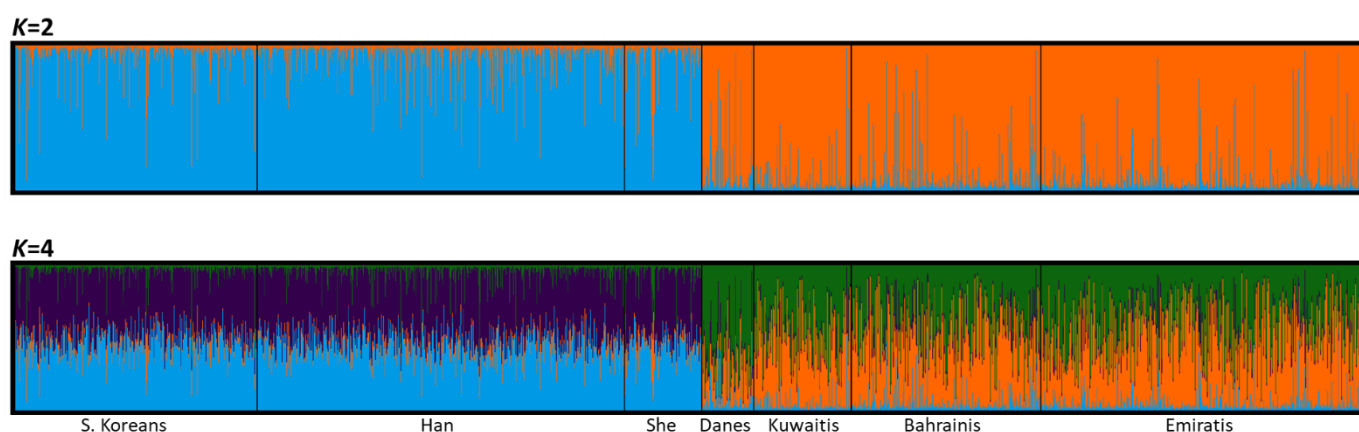


Figure 3.11. Analysis of the structure of seven populations using 30 indel markers.

Assessment of population differentiation using the model-based approach implemented in *Structure* software. The number of colours in each bar represents the number of assumed populations ($K = 2$ and $K = 4$), each vertical line represents an individual, and the amount of each colour in each line represents the proportion of that individual's genetic ancestry that has been assigned to each of the K populations.

3.4.12 Population differentiation

To determine how genetically distant the Kuwaiti population is to other global populations, PCA was carried out using the allele frequencies of the DIPplex markers from data sets obtained from 28 different global populations, including Kuwait. As shown in **Figure 3.12**, the PCA revealed five distinct clusters representing different continents and geographical locations, namely Asia, Europe, South America, Africa, and the Middle East.

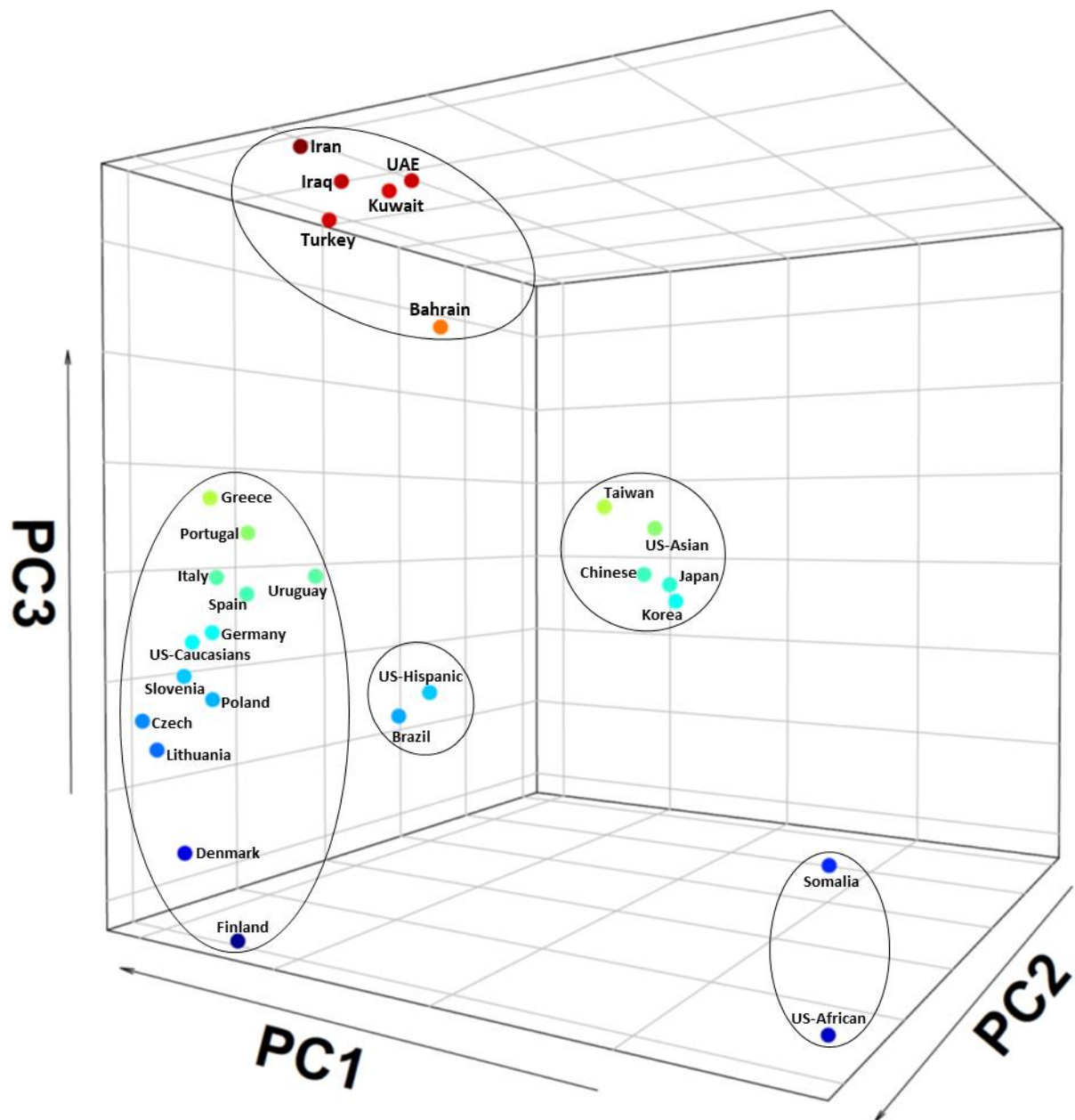


Figure 3.12. 3D-PCA plot of genetic distance based on DIPplex allele frequencies in 28 populations.

PCA (principal component analysis) shows five clusters resulting from the distribution of allele frequencies of 28 populations typed for the 30 DIPplex indel markers. Each circle indicates a group of populations clustered together, representing a continent/geographic region.

3.5 Discussion

In this study, 30 indel markers and the amelogenin sex marker were analysed in a sample of 150 unrelated individuals from the Kuwait population using the Investigator® DIPplex kit. DNA was extracted using a magnetic bead-based extraction technique and quantified using quantitative PCR. Amplification was carried out using DIPplex markers, and indel detection

was performed by CE using the DIPplex allelic ladder and DNA size standard 550 (BTO). Population genetic statistics and forensic parameters were calculated to evaluate the forensic utility of the DIPplex markers in the Kuwait population. The ability of the marker to detect population differentiation was further investigated.

3.5.1 Allelic imbalance

Allelic imbalance usually occurs when a variant (mutation) such as an SNP or indel is located within the primer binding region for a marker, which disrupts the primer binding site [13]. In this study, HLD128 was the most notably imbalanced locus, with 15 individuals exhibiting imbalance of its alleles. A similar observation was reported for individuals in population samples from Turkey, Iraq, Slovenia [137] and Greece [331]. This was explained by two mutations (SNPs) found in the primer binding region of HLD128, 21 bp and 22 bp upstream of the locus [331]. In contrast, allelic imbalance at the HLD125 locus, which was found in two Kuwaiti individuals, was also seen in individuals of Asian origin (living in Texas, USA), but the mutation behind this imbalance is unknown [145]. Imbalance artefacts at HLD58 and HLD101 have not previously been reported in other populations.

Another possible explanation for allelic imbalance is a small amount of input DNA [13]. However, it is not likely to be the case in this study, in which freshly collected samples were used, and DNA was adjusted to the optimal concentration required for amplification. The fundamental problem associated with these phenomena is that when attempting to interpret mixed DNA profiles (containing DNA from more than one contributor), allele peak heights are the best source of information for distinguishing between contributors; therefore, this imbalance will make the interpretation unreliable. To overcome this, it has been suggested that lowering the annealing temperature would reduce the stringency of binding [13]. However, lowering the optimum temperature could result in amplification of non-targeted sequences.

3.5.2 Off-ladder alleles

An off-ladder allele was observed at the HLD84 marker, which could not be identified by the GeneMapper™ software. The allele is referred to as an off-ladder allele because the allelic ladder did not recognise the call for this allele. This off-ladder allele has been reported previously in both Hispanic and African American individuals [143, 145]. The presence of an

off-ladder peak has been explained as being due to the presence of an indel variant located 10 bp downstream of the marker site [143]. This indel has also been reported in African individuals (based on the 1000 Human Genomes Project).

3.5.3 Null alleles

Null (silent) alleles can arise when mutations occur at or near the end of the primer binding region, which disrupts the primer from annealing to its target [13]. Therefore, the allele that is supposed to be amplified will drop out of the marker set and, consequently, will be undetected by electrophoresis. In the Kuwait population, 16 individuals had a partial null allele at the insertion allele for the HLD97 marker.

The HLD97 allele dropout has previously been reported in Danish [124], US Asian, US Caucasian, US Hispanic and African American [143-145], Spanish and Basque [142], Persian [330], Iraqi, Lithuanian, Slovenian, and Turkish individuals [137]. This poorly genotyped allele has been linked to an SNP located 61 bp downstream of the marker within the primer binding region [124, 137, 142-144, 330].

To prevent the occurrence of null alleles, the primer can be either adjusted by moving it away from the mutation site or redesigned in a way that allows it to amplify from the mutation region [338]. However, the former solution is not feasible in this case, because the redesigned primer may interfere with other primers, thereby needing all primers contained in the multiplex to be redesigned. In contrast, the latter approach is more feasible, especially in cases where the mutation causing the null allele is commonly observed [13]. Otherwise, it is advisable to remove the marker completely, due to the issue of observing homozygous genotypes that may actually contain heterozygous alleles where one of the alleles is null.

3.5.4 Heterozygosity and Hardy-Weinberg equilibrium

Heterozygosity is a measure of genetic variation and shows how much genetic variability exists within the population for the tested markers. The average observed heterozygosity across markers in this study was 0.461, indicating moderate diversity in the Kuwait population. Moreover, heterozygosity ranged from 0.333 to 0.560, which shows reasonably high levels of polymorphism.

HWE can be described as a model for predicting the next generation's genotype frequencies based on the allele frequencies observed in the parental generation. The model depends only

on Mendelian laws of allele segregation in diploid organisms, assuming mating is random in a large population which is not influenced by any evolutionary processes such as mutation, genetic drift, natural selection, or migration.

When applying the HWE equation $p^2 + q^2 + 2pq = 1$ for indel markers, p and q represent the allele frequencies of deletion and insertion alleles, respectively, and p^2 , q^2 and $2pq$ are the expected genotype frequencies – the probability of finding homozygotes (insertion or deletion) and heterozygotes in the next generation's offspring – if all the above assumptions of the HWE model are met. When the observed and expected genotype frequencies are similar, the alleles present at a locus or genetic marker are independent, and the locus is in equilibrium.

The p -value of an HWE test represents the probability of observing a given deviation between observed and expected frequencies by chance. The convention is to use a p -value less than 0.05 to reject the null hypothesis that there is no statistically significant difference between observed and expected frequencies, i.e. the probability is less than 5% of finding the observed deviation between observed and expected values by chance. In this study, three markers were observed to violate this null hypothesis at a significance level of 0.05. Similar departures from equilibrium in these three markers were also reported in a number of other populations, as shown in **Table 3.8**.

Table 3.8. Comparison of p -values in tests of the Hardy-Weinberg equilibrium (HWE) between the Kuwait population and other populations.

The departure from HWE of three markers in the Kuwait population that have been observed in other population studies.

Marker	Population	P -value
HLD56	Kuwaiti (this study)	0.0119
	Kazakh (Chinese) [139]	0.0048
HLD88	Kuwaiti (this study)	0.0015
	Miao (Chinese) [339]	0.0190
	Japanese [140]	0.0116
	Taiwanese [133]	0.0301
	Kuwaiti (this study)	0.0141
	Brazilian immigrant [131]	0.0124
HLD97	Greek [331]	0.0230
	Uyghur (Chinese) [129]	0.0306
	Tibetan (Chinese) [129]	0.0296
	She (Chinese) [323]	0.0293
	Basque [142]	0.0475
	Bangladeshi [340]	0.0085
	Somali [128]	0.0160
	Emirati [324]	0.00002

When the Bonferroni correction was applied to the deviating markers, HLD56 and HLD97 were statistically shown to conform to HWE expectations ($p > 0.0016$). However, a significant departure from HWE was still observed at the HLD88 locus. Deviation from HWE can be explained by an excess of homozygotes, which could be due to non-random mating. Inbreeding coefficient (F_{IS}) could be an explanation for the departure from HWE; F_{IS} values approaching +1 reflect inbreeding (excess homozygotes) whereas those approaching -1 reflect outbreeding (excess heterozygotes; see **Table 3.3**). The average F_{IS} across markers was 0.0435, which indicates a moderate level of inbreeding in the Kuwaiti population. This could be explained by consanguineous marriage, which is frequent between Kuwaitis and shows a rate of 54.3% [341]. Allele dropout is another explanation for the excess of homozygotes, which was confirmed at the HLD97 locus, and could also be the reason behind the initial departure of HLD56 and HLD88 from HWE.

3.5.5 Linkage disequilibrium

Linkage disequilibrium describes the level of association between alleles at different loci, and can occur when loci are located on the same chromosome. It can also occur between unlinked loci (on different chromosomes) due to selection, genetic drift, inbreeding or population

substructure [342]. The dataset obtained in this study contains 21 pairs of markers (out of 435 tested pairs) that were in linkage disequilibrium ($p < 0.05$). However, none of these pairs of markers showed significant linkage disequilibrium after applying the Bonferroni correction ($p > 0.0001$), indicating that all loci are statistically independent and can, therefore, be used for the calculation of genotype frequencies in the Kuwaiti population.

3.5.6 Forensic parameters

The PIC is used to estimate the level of polymorphism of a marker. The markers with $PIC > 0.5$ are highly polymorphic, whereas low levels of polymorphism are indicated by a $PIC < 0.25$ [314]. The PIC values for the DIPplex markers considered in this study were in the intermediate range, between 0.25 and 0.50.

The random match probability is the probability that two unrelated individuals selected randomly from the studied population possess identical genotypes. The lower the probability, the smaller the likelihood of a match between an individual and crime scene sample. The combined match probability for the 30 markers was 2.736×10^{-13} , i.e. the likelihood of finding two identical profiles in the population is 1 in 3.70×10^{12} . In comparison, the combined match probability obtained from typing 15 STR markers in the Kuwaiti population was considerably smaller (1.37×10^{-16}) [298], meaning that the probability of a random match is lower when using STR markers than when using indel markers.

Discriminatory power is the probability that two unrelated individuals selected randomly from the studied population possess different genotypes. Markers with high discriminatory power are better able to distinguish between two unrelated individuals. The combined power of discrimination of the DIPplex markers in the Kuwaiti population was calculated as $> 99.999\%$, i.e. there is a 99.999% chance of these markers discriminating between two individuals.

The power of exclusion is used in paternity testing to assess the efficiency of a locus to exclude a random individual as being the alleged parent. In this study, the combined power of exclusion of 30 indels was 99.46%. In contrast, the combined power of exclusion exceeds 99.999% using only 15 STRs in the Kuwaiti population [298].

The typical paternity index (TPI) used in paternity testing is the likelihood ratio of the probability of the alleged man being the biological father of an offspring compared to a

random individual in the population. The larger the value is, the greater the chance that the alleged father is the biological father compared to a random male. In this study, the TPI ranged between 0.750 and 1.136 (average = 0.936) for the 30 indels. In contrast, STR markers have shown a high TPI of 2.475 in the Kuwaiti population [298]. Therefore, indel markers are not informative for paternity testing as a stand-alone system. However, the advantages of their low mutational rate and simple genotyping technique allow them to be useful in difficult paternity testing scenarios, such as when STR markers have been inconclusive (for example if mutations exist in families) or in cases where additional markers are needed to provide supporting evidence.

3.5.7 Genotype assignment and ancestry inference

Although DIPplex markers were originally chosen for the purposes of human identification, some of them can be used as ancestry informative markers (AIMs) due to the similarity of their allele frequencies between specific populations and differences between others [133-136]. Provision of intelligence on the origin/ancestry of unknown/unidentified samples can be beneficial in forensic cases, especially when there is no human eyewitness description, when the description is unreliable, or when the DNA profile has no match in the database. It must be noted that 'ancestry' in this context is biogeographic ancestry, i.e. individuals' origin is associated with the geographical location.

On this basis, two tests were carried out, bounded by the availability of the genotypes in the literature, to investigate whether these indels can serve as informative ancestry markers to identify Kuwaiti individuals from the other populations. Cross-validation tests (each genotype was removed and classified by the remaining dataset) were performed using the Snipper. The analysis showed that indel markers could strongly discriminate between Kuwaiti individuals and individuals from East Asian populations, but their probabilities were low in distinguishing between Kuwaiti individuals and individuals from closely related populations (Bahrain and UAE). This was confirmed by PCA and population structure analysis, which showed a clear signal of genetic differentiation between the East Asian and Middle Eastern populations, but the Danes and Middle Eastern populations could not be successfully classified.

Although the test was nearly 99% successful in classifying Kuwaiti genotypes (**Table 3.6**), more than half of the Danish individuals were misclassified with Kuwaiti individuals. The

distributional pattern observed in PCA and the genetic similarity seen in the STRUCTURE analysis at $K = 2$ (**Figure 3.11**) suggest that the 30 markers are not informative enough to discriminate between the Kuwaiti population and European populations (Danes), even with the slight genetic differences seen between the two samples at $K = 4$ (**Figure 3.11**). However, data from more populations would be useful in getting a better understanding of the utility of the DIPplex markers as AIM markers. Even when considering other assays that have been specifically designed for ancestry identification like AIM-Indels [343], the lack of data from a large number of populations, especially for the Middle East region, would still be a significant limitation.

3.5.8 Population differentiation

To investigate the genetic distance between the Kuwaiti population and other global populations, PCA was performed using allele frequencies of the 30 indel markers from 28 populations (including Kuwait). As shown in **Figure 3.12**, the markers permitted the clustering of the populations into five groups according to their geographical location. These clusters indicate that the geographical distance between locations restricts gene flow between populations. Thus, there is more gene flow between geographically close countries, which consequently causes more similar distributions of allele frequencies among them. For example, the Kuwaiti population studied here is more closely genetically related to the populations in surrounding countries, such as Iran, Iraq and Turkey, than it is to populations from other countries on different continents. Moreover, in this study, Uruguay was found to be clustered with European populations. This is in agreement with the previous finding that Uruguayans are European descendants sharing similar patterns of allele frequencies to European populations [332]. Overall, the allele frequencies of the indel markers in the DIPplex kit can be used to successfully separate geographically distant global populations on the basis of their genetic distance, making them a good choice for determining genetic differentiation between populations.

3.6 Conclusion

This study was undertaken to evaluate the forensic utility of the 30 indel markers in the Qiagen Investigator® DIPplex kit. The genotyping results showed that the majority of the DIPplex markers were amplified successfully, but some loci exhibited artefacts such as allelic imbalance, an off-ladder allele, and partially null alleles. These artefacts were most likely due

to the presence of genetic variants in the relevant primer binding sites. Only one marker deviated from the expectations of HWE after Bonferroni correction, due to an observed excess of homozygotes. A high level of homozygosity could be due to either allele dropout or inbreeding, and the latter is supported by a moderate level of inbreeding between Kuwaiti individuals. Results also showed no significant linkage disequilibrium in all possible pairwise combinations of loci (after applying the Bonferroni correction), meaning that the DIPplex loci were independent, allowing them to be used to calculate genotype frequencies.

The high combined power of discrimination and low combined match probability of the indel markers indicate that they are reasonably informative, allowing them to be used as forensic markers for identification in the Kuwaiti population. The low paternity indices yielded by the DIPplex markers were not informative, which limits their usefulness as a typing system for paternity/kinship cases in the Kuwaiti population. However, they can be used in addition to STR markers in challenging kinship cases where mutations exist in familial STR profiles. This study confirms previous findings that diallelic markers are less informative than STRs, and contributes additional evidence suggesting that more indel markers are needed to increase the power of DNA evidence. The indel markers in the DIPplex kit can separate geographically distant global populations, such as between East Asian and Middle Eastern populations, based on their genetic distance. However, the results suggest that the 30 indels do not discriminate between the Kuwaiti population and European populations (Danes). This justifies that more studies involving more populations are needed to allow an accurate assessment of genetic differentiation.

Chapter 4 Analysis of 23 Autosomal STR markers in the Kuwaiti population

4.1 Introduction

Currently, short random repeats (STRs), dominantly typed using the polymerase chain reaction (PCR)-capillary electrophoresis (CE) system, are the primary genetic markers used in almost every forensic laboratory worldwide to identify evidentiary materials and investigate paternity cases. They became the gold standard for DNA investigation due to their highly polymorphic repeat units making them highly informative, the simple workflow (PCR-CE) involved in their typing, as well as the existence of well-established global databases of allele frequencies that permit the statistical evaluation of STR profiles.

At present, forensic DNA analysis in Kuwait is carried out by the Kuwaiti Identification DNA Laboratory (KIDL) using only STR markers, including autosomal STRs, Y-chromosome STRs (Y-STRs) and X-chromosome STRs (X-STRs). Autosomal STRs are routinely used both for identification of individuals and paternity testing, whereas Y-STRs and X-STRs are used less frequently, and only for specific scenarios. To date, few papers have been published investigating the forensic utility and genetic diversity of autosomal STR markers in the Kuwaiti population. In 2008, Alenizi and colleagues reported the allele frequencies of 15 STR loci included in the AmpF ℓ STR $^{\circ}$ Identifiler $^{\circ}$ kit) [298]. Based on these 15 STRs, the genetic distances between Kuwaiti nationals and foreign nationals from seven other populations residing in Kuwait were found to be consistent with their geographical distances [299]. Another recent study investigated the forensic utility of 25 autosomal STRs included in two separate kits: the PowerPlex $^{\circ}$ CS7 system and the PowerPlex $^{\circ}$ 21 system (Promega Corporation, WI, USA) [300]. Although these existing STRs are efficient for analysing cases of simple relationships, more STRs are increasingly required, particularly for complex paternity cases, and to increase the discrimination power in cases of partial DNA profiles and DNA mixtures.

Recently, Promega launched the PowerPlex $^{\circ}$ Fusion 6C kit, a six-dye kit that can amplify 27 loci, including the 20 autosomal STR loci in the expanded CODIS set (CSF1PO, FGA, TH01, TPOX, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, and D22S1045) [344], three

additional autosomal STRs (PentaE, PentaD, and SE33) to increase the power of discrimination, two sex chromosome markers (amelogenin and DYS391), and two rapidly mutating Y-STRs (DYS570 and DYS576) [345]. The PowerPlex® Fusion 6C kit was validated by multi-laboratory evaluation following SWGDAM guidelines [345]. This part of the research will focus on assessing the population genetics and forensic utility of the 23 autosomal STRs before implementing this kit for criminal and relationship cases in Kuwait.

4.2 Aims and objectives

The main aim of this chapter is to evaluate the forensic utility of 23 autosomal STR markers in the PowerPlex® Fusion 6C kit, which uses the PCR-CE method, using 400 samples from the Kuwaiti population. Four markers (D10S1248, D22S1045, D2S441 and SE33) will be reported for the first time in the Kuwaiti population. A second aim is to generate data comparable to those generated by NGS technology (reported in chapter 5) to examine the concordance of the markers collected using different methodologies. This study involves:

- Generating allele frequency data for the STR makers to increase the amount of genetic data available for the Kuwaiti population.
- Calculating population and forensic parameters to assess the forensic utility of the STR markers in this population.
- Examining the genetic relationship between the Kuwaiti population and other populations at a global and regional level.

4.3 Materials and methods

Sample collection was carried out in the Kuwait Identification DNA Laboratory (KIDL), which is part of the General Department of Criminal Evidence. Sample collection was carried out by PhD student Mahdi Haidar. STR profiling was carried out by MSc student Fatimah Abbas at the University of Strathclyde. Data and statistical analysis were carried out by PhD student Mahdi Haidar at the University of Strathclyde.

4.3.1 Samples and genotyping

Blood samples were collected on Whatman® FTA® cards (GE Healthcare Life Sciences, IL, USA) from 400 unrelated Kuwaiti (253 males and 147 females). Sample collection and ethical approval are described in section 2.2., and the participant information sheet and consent form are shown in **Appendix 1** (A1.2). DNA was amplified directly from a 1.2 mm FTA® card

punch as described in section **2.6.1**. Detection and separation of the DNA fragments were carried out using an Applied Biosystems 3500 Genetic Analyzer with the internal lane standard WEN ILS 500 and allelic ladder provided with the PowerPlex® Fusion 6C kit as described in section **2.6.2**. Genotype determination and allele calling for only the 23 autosomal loci were carried out using GeneMapper™ ID-X software version 1.4.

4.3.2 Statistical analysis

Arlequin statistical software version 3.5 was used to calculate allele frequencies, to test for linkage disequilibrium, and to test for deviation from the Hardy-Weinberg Equilibrium [309]. Forensic parameters, including the random match probability (RMP), discrimination power (DP), power of exclusion (PE), typical paternity index (TPI) and polymorphic information content (PIC), were calculated using *STRAF*, an online tool for STR data analysis [310].

4.3.3 Intra-population genetic structure among Kuwaitis

Countries in the Arabian Peninsula, including Kuwait, have a high rate of consanguineous marriage, which causes differential distribution of alleles among families and tribes; this commonly results in genetic stratification in populations in this region [289, 346]. Since the subpopulations within a large population may have different distributions of allele frequencies, it is essential for newly introduced markers to be assessed for the presence of any hidden population structure, in order to avoid calculation of forensic parameters using inaccurate allele frequencies taken from the total population, rather than the relevant subpopulation. The presence of such stratification would negatively impact the accuracy of forensic indices, such as the random match probability of the markers, because there would be a higher chance of random individuals possessing similar genotypes within a subpopulation than within the total population [347]. Therefore, two different methods were used to detect any hidden genetic structure in the population, principal component analysis (PCA), and a Bayesian-based method implemented in *STRUCTURE* version 2.3.4 [315, 348, 349].

In order to demonstrate whether these two methods were able to cluster the samples into their real subpopulations, each sample was categorised into one of three ancestral subgroups ($K=3$) based on the donor's surname. Generally, according to the Kuwaiti naming convention, a person's full name can consist of four names: the first is the given name, the second is the father's name, the third is the grandfather's name, and the fourth is the surname indicating

the tribal/family lineage. However, specific personal names can also provide links to the ancestry of the individual. Therefore, the procedure taken to identify origin was mainly based on surnames and sometimes on the other names. Because Kuwait is a small country, most of the individuals' origins were identified. For those that were not identified, two approaches were undertaken to ascertain the origin: consultation of native Kuwaitis who are familiar with tribal names/surnames; and use of websites/forums populated with information on genealogy and surnames.

The Kuwaiti population was previously found to be mainly composed of settlers originating from three different regions: the Arabian Peninsula (from Saudi Arabia), the desert (representing nomadic tribes), and Persian countries (mainly Iran) [289-292, 350]. This is supported by the migration of tribal families from Najd (in Saudi Arabia), nomadic tribes from neighbouring countries such as Iraq and Saudi Arabia wishing to settle permanently instead of living an itinerant life in the desert, and the migration of Persian families crossing the sea to work as merchants and labourers. The modern genetic structure of the Kuwaiti population has been significantly influenced by these migrations.

On this basis, the samples were categorised into three groups: KW-1 (n=162) representing individuals originating from the Arabian Peninsula, KW-2 (n=163), which consisted of those originating from Persian countries, and KW-3 (n=75), which was composed of Bedouin individuals from nomadic tribes. PCA was carried out on allele frequencies at the 23 autosomal STR loci for the three groups using R software and visualised using the factoextra package.

In contrast to PCA, which is an unsupervised clustering algorithm, STRUCTURE (a Bayesian-based approach) takes a range of different possible numbers of populations (K) in order to calculate the proportion of the genome of each individual in the sample originating from each inferred population [315]. STRUCTURE software calculates the likelihood of the data (X) for range of K values, and the true number of K is determined by the maximal value of $\ln P(X|K)$. However, it was found by Evanno et al. that the maximal value does not always provide the correct number of K in the data [316]. Instead, the maximal value of the rate of change (ΔK) in the $\ln P(X|K)$ between successive K values accurately infers the true number of genetic clusters in the data [316]. As such, both $\ln P(X|K)$ and ΔK at each K were calculated and reported. STRUCTURE was run without population information, as recommended by the

STRUCTURE documentation, in order to check whether the results approximately agreed with the separation of samples into their subgroups. Thus, the predefined groups (KW-1 to KW-3) were only included as a population label rather than as prior information for the analysis. The parameters for the analysis were set as follows: 'admixture' and 'correlated allele frequencies' models using 100,000 Markov Chain Monte Carlo (MCMC) steps for each run, with the first 100,000 discarded as a burn-in, and the inferred number of K was set from 2 to 10. At each K , the analysis was repeated five times in order to test the results for consistency. The results were visualised using CLUMPAK (Clustering Markov Packager Across K) [326, 327], and the best K was calculated using STRUCTURE HARVESTER [328, 329].

4.3.4 Inter-population genetic structure and population relationship

To assess the genetic distance between the Kuwaiti sample and other global populations, PCA was conducted based on allele frequencies of the 23 autosomal loci for 57 global populations grouped into seven continental regions: Africa (AFR), America (AMR), Central and South Asia (C_S_ASIA), the Middle East (ME), Europe (EUR) and East Asia (E_ASIA). Data for the global populations were obtained from the HGDP-CEPH Human Genome Diversity Panel (HGDP-CEPH) [351-353], using the online forensic STR frequency browser, popSTR [354]. Data from Lebanon (LEB) and an Indian (IND) population from Madhya Pradesh typed for the 23 autosomal loci [355, 356] were also included in the analysis. Genetic distance was also assessed at a regional level using allele frequencies for 13 of the 23 autosomal loci (CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA) that are shared between the data reported in this study (KW) and other studies of Kuwait and neighbouring countries: Kuwait (KW1 [300] and KW2 [299]), Iran (IRN [357] and IRN1 [299]), Saudi Arabia (SA [358] and SA1 [359]), Qatar (QAT [360]), Oman (OMN [361]), Yemen (YEM [361]), United Arab Emirates (UAE [362]), Bahrain (BAH [363]), and Iraq (IRQ [364] and IRQ1 [299]). PCA analysis was conducted using R software and visualised using the *factoextra* package.

In addition to the PCA, the genetic relationship between the Kuwaiti samples and the other populations, at both the continental and regional levels, was investigated using phylogenetic trees. These trees were constructed using pairwise genetic distances (D_A) based on Nei et al. [365], which were calculated from the allele frequencies of the populations using *POPTREE2*

software [366]. The type of phylogenetic trees used were Neighbour-joining (NJ) trees, constructed using *Mega X* software version 10.0.5 [367].

4.4 Results and Discussion

4.4.1 STR genotyping

The PowerPlex® Fusion 6C kit uses specific primers designed to amplify each STR locus and labelled with one of five fluorescent dyes that are detected by CE. Samples were analysed using the Applied Biosystems 3500 Genetic Analyzer, in which amplicons are separated according to their size and arranged based on the dye colour emitted. Peak sizes are determined by a calibration curve, which is produced from internal size standard fragments of known length; the internal size standard used here was WEN ILS 500, which contains 21 fragments (**Figure 4.1**). Alleles/genotypes are determined by comparing peaks with an allelic ladder, which contains all the possible alleles for each locus. No sign of contamination was detected in any of the negative controls. The positive control DNA 2800M profile is shown in **Figure 4.1** and was in concordance with the positive control profile given in the manufacturer's protocol. The genotyping resulted in full profiles for all 400 Kuwaiti individuals using the PowerPlex® Fusion 6C kit.

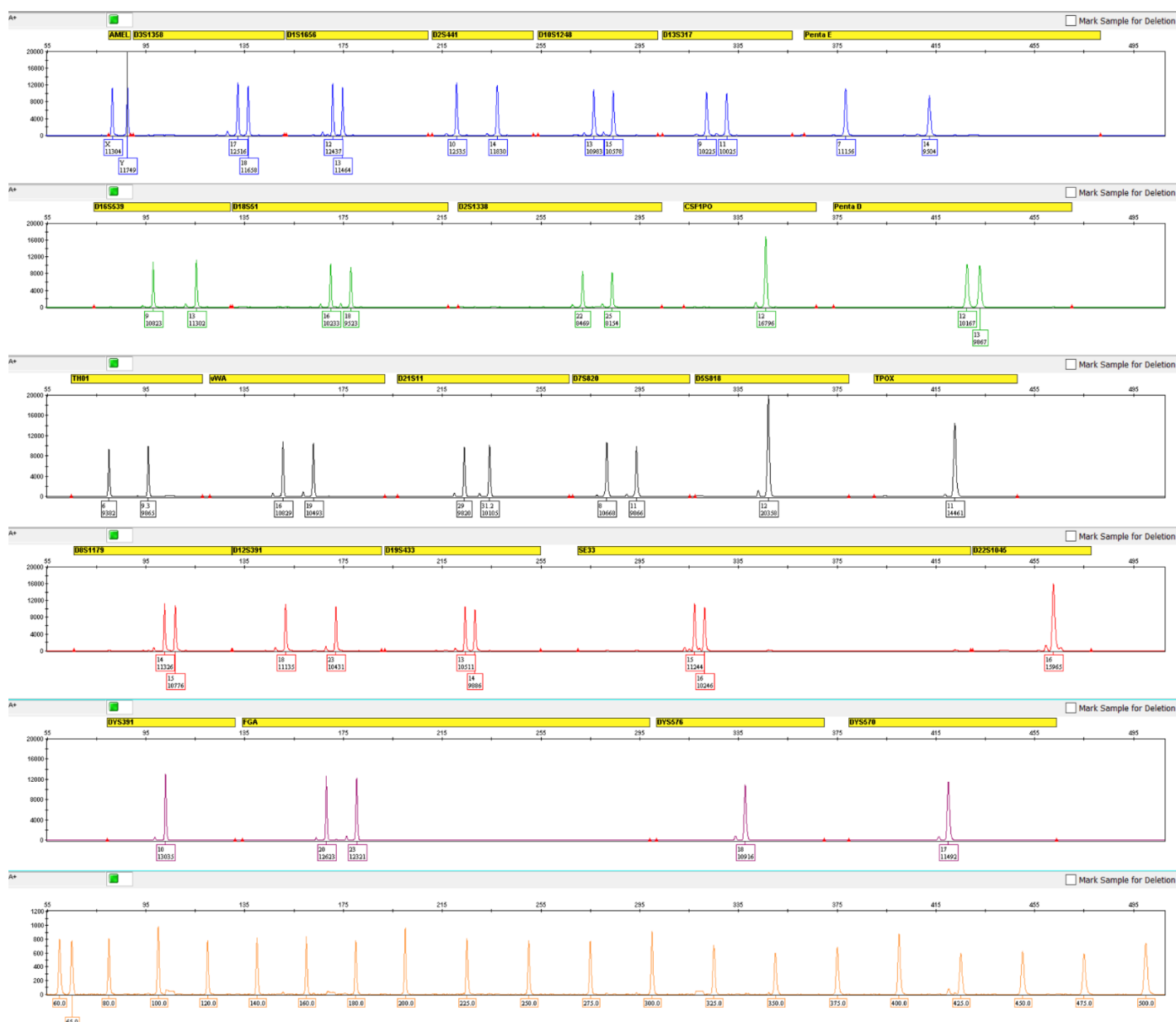


Figure 4.1. Electropherogram of the PowerPlex® Fusion 6C positive control 2800M.

Loci are labelled with five colors (fluorescent dyes) and distributed across five lanes. The last shows the 21 fragments of the internal size standard WEN ILS 500, labelled with a sixth dye.

4.4.2 STR Off-ladder and novel alleles

Alleles that could not be identified using the *GeneMapper™* allelic ladder for the PowerPlex® Fusion 6C kit were assigned as off-ladder (OL) alleles, and these were observed in 13 samples. These samples were re-amplified for confirmation and all OL alleles were confirmed. Five different OL alleles were observed at both PentaE (**Figure 4.2**) and SE33 (**Figure 4.3**), and one each at the PentaD (**Figure 4.4**), D22S1045 (**Figure 4.5**) and D18S51 (**Figure 4.6**) loci. The

samples were also sequenced using the Verogen ForenSeq™ DNA Signature Prep kit (see Chapter 5), and these data were examined to determine whether the undesigned alleles at the PentaE, PentaD, D22S1045, and SE33 loci could be identified. As a result, these alleles were all identified by their sequences, as shown in **Table 4.1**. The D18S51 locus is not included in the Verogen ForenSeq™ DNA Signature Prep kit, therefore its OL allele was identified using the allelic ladder bins created in *GeneMapper™* software. All of the identified alleles have been reported previously in the STRBase database (an online STR database created by the United States National Institute of Standards and Technology (NIST) [368]), except for the PentaD 11.2 allele, which is a novel allele not reported before in the literature.

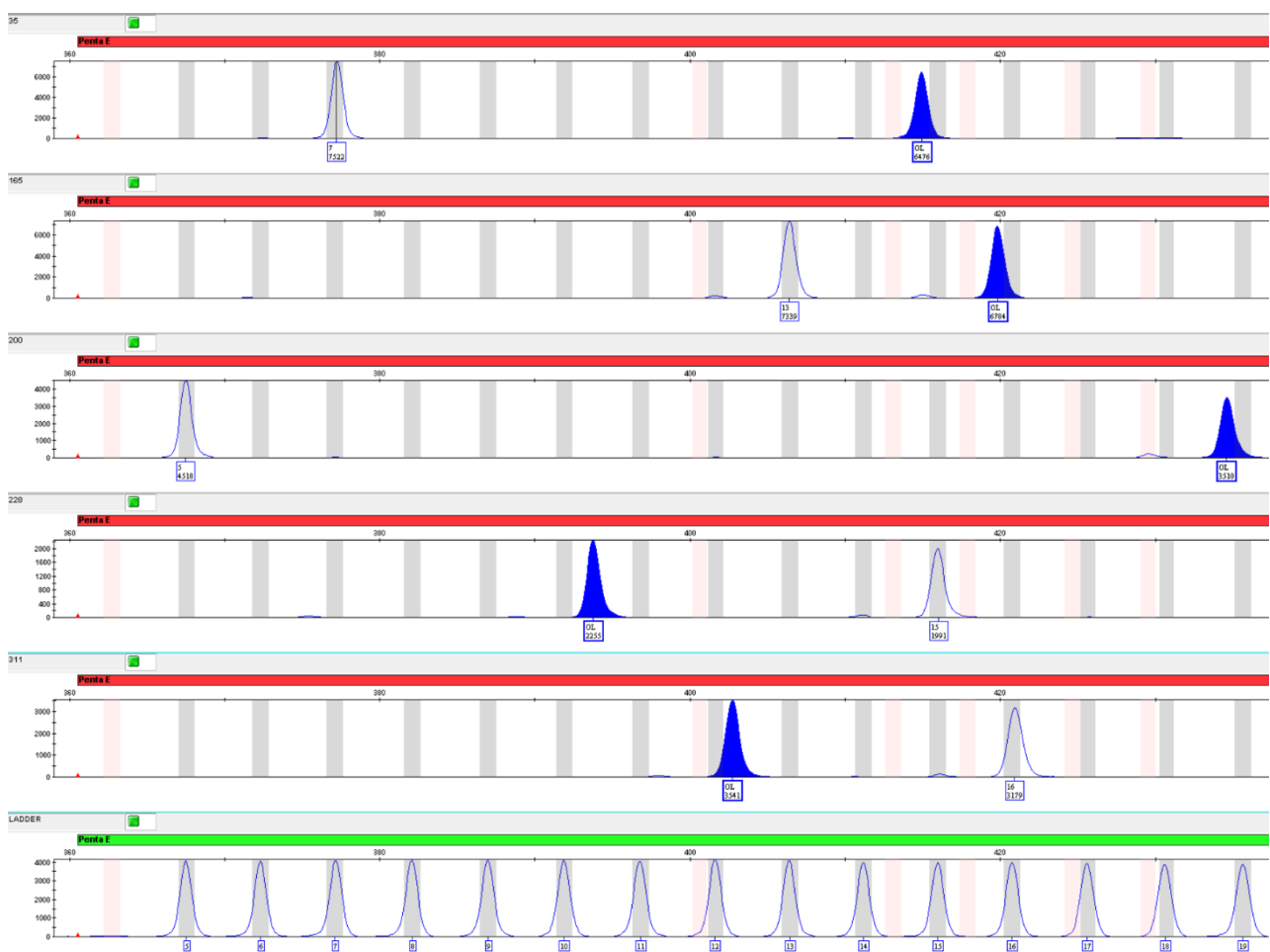


Figure 4.2. PentaE off-ladder alleles.

Five off-ladder (OL) alleles observed at the PentaE locus highlighted in blue, compared to the allelic ladder of the associated locus in the bottom panel.

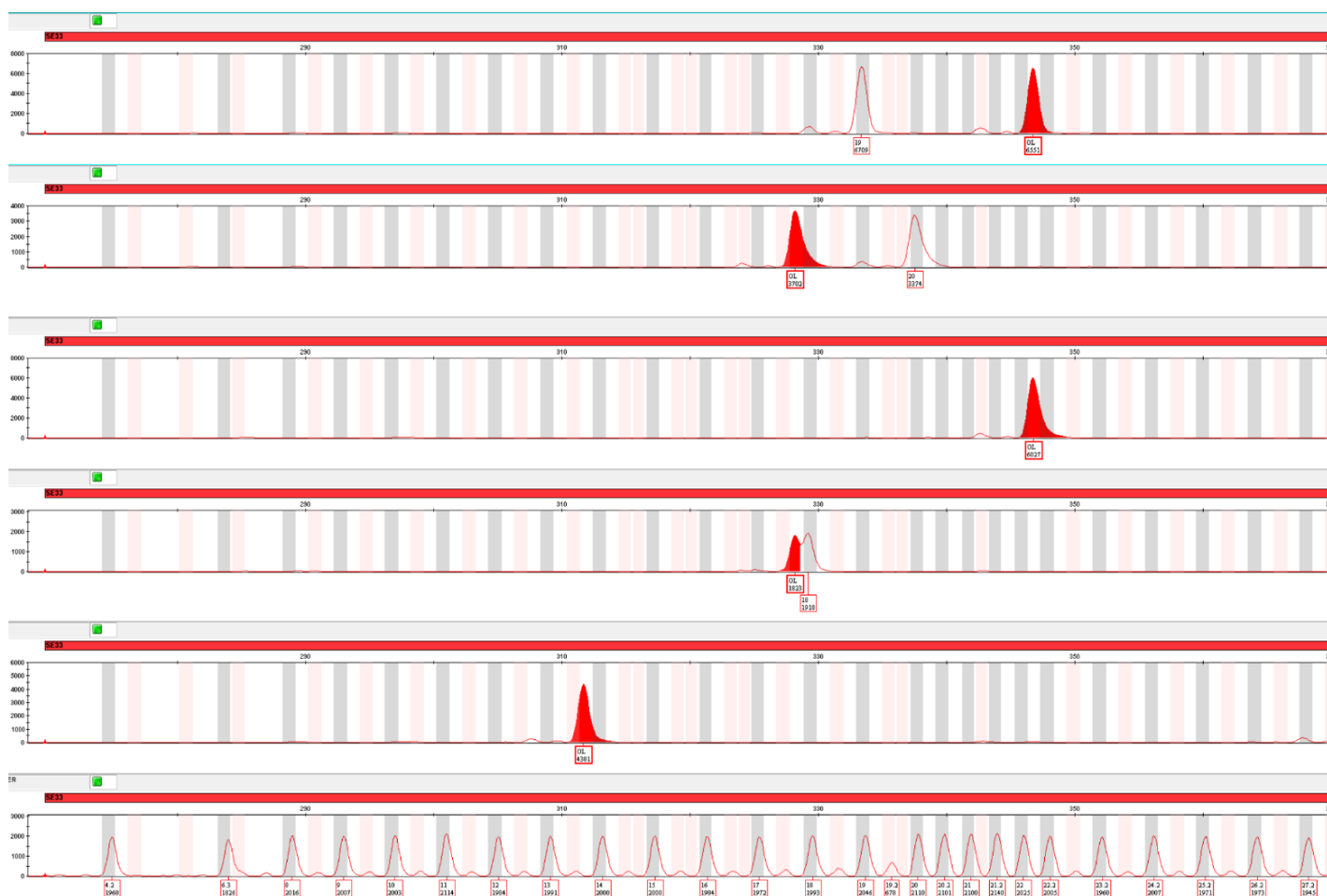


Figure 4.3. SE33 off-ladder alleles.

Five off-ladder (OL) alleles observed at the SE33 locus highlighted in red, compared to the allelic ladder of the associated locus in the bottom panel.

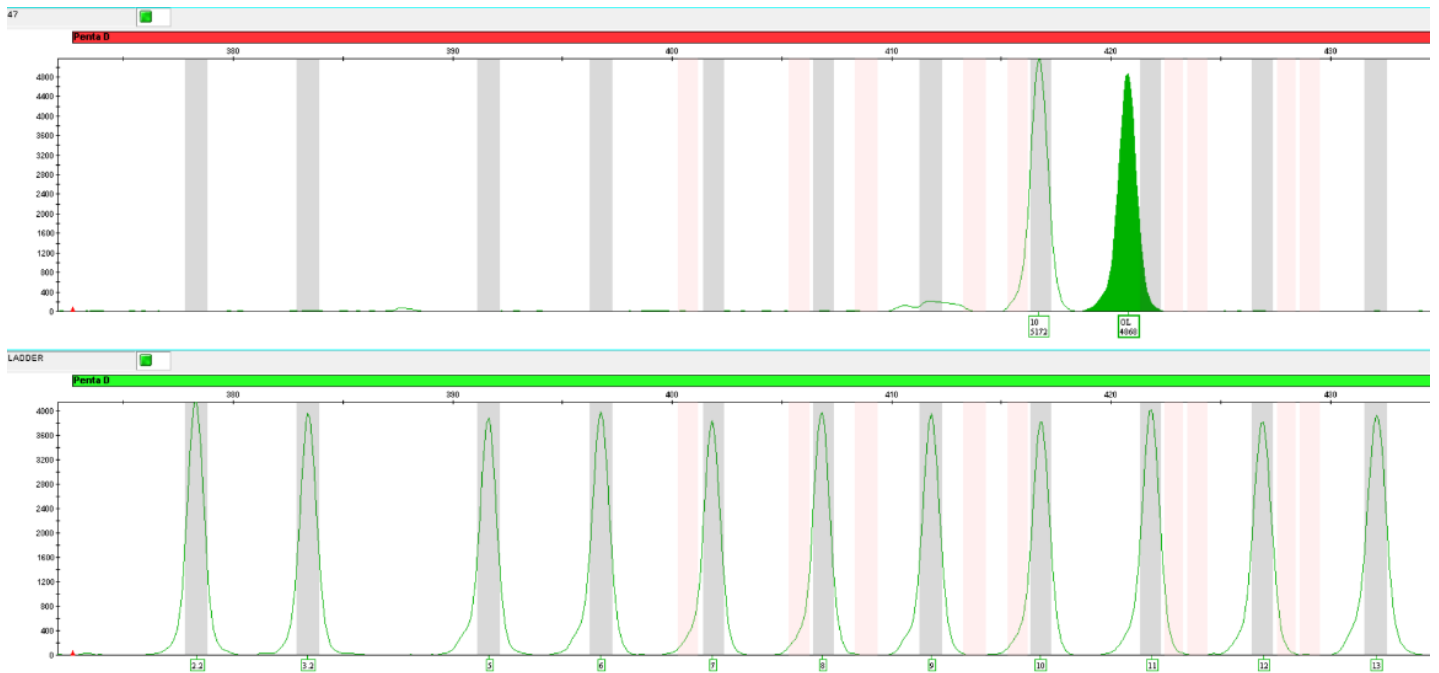


Figure 4.4. PentaD off-ladder allele.

One off-ladder (OL) allele observed at the PentaD locus highlighted in green, compared to the allelic ladder of the associated locus in the bottom panel.

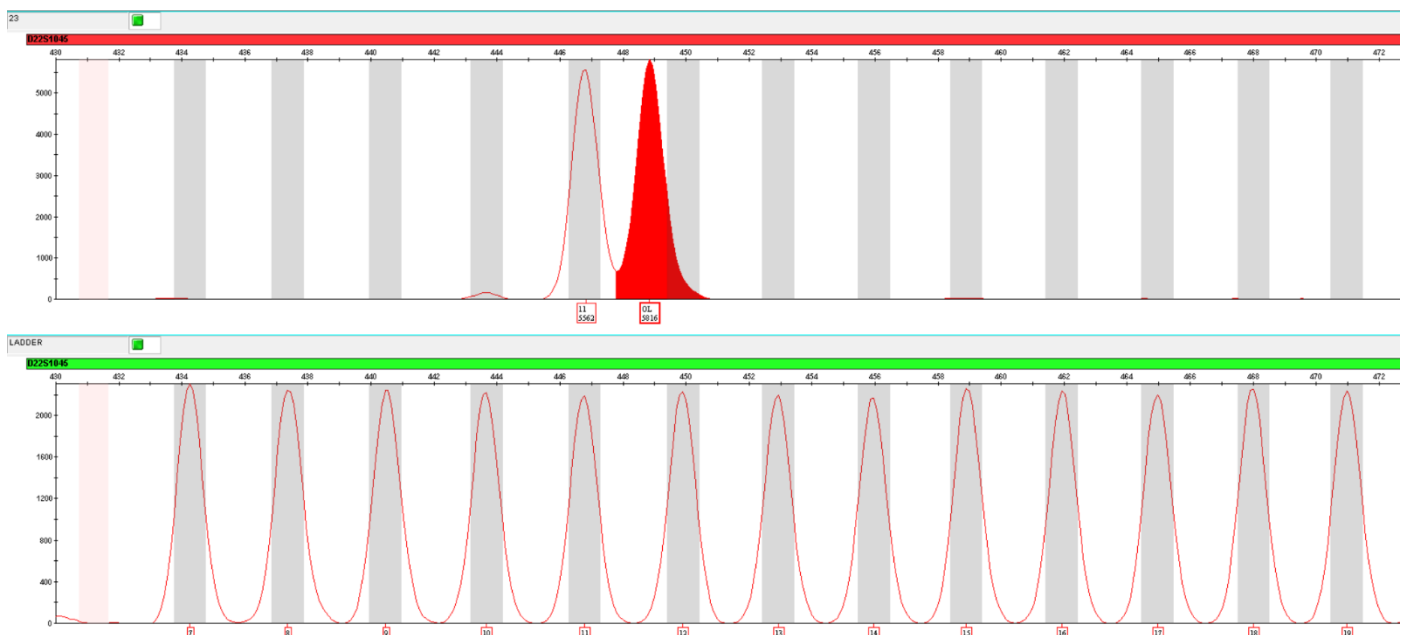


Figure 4.5. D22S1045 off-ladder allele.

One off-ladder (OL) allele observed at the D22S1045 locus highlighted in red, compared to the allelic ladder of the associated locus in the bottom panel.

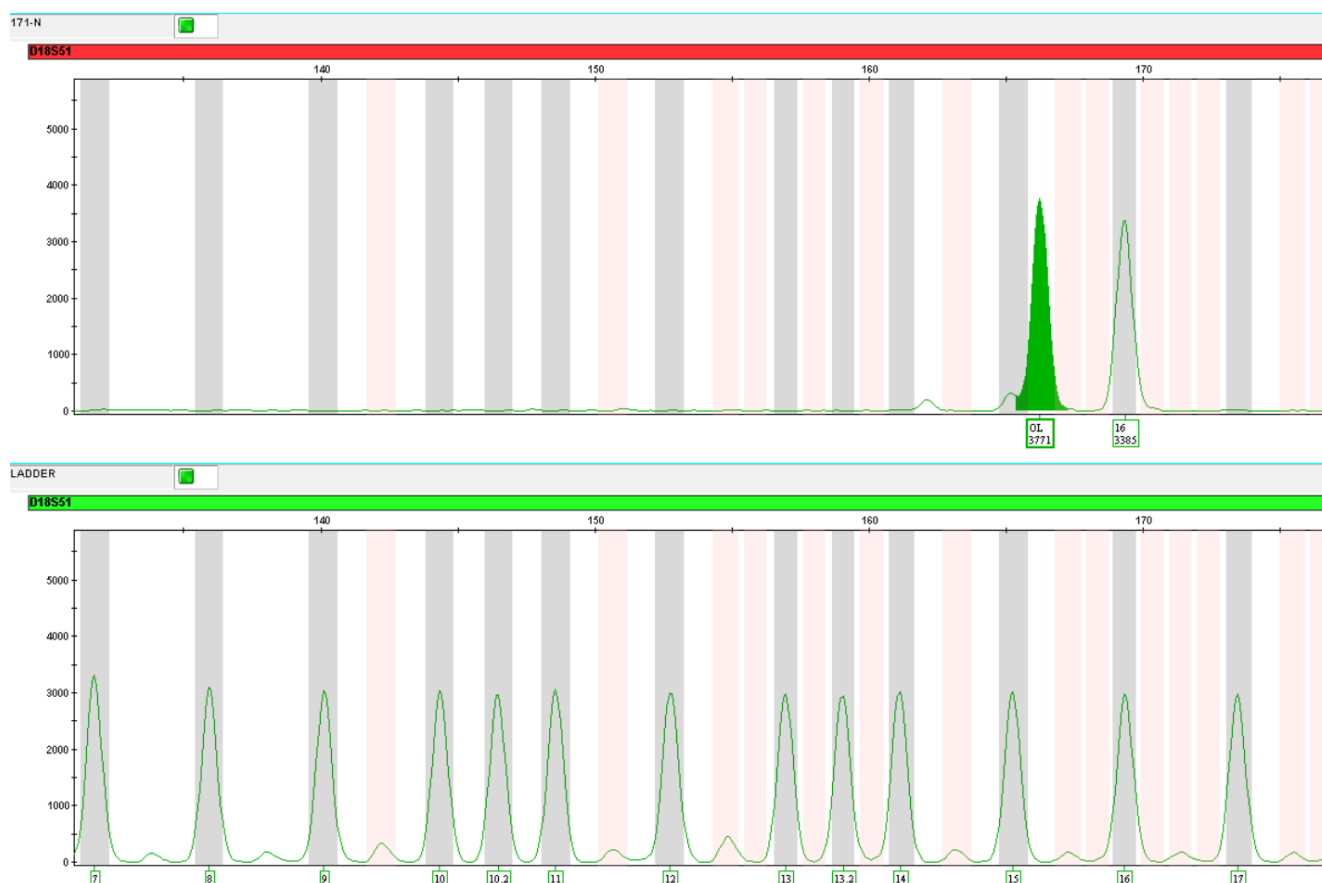


Figure 4.6. D18S51 off-ladder allele.

One off-ladder (OL) allele observed at the D18S51 locus highlighted in red, compared to the allelic ladder of the associated locus in the bottom panel. The OL falls between the alleles 15 and 15.2, thus it was assigned as 15.1.

Table 4.1. Off-ladder alleles identified using the Verogen ForenSeq™ DNA Signature Prep kit.

Locus	Repeat Structure	Identified allele
PentaE	AAGA (AAAGA)15	15.4
	(AAAGA)7 AAAA (AAAGA)7	14.4
	AAAGA AA (AAAGA)9	10.2
	AAGA (AAAGA)18	18.4
PentaD	AAAAA AAAG (AAAGA)9	10.4
D22S1045	(ATT)4 AT (ATT)4 ACT (ATT)2	11.2
SE33*	CT (CTTT)3 C (CTTT)3 C (CTTT)11 (CTTT)3 CT (CTTT)2	13.3
	CT (CTTT)3 C (CTTT)9 T (CTTT)13 CT (CTTT)3 CT (CTTT)2	22.1
	CT (CTTT) C (CTTT)3 C (CTTT)15 (CTTT)2 CT (CTTT)2	17.3

*SE33 length allele designation according to Borsuk et al [369].

4.4.3 Allele frequencies and forensic statistics

Full PowerPlex® Fusion 6C STR profiles were recovered from blood samples taken from 400 Kuwaiti individuals. **Table 4.2** shows the allele frequencies and forensic parameters calculated

for these samples. SE33 was the most discriminative locus in the Kuwaiti population, having 45 different alleles (PIC = 0.945), whereas TPOX was the least discriminative (PIC = 0.616). The calculated combined match probability (CMP) was 7.37×10^{-30} , meaning that the probability of observing two identical profiles for the 23 autosomal loci in the Kuwaiti population was 1 in 1.36×10^{29} . The TPI ranged between 1.439 (TPOX) and 8.333 (SE33), and the combined PE was > 99.9999%. These high values indicate the usefulness of the PowerPlex® Fusion 6C kit for both human identification and paternity testing in the Kuwaiti population.

Table 4.2. Allele frequencies and forensic statistics among 400 Kuwaiti individuals typed at 23 autosomal STR loci in the PowerPlex® Fusion 6C kit.

Allele	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D1S1656	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D5S818	D7S820	D8S1179	FGA	PentaE	PentaD	TH01	TPOX	vWA	SE33
2.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.025	-	-	-	-
3.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.070	0.001	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	0.279	0.003	-	-
6.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
7	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	0.020	-	-	0.116	0.009	0.224	-	-	-
8	0.003	-	-	0.119	0.043	-	-	-	-	-	-	0.001	-	0.016	0.170	0.009	-	0.041	0.023	0.121	0.510	-	-
9	0.024	0.006	-	0.043	0.168	-	0.001	0.001	-	-	-	-	-	0.063	0.093	0.009	-	0.014	0.214	0.228	0.149	-	-
9.1	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	0.001	-	-	-	-	-	-	-	-
9.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.123	-	-	-
10	0.290	-	-	0.059	0.088	0.008	0.001	0.006	-	0.009	-	0.140	-	0.103	0.286	0.078	-	0.059	0.183	0.025	0.098	-	-
10.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-
10.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-
10.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-
11	0.264	0.016	-	0.310	0.313	0.024	0.008	0.098	-	0.159	-	0.395	-	0.279	0.255	0.085	-	0.115	0.190	-	0.218	-	0.003
11.2	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-
11.3	-	-	-	-	-	-	-	-	-	-	-	0.070	-	-	-	-	-	-	-	-	-	-	-
12	0.348	0.026	-	0.328	0.249	0.115	0.093	0.114	-	0.014	-	0.055	-	0.333	0.151	0.120	-	0.164	0.118	-	0.021	-	0.006
12.2	-	-	-	-	-	-	0.010	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003
12.3	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	-
13	0.061	0.210	-	0.106	0.120	0.170	0.249	0.094	-	0.010	-	0.025	0.005	0.200	0.023	0.253	-	0.100	0.125	-	0.001	0.004	0.023
13.2	-	-	-	-	-	-	0.030	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
13.3	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	0.001
14	0.010	0.345	-	0.035	0.019	0.144	0.235	0.123	-	0.076	0.004	0.281	0.061	0.008	0.001	0.189	-	0.046	0.069	-	0.001	0.083	0.041
14.2	-	-	-	-	-	-	0.049	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
14.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-
15	-	0.224	0.018	0.001	0.003	0.148	0.126	0.139	-	0.436	0.003	0.021	0.234	-	-	0.191	-	0.071	0.035	-	-	0.123	0.031
15.1	-	-	-	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15.2	-	-	-	-	-	0.001	0.084	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
15.3	-	-	-	-	-	-	-	0.029	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-
16	-	0.134	0.015	-	-	0.133	0.061	0.226	-	0.248	0.041	0.005	0.266	-	-	0.050	-	0.048	0.004	-	-	0.245	0.060
16.2	-	-	-	-	-	0.001	0.040	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16.3	-	-	-	-	-	-	-	0.043	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
16.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-
17	-	0.030	0.128	-	-	0.100	0.010	0.058	-	0.043	0.189	-	0.286	-	-	0.015	0.056	-	-	-	0.249	0.073	-
17.2	-	-	-	-	-	0.001	0.004	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-
17.3	-	-	0.004	-	-	-	-	0.059	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003
17.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-
18	-	0.008	0.194	-	-	0.085	-	0.001	-	0.005	0.120	-	0.140	-	-	0.003	0.008	0.033	0.001	-	-	0.205	0.089
18.3	-	-	0.019	-	-	-	-	0.009	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-
19	-	0.001	0.126	-	-	0.046	-	-	-	-	0.119	-	0.006	-	-	-	0.060	0.025	-	-	-	0.081	0.088
19.1	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	0.001
19.3	-	-	0.010	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
20	-	-	0.095	-	-	0.010	-	-	-	-	0.156	-	0.001	-	-	-	0.086	0.025	-	-	-	0.011	0.044

Table 1 (continued)

Allele	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D1S1656	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D5S818	D7S820	D8S1179	FGA	PentaE	PentaD	TH01	TPOX	vWA	SE33
20.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003
21	-	-	0.103	-	-	0.008	-	-	-	-	0.048	-	-	-	-	-	0.143	0.005	-	-	-	-	0.016
21.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	0.020
22	-	-	0.105	-	-	0.003	-	-	-	-	0.050	-	-	-	-	-	0.171	0.003	-	-	-	-	0.008
22.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.004
22.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	0.014
23	-	-	0.114	-	-	0.004	-	-	-	-	0.115	-	-	-	-	-	0.179	0.001	-	-	-	-	-
23.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	0.033
24	-	-	0.039	-	-	-	-	-	-	-	0.085	-	-	-	-	-	0.170	-	-	-	-	-	0.001
24.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	0.028
25	-	-	0.024	-	-	-	-	-	-	-	0.054	-	-	-	-	-	0.118	-	-	-	-	-	-
25.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	0.026
26	-	-	0.005	-	-	-	-	-	-	-	0.011	-	-	-	-	-	0.040	-	-	-	-	-	-
26.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.059
27	-	-	0.003	-	-	-	-	-	0.020	-	0.006	-	-	-	-	-	0.008	-	-	-	-	-	-
27.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.074
28	-	-	-	-	-	-	-	-	0.144	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-
28.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.058
29	-	-	-	-	-	-	-	-	0.230	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-
29.2	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053
30	-	-	-	-	-	-	-	-	0.224	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-
30.2	-	-	-	-	-	-	-	-	0.019	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053
31	-	-	-	-	-	-	-	-	0.051	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
31.2	-	-	-	-	-	-	-	-	0.114	-	-	-	-	-	-	-	-	-	-	-	-	-	0.036
32	-	-	-	-	-	-	-	-	0.006	-	-	-	-	-	-	-	-	-	-	-	-	-	-
32.2	-	-	-	-	-	-	-	-	0.119	-	-	-	-	-	-	-	-	-	-	-	-	-	0.019
33	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	-	-	-	0.006
33.2	-	-	-	-	-	-	-	-	0.054	-	-	-	-	-	-	-	-	-	-	-	-	-	0.005
34	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	-	-	-	0.006
34.2	-	-	-	-	-	-	-	-	0.009	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
35	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	0.003
35.2	-	-	-	-	-	-	-	-	0.003	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	0.001	-	-	-	-	-	-	-	-	-	-	-	-	-	0.004
38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.001
CSF1PO	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800	800
Alleles(n) ^a	8	10	17	8	8	18	15	15	17	10	14	12	8	7	9	11	20	24	16	7	8	8	45
Alleles(v) ^b	0.725	0.763	0.870	0.753	0.788	0.860	0.823	0.895	0.853	0.705	0.848	0.720	0.780	0.728	0.775	0.835	0.838	0.900	0.838	0.780	0.653	0.808	0.940
Ho ^c	0.722	0.768	0.885	0.766	0.789	0.879	0.844	0.875	0.844	0.716	0.883	0.737	0.770	0.758	0.793	0.835	0.864	0.912	0.849	0.791	0.661	0.809	0.949
He ^d	0.919	0.052	0.599	0.719	0.084	0.108	0.957	0.380	0.478	0.558	0.559	0.578	0.507	0.125	0.136	0.077	0.366	0.359	0.484	0.593	0.339	0.961	0.778
HWE ^e	0.130	0.094	0.026	0.090	0.078	0.029	0.042	0.030	0.046	0.122	0.026	0.109	0.097	0.098	0.076	0.055	0.035	0.016	0.041	0.077	0.161	0.067	0.007
RMP ^f	0.870	0.906	0.974	0.910	0.922	0.971	0.958	0.970	0.954	0.878	0.974	0.891	0.903	0.902	0.924	0.945	0.965	0.984	0.959	0.923	0.839	0.933	0.993
DP ^g	0.468	0.531	0.735	0.514	0.576	0.715	0.641	0.785	0.700	0.436	0.690	0.460	0.562	0.472	0.553	0.666	0.670	0.795	0.670	0.562	0.359	0.613	0.878
PE ^h	1.818	2.105	3.846	2.020	2.353	3.571	2.817	4.762	3.390	1.695	3.279	1.786	2.273	1.835	2.222	3.030	3.077	5.000	3.077	2.273	1.439	2.597	8.333
TPI ⁱ	0.670	0.732	0.873	0.731	0.758	0.865	0.825	0.861	0.824	0.675	0.870	0.698	0.731	0.718	0.761	0.813	0.848	0.905	0.830	0.758	0.616	0.780	0.945
PI ^j																							

a) number of alleles surveyed, b) number of different allelic variants, c) observed heterozygosity, d) expected heterozygosity, e) Hardy-Weinberg Equilibrium test (p-value), f) random match probability, g) discrimination power, h) power of exclusion, i) typical paternity index, j) polymorphic information content.

4.4.4 Statistical analysis of populations

No significant deviation from the expectations of the Hardy-Weinberg Equilibrium was detected at any locus in the Kuwaiti genotypic data (**Table 4.2**), therefore, the PowerPlex® Fusion 6C autosomal STR alleles are independent and can be used to estimate allele frequencies from their genotype frequencies. Association between alleles at all possible pairwise combinations of loci was evaluated using the linkage disequilibrium test. Significant linkage disequilibrium was detected between 22 (of a total of 253) pairs of loci ($p < 0.05$). However, after Bonferroni correction of the significance level using the number of tests ($0.05/253 = 0.000198$), none of the pairs of loci showed significant linkage disequilibrium, indicating that all loci are statistically independent. Therefore, their allele frequencies can be multiplied together to estimate match probabilities in the Kuwaiti population.

4.4.5 Intra-population genetic structure

Markers that are used for human identification may have weaker discrimination power in populations with genetic structure than in unstructured populations, due to the impact that the presence of subpopulation groups has on the random match probability. This is due to the fact that individuals coming from the same subpopulation groups tend to possess similar alleles, which means the likelihood of seeing random individuals possessing similar genotypes would increase in the presence of genetic structure [347]. Despite the fact that, in this study, no significant deviation from the expectations of the Hardy-Weinberg Equilibrium was detected between the markers, indicating that there is no genetic stratification, it is useful to assess the markers to see if they reveal any genetic clusters within the data. To achieve this, PCA was carried out on the DNA profiles obtained from the Kuwaiti samples for the 23 autosomal PowerPlex® Fusion 6C STR markers. PCA is an unsupervised clustering method that does not require any prior information about the ancestral origin of the samples. Simply, it clusters the samples based on their similarities to each other, forming homogenous clusters of individuals that can be seen on a PCA plot. As expected, the PCA plot (**Figure 4.7**), did not show any pattern of segregation that could be related to the ancestral population of origin of the individuals in the data, indicating that there is no genetic structure within the sample.

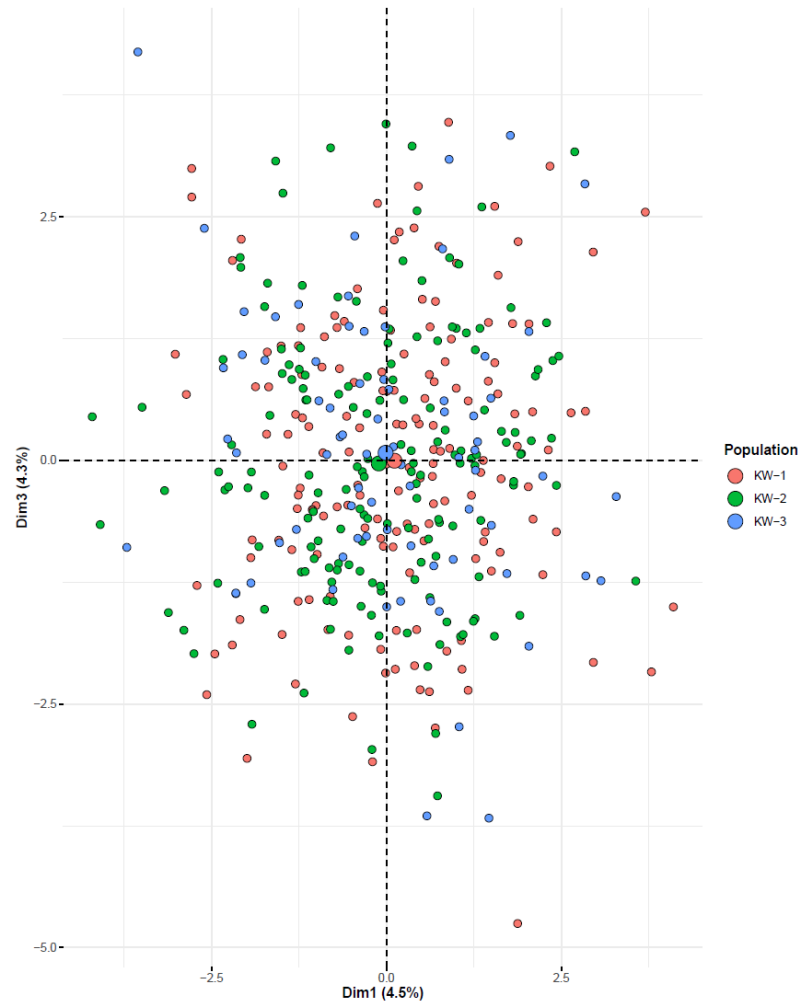


Figure 4.7. Principal component analysis (PCA) of three Kuwaiti subpopulations using 23 autosomal STR markers.

PCA plot showing samples from 400 individuals from the Kuwaiti population, typed for the 23 autosomal STR markers in the PowerPlex® Fusion 6C kit. The samples were colour-coded based on their subpopulation of origin KW-1 to KW-3. The size of the point represents the number of samples.

Another widely used method to infer population structure in genetic data is the Bayesian-based model implemented in the *STRUCTURE* software, which calculates how likely each individual in the data is to belong to each of a number of K (predetermined by the user) populations, and then uses this information to assign individuals into population subgroups [316]. The analysis was run without population information, and the mean log likelihood across five repeated runs of the analysis for each value of K (from 2 to 10) was estimated. The results showed inconsistency in estimating the log likelihood at $K=5$ and over, which is indicated by the high standard deviation (SD), as seen in **Figure 4.8A**. Based on the method

described in Evanno et al. [316], the most likely inferred value of K was 7, as this is the number of populations at which the highest Delta K value was recorded (**Figure 4.8B**).

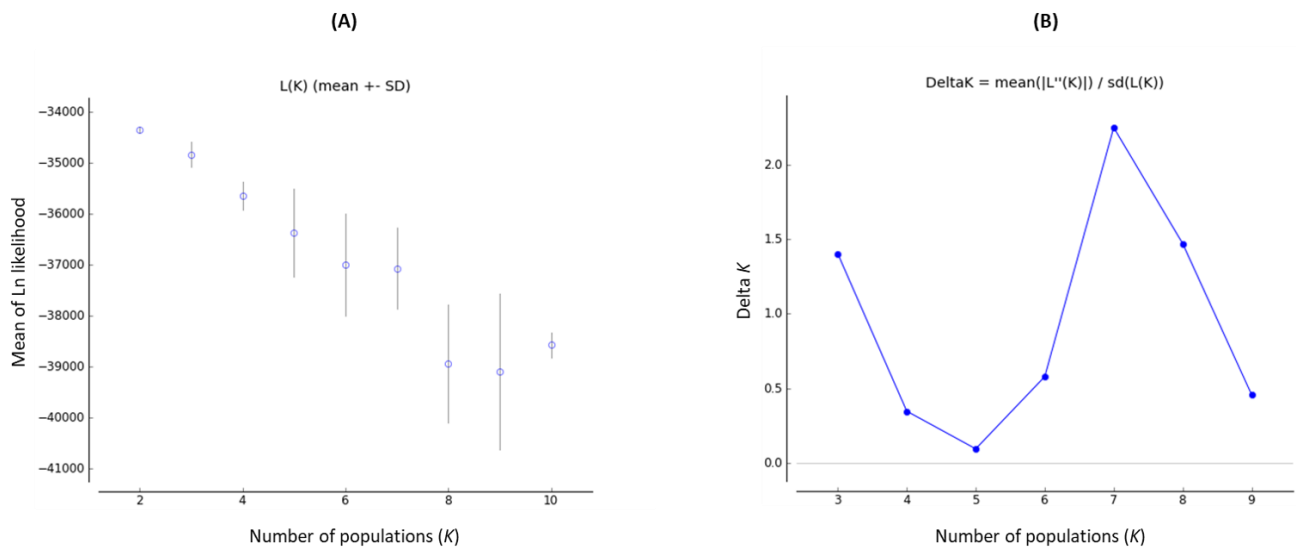


Figure 4.8. Mean likelihood and Delta K values generated from *STRUCTURE*.

Estimated log likelihood plots of *STRUCTURE* results. (A) Mean log likelihood ($\pm$ SD) of the data at each of K populations ($\ln P(X|K)$), based on five repeat runs at each value of K . (B) Rate of change (Delta K) of the log likelihood between each value of K .

However, whilst the results indicated that the data is most probable at $K=7$, there was no clear genetic differentiation between individuals in the sample. This can be seen in **Figure 4.9**, which shows that there is no clear signal of structuring between the three subpopulation groups, in terms of the proportion of each individual's genetic ancestry that is assigned to each of the K populations, regardless of the number of populations assumed. That is, all of the three predefined Kuwaiti subpopulations (KW-1 to KW-3) has approximately equal proportions of ancestry assigned into each of the K populations, indicating that there is no genetic structure in the data. This is further supported by the relatively small increases in mean log likelihood and Delta K values from $K=2$ to $K=3$, suggesting that there is limited evidence for any genetic structuring within the Kuwaiti population sample, in agreement with the results of the PCA analysis above.

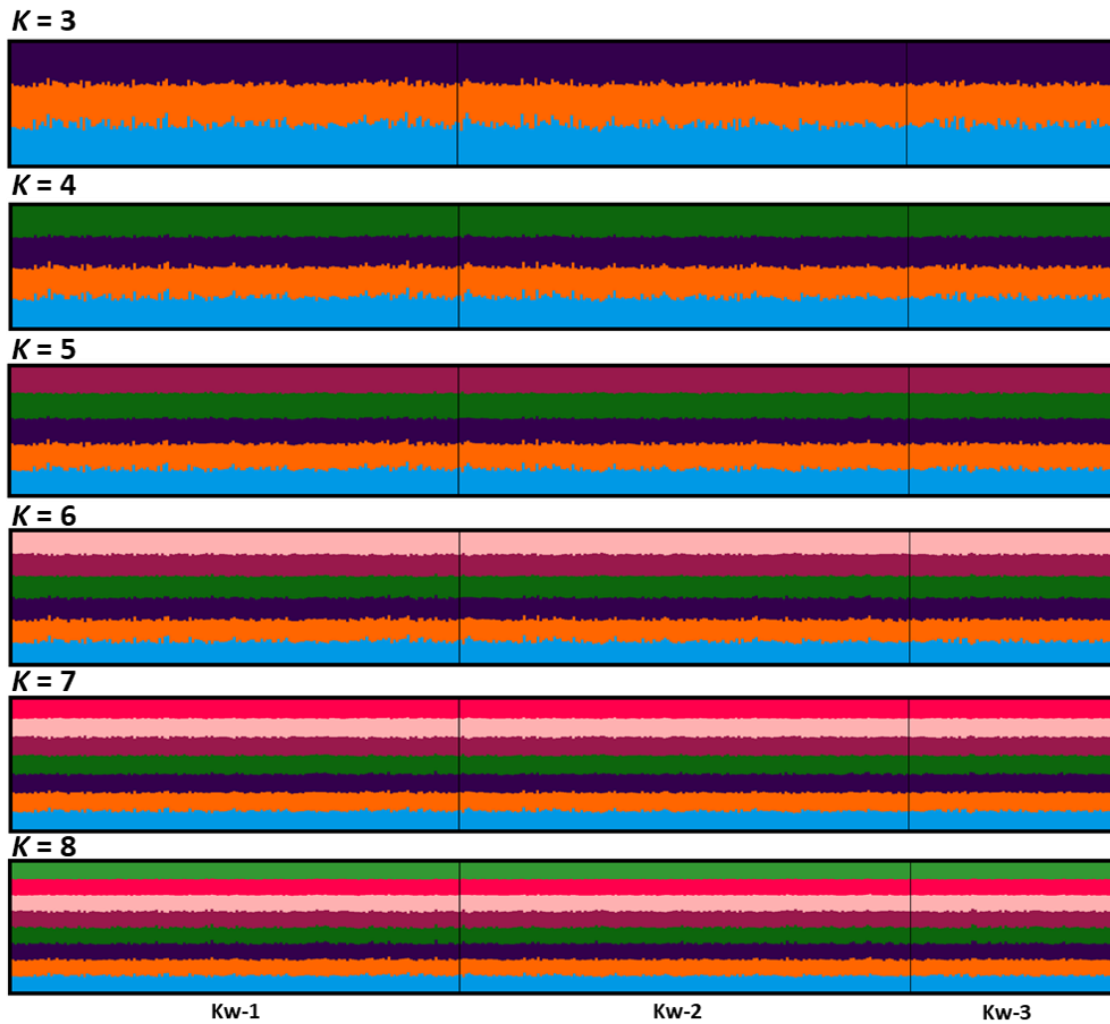


Figure 4.9. Analysis of the structure of the Kuwaiti population using 23 STR markers.

Assessing population stratification in the Kuwaiti population using the model-based approach implemented in *STRUCTURE*. The number of colours in each bar represents the number of assumed populations from $K=3$ to $K=8$ from, each vertical line represents an individual, and the amount of each colour in each line represents the proportion of that individual's genetic ancestry that has been assigned to each of the K populations.

4.4.6 Genetic distance

To investigate the genetic distance between the Kuwaiti population and other global populations, allele frequencies for the 23 autosomal STRs in the PowerPlex® Fusion 6C kit were pooled from the HGDP-CEPH global panel, which contains 57 populations grouped into seven global regions, and consists of eight African ($N = 507$), six American ($N = 551$), nine Central and South Asian ($N = 202$), four Middle Eastern ($N = 160$), 11 European ($N = 2135$), and 17 East Asian ($N = 227$) populations. An Indian population ($N = 374$) and a Lebanese population ($N=505$) have also been typed for the 23 loci, thus were added to the analysis. Both PCA and phylogenetic analyses were carried out, and the resulting plots characterise the

genetic differentiation between populations. The distribution of the populations on the PCA plot (**Figure 4.10**), and the genetic distances between them on the NJ tree (**Figure 4.12A**) show that the Kuwaiti population is genetically close to the Lebanese and Middle Eastern groups, which includes Mozabite, Druze, Palestinian and Bedouin populations. This is likely to be explained by the gene flow between these geographically close locations, which consequently leads to more similar allele frequency distributions among them.

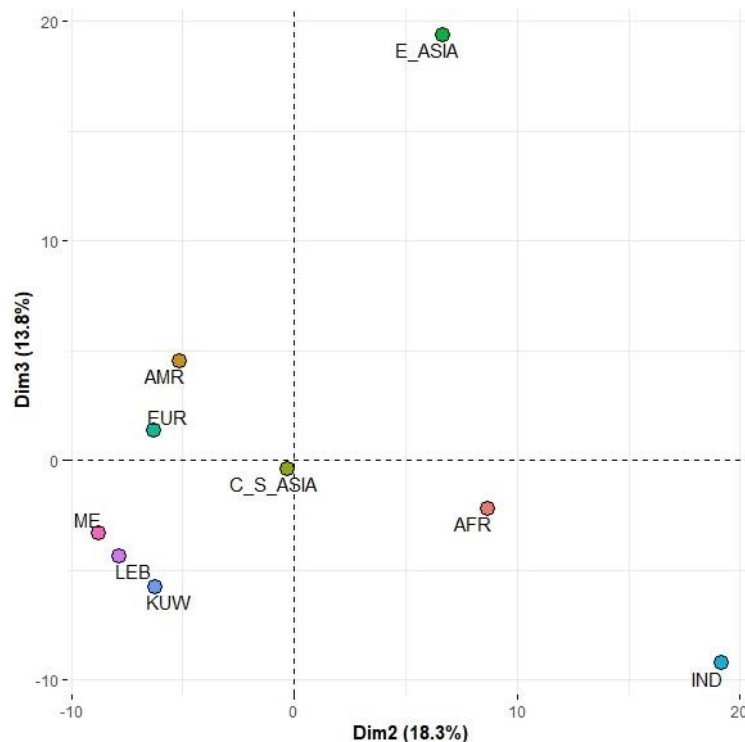


Figure 4.10. Principal component analysis of nine populations genotyped for 23 STR markers.

PCA of genetic distance based on allele frequencies of 23 autosomal PowerPlex® Fusion 6C STR loci shared between the population from Kuwait (KUW) and those from Africa (AFR), America (AMR), Central and South Asia (C_S_ASIA), the Middle East (ME), Europe (EUR), East Asia (E_ASIA), Lebanon (LEB) and India (IND; Madhya Pradesh).

At the regional level, genetic distance was assessed based on the 13 loci shared between this study and studies examining other populations from the Arabian Peninsula. The resulting PCA plot (**Figure 4.11**), and NJ tree (**Figure 4.12B**) show that the Kuwaiti dataset broadly clustered with the previously published Kuwaiti data, and was genetically closer to the countries in the north of the region such as Iraq and Iran. In contrast, there was a higher level of genetic differentiation between Kuwait and Saudi Arabia, Yemen and Qatar, which were clustered in

the upper-right part of the PCA plot, and Bahrain, Oman and UAE, which were clustered in the lower part of the plot.

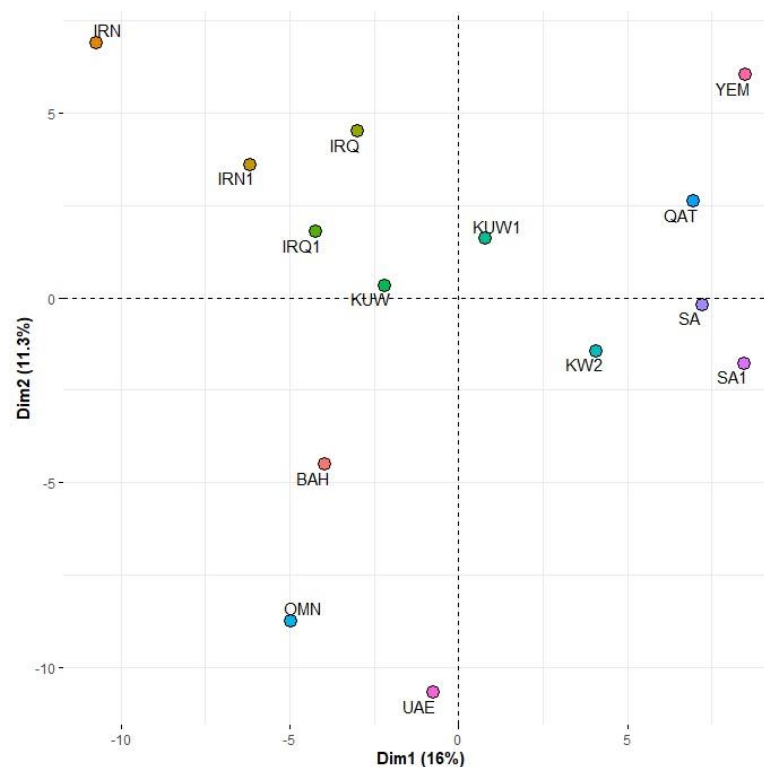


Figure 4.11. Principal component analysis of 13 populations using 13 autosomal STR markers.

Principal component analysis of genetic distance based on allele frequencies of 13 autosomal PowerPlex® Fusion 6C loci (CSF1PO, D13S317, D16S539, D18S51, D21S11, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA) shared between the dataset reported here (KUW), previously published Kuwaiti studies (KUW1 and KUW2), and studies of the populations of neighbouring countries Iraq (IRQ and IRQ1), Iran (IRN and IRN1), Saudi Arabia (SA and SA1), United Arab Emirates (UAE), Oman (OMN), Yemen (YEM), and Bahrain (BAH).

In this study, 30% of individuals declared their origins as being from the north (Iraq and Iran), 39% from the south region (Saudi Arabia, Bedouin and Bahrain), and 24% had parents of different origin (admixed). Therefore, the closer genetic relationship of the samples in this study to the northern region might be due to the presence of these individuals. Overall, it can be seen that based on the differential allele frequencies of the 23 autosomal loci in different populations, both distant and closely related populations were genetically separated from each other, as illustrated in the PCA plots in **Figure 4.10** and **Figure 4.11**.

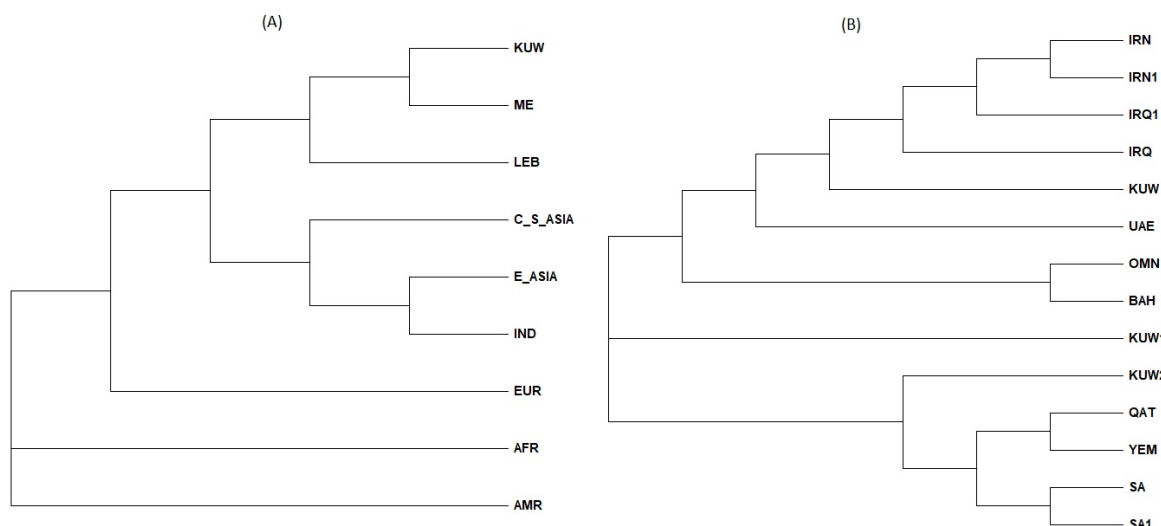


Figure 4.12. Neighbour joining (NJ) tree based on autosomal STR markers.

Neighbour-joining tree based on pairwise D_A values from allele frequencies of (A) the 23 autosomal PowerPlex® Fusion 6C STR loci in Kuwait and populations from Africa (AFR), America (AMR), Central and South Asia (C_S_ASIA), the Middle East (ME), Europe (EUR), East Asia (E_ASIA), Lebanon (LEB) and India (IND; Madhya Pradesh), and (B) the 13 autosomal STR loci shared between the dataset reported here (KUW), previously published Kuwaiti studies (KUW1 and KUW2), and studies of the populations of neighbouring countries Iraq (IRQ and IRQ1), Iran (IRN and IRN1), Saudi Arabia (SA and SA1), United Arab Emirates (UAE), Oman (OMN), Yemen (YEM), and Bahrain (BAH). The nodes represent a common ancestor, and the branching tips are descendants of that common ancestry.

4.5 Conclusion

This chapter evaluated the forensic utility of the 23 autosomal STR loci included in the Promega PowerPlex® Fusion 6C™ kit for the Kuwaiti population. Among these loci, D10S1248, D22S1045, D2S441 and SE33 are reported for the first time for Kuwait. The genetic data indicate that these 23 autosomal STRs are highly polymorphic in the Kuwaiti population and are of high value for human identification and paternity testing. STRUCTURE and PCA analysis show no signature of genetic structuring of the Kuwaiti population into subpopulations. Comparison of the Kuwaiti population to other global populations indicates that Kuwait clusters with other Middle Eastern populations, and shows a close relationship with Iran and Iraq, suggesting that they may share common ancestry.

Chapter 5 Massively parallel sequencing of short tandem repeat (STR) markers in the Kuwaiti population

5.1 Introduction

The last decade has seen the introduction of several technologies for massively parallel sequencing (MPS), also termed 'next-generation sequencing' (NGS), which out-perform the traditional Sanger sequencing method in terms of cost and throughput. MPS technologies work in a massively parallel manner to generate millions of DNA sequences (reads) [246], and are now used routinely to sequence whole human genomes as well as targeting regions to permit sequencing of a subset of the whole genome. Targeted MPS technologies are capable of analysing large numbers of markers in a single run, including single nucleotide polymorphisms (SNPs), short tandem repeats (STRs), insertion and deletion polymorphisms (indels), structural variants (SVs) and copy number variants (CNVs) [236].

More recently, MPS technology has been developed for use in forensic genetics and offers a number of benefits over the current polymerase chain reaction (PCR)-capillary electrophoresis (CE) typing system for forensic STR and SNP markers. The main advantage of MPS is that it defines the nucleotide sequences of targeted amplicons, which include any sequence variation within a STR repeat region as well as variants occurring in the regions flanking the repeat array. In addition, MPS has overcome the limited multiplexing capacity of CE by allowing high numbers of different classes of marker to be combined into a single assay. This in turn saves both time and the often-limited DNA material when additional DNA investigations are required, as well as augmenting the discrimination power of forensic testing. When using PCR-CE typing of STRs, partial profiles are often observed, especially in degraded or old samples, due to a failure to amplify large STR fragments. MPS offers an improved system in which the amplicons for different classes of markers are designed to be as short as possible, to improve information recovery from challenging samples.

Verogen, Inc. (CA, USA) has recently focused on forensic genomics and launched the MiSeq FGx®, a massively parallel sequencing system utilising Illumina's (CA, USA) sequencing technology for forensic applications [247]. The key kit that the FGx® system runs is the ForenSeq™ DNA Signature Prep Kit, which contains two primer sets: primer mix (A) can amplify 153 markers including the amelogenin sex chromosome marker, 27 autosomal STRs,

24 Y-STRs, 7 X-STRs and 94 identity SNPs. Primer mix (B) contains all the markers in the primer mix (A), in addition to 22 phenotypic SNPs and 56 biogeographical ancestry SNPs. An additional five markers (SE33, DXS8377, DXS10148, DYS456, and DYS461), are also present in the FASTQ files but are not recovered or reported in the ForenSeq™ Universal Analysis Software (UAS) [267, 369]. A considerable amount of literature on the ForenSeq™ kit has been published, assessing the forensic performance of both the kit and the sequencing system, including tests on sensitivity, reproducibility, concordance and mixture analysis [251-259]. Other studies have focused on the applications of the ForenSeq™ kit in forensic DNA analysis, including its performance on degraded DNA [260, 261], and for kinship testing [262, 263] and ancestry prediction [264, 265].

For this system to be implemented in forensic laboratories, population studies must be carried out to produce allele frequencies that facilitate the calculation of statistics to estimate the strength of DNA profile evidence, as well as to check for population-specific variants and null alleles. To date, several studies have been published testing the performance of the ForenSeq™ markers on different populations [267-275]. These studies have demonstrated that sequence variants detected within both STRs and their flanking regions (indel or SNP variants) provide additional allelic diversity compared to CE-based methods. This increased number of genetic variants detected within forensic marker regions using MPS increases the power of this kit as a forensic DNA test.

The abundance of genetic variants within STRs suggests that there are significant numbers of variants within populations yet to be detected. Population studies using new sequencing technologies will gradually increase the number of novel STR genetic variants and determine their frequencies in different populations. However, the current reported population data are not sufficient to produce allele frequencies for existing known variants, or the new variants discovered by sequencing [266]. Therefore, comprehensive databases of sequencing data from large-scale population studies are needed to characterise STR variants, in order that STR/SNP frequency calculations can be carried out, to permit the assessment of the evidential value of DNA profiles generated using these technologies. To date, no studies have investigated MPS of forensic STR markers in the Kuwaiti population. In this chapter, the forensic utility of the 27 autosomal STR markers included in the ForenSeq™ kit was examined for the Kuwaiti population.

5.2 Aims and objectives

The main aim of this part of the study is to evaluate the forensic performance of 27 STR markers contained within the Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics system, in the Kuwaiti population. This study involves:

- Generating STR profiles by sequencing 376 unrelated Kuwaiti individuals using the ForenSeq™ DNA Signature Prep Kit (primer set A).
- Comparing the results to a CE-based STR genotyping kit, the Promega PowerPlex® Fusion 6C kit (reported in chapter 4) to examine concordance between the two systems.
- Calculating allele frequencies and assess a range of forensic parameters for the sequenced STR makers.
- Cataloguing the underlying genetic variation in the sequenced markers in the Kuwaiti population.
- Identifying and evaluate the performance of the SE33 locus sequences in this population, which were obtained from raw sequencing files using a bioinformatics approach.

5.3 Materials and methods

Sample collection was carried out at the Kuwait Identification DNA Laboratory (KIDL), which is part of the General Department of Criminal Evidence. Sample sequencing was carried out at the Alec Jeffreys Forensic Genomics Unit, a part of the Forensic Science Institute at the University of Leicester. The whole procedure starting from template and library preparation to sequencing was carried out by PhD student Mahdi Haidar at the University of Leicester. All the data was transferred to a portable hard drive and data analysis was carried out at the University of Strathclyde.

5.3.1 Sample collection and preparation

Blood samples were collected on Whatman® FTA® cards (GE Healthcare Life Sciences, IL, USA) from 376 unrelated Kuwaiti individuals (248 males and 130 females). Samples were prepared as described in section **2.8**.

5.3.2 Template and library preparation using two-step PCR

The first PCR was carried out to amplify and tag STR and SNP sequences in the DNA samples using the ForenSeq primer mix A. PCR1 master mix was prepared according to the volumes recommended by the manufacturer, as described in 2.8.1. Target amplification was performed in a GeneAmp PCR System 9700 instrument (Thermo Fisher Scientific, USA) using the PCR1 thermal profile described in section **2.8.1**.

The second PCR was performed to attach sample-specific indexed adapters (barcodes) to the amplified DNA, which allow separation of reads and identification of samples during the data analysis phase. In order to ensure that samples can be accurately separated, each sample should be amplified with a specific barcode. PCR2 master mix was prepared according to the volumes recommended by the manufacturer, and barcodes added to the samples following the procedures described in section **2.8.1**.

5.3.3 Library purification and normalisation

The amplified indexed libraries were purified twice using Sample Purification Beads. The magnetic beads capture the DNA strands to form a magnetic bead-DNA compound. Inhibitors and other biological remnants were eliminated through two washing steps on a magnetic stand using 80% absolute ethanol. This was followed by the addition of Resuspension Buffer to release the DNA molecules from the beads. The purified libraries were then normalised by taking a fixed quantity of magnetic beads, which capture the same amount of DNA from each library. This assures equal amounts of DNA is input into the sequencing reaction for all the libraries while preserving their individual genetic content and ensuring a consistent cluster density. Purification and normalisation were carried out as described in section **2.8.2**.

5.3.4 Library preparation and sequencing

Once normalisation had been carried out, 5 µL of each library was transferred to a single tube, and the pooled libraries were diluted in Hybridization Buffer. This was followed by the addition of Human Sequencing Control (HSC) which is a positive sequencing control for the FGx® instrument into the tubes containing the pooled libraries. Reagent volumes and detailed procedures are described in section **2.8.3**. Four sequencing runs were carried out, each of which included 94 samples, a positive control (2800M) and a negative control (nuclease-free water) to test for contamination.

5.3.5 Data analysis

Sequencing data were initially analysed using the graphical interface program, ForenSeq™ Universal Analysis software (UAS) version 1.3. The default analysis setting of 1.5% for the analytical threshold (AT) and 4.5% for the interpretation threshold (IT) were used for data interpretation. The values of the AT and IT are calculated by multiplying the threshold percentage by 650, which is the default minimum number of sequencing reads for each autosomal locus in the UAS. Thus, the default AT is 10 reads ($1.5\% \times 650$ reads), meaning that alleles with fewer than 10 reads are not recognised using the default software settings, and the default IT is 30 reads. The default stutter filter percentages are shown in **Appendix 3**.

Quality metrics, and the positive and negative controls, were all checked after the completion of each run. Allele/genotype calls were viewed through the UAS graphical interface, which also includes quality control flags (indicators) for each locus. Manual Interpretation/editing was carried out when, for example, a flag showed a low coverage allele at a locus or when a user action was required to designate a stutter (when a stutter had a higher number of reads than the recommended percentage) from parent alleles. A detailed genotype report was generated for each sample containing the STR allele/genotype calls, the number of reads, STR core repeated nucleotide sequence and quality control indicators for all markers. All the reports were manually reviewed and then exported to a single Excel file prior to carrying out statistical analysis. In addition, the flanking sequence report option, implemented in the UAS, was used to generate the flanking region sequences for all samples.

5.3.6 Bioinformatics analysis

FASTQ files were further analysed using STRait Razor version 3.0 [306] by mapping to the GRCh38 reference genome. This was carried out to examine the bioinformatic concordance of allele designation between length- and sequence-based alleles and to determine the nomenclature for the sequenced alleles according to ISFG recommendations [266]. The nomenclature system in the database used by STRait Razor is based on the data published by Novroski et al. [267]. Therefore, alleles which could not be identified in the database were designated as 'NovelSeq'.

Novel alleles (NovelSeq) were manually given the appropriate nomenclature using the Forensic STR Sequence Structure Guide [370]. The guide contains the forward strand

reference sequence strings with the flanking regions of 35 autosomal STR markers as well as the most commonly occurring variants found within the flanking regions. In samples that contained a variant not included in the guide, FASTQ files were manually aligned to the GRCh38 genome using BWA (Burrows-Wheeler Aligner) version 0.7.15-r1140 and the resulting bam files were visualised using IGV (Integrative Genomics Viewer) version 2.8.0. This provides a view of the variant location, which was then identified using Variation Viewer 2.1.4 [371] and given the appropriate variant code (rs-number), if available.

5.3.7 SE33 locus analysis

SE33 sequence data were extracted from the FASTQ files using STRait Razor with a modification in its configuration file as follows: 5' anchor (TATAGTAACTTG), 3' anchor (CTTTTCTTTCTTTTCTTC) and motif (CTTT) [369]. Sequences based on >10 reads were generated automatically, and to capture more data, those with <10 reads were recovered manually. Nomenclatures for the sequenced alleles were assigned manually according to the SE33 allele-classification system reported by Borsuk et al. [369].

5.3.8 Concordance test

Comparison between CE and MPS systems was carried out by comparing the genotypes generated using the Promega PowerPlex® Fusion 6C kit (reported in **Chapter 4**) with the genotypes based on the allele lengths in the sequence data.

5.3.9 Forensic statistical analysis

Arlequin version 3.5 was used to calculate population genetic parameters including allele frequencies, and observed and expected heterozygosities, as well as to test for departure from Hardy-Weinberg Equilibrium (HWE) and for evidence of linkage disequilibrium (LD) [309]. Forensic parameters including the random match probability (RMP), discrimination power (PD), power of exclusion (PE), typical paternity index (TPI) and polymorphic information content (PIC) were calculated using STRAF, an online tool for STR data analysis [310].

5.4 Results and discussion

5.4.1 Sequencing quality

The MiSeq sequencing run quality is evaluated by four metrics: cluster density (K/mm²) i.e., the number of thousand (K) clusters per square millimetre (mm²), and the percentage of clusters passing filter, phasing, and pre-phasing. Quality metrics from the four runs performed for this project are shown in **Appendix 4**. The 'cluster density' measure is the number of clusters generated per square millimetre. All the runs were in the recommended range of 400-1650 K/mm².

The 'clusters passing filter' measure assesses the quality of the reads. The filter detects high quality base calls based on their signal intensity and eliminates low quality reads. From the four runs, more than 94% of the overall clusters passed the filter, meaning that high quality reads were generated from all the runs. The 'phasing' and 'pre-phasing' metrics rate the quality of the sequencing chemistry. These metrics indicate the rate of bases that lose synchronisation with each other within a cluster. Phasing is the proportion of bases that fall behind a cluster, whereas pre-phasing indicates the proportion of bases that go ahead of a cluster, when DNA strands extend in a cluster during sequencing. Three of the four sequencing quality metrics ('cluster density', 'clusters passing filter', and 'phasing') generated from the MiSeq FGx[®] were within the recommended range. However, the fourth metric ('pre-phasing') was higher than the recommend value of 0.15%, in all the runs performed. This slightly elevated pre-phasing has no effect on the quality of the data produced and results were sufficient for analysis (Richard Kessell, Verogen, personal communication).

5.4.2 Sequencing performance

Sequencing depth, also called depth of coverage (DoC), is the number of high quality aligned reads generated from a target genomic region. The average DoC for individual STR markers was within the range 79X– 1800X. As shown in **Figure 5.1**, PentaE and D1S1656 yielded the lowest coverage, while TH01 and D20S482 exhibited the highest coverage among all the markers. Variation in DoC among the STR loci was observed in this study in a consistent pattern over the four sequencing runs (**Appendix 5**).

It is common to see variation in DoC in such large multiplexes containing both STR and SNP markers, particularly when compared to the balance seen in CE-typed STRs. This is because, during PCR amplification of over 230 targets, some loci will amplify more efficiently than others and subsequently yield higher DoC, e.g. TH01. Even when the primer balance and cycling conditions are optimised, the DoC will never be entirely even across all loci. However, although low DoC was observed for some individual samples at some loci, the average DoC across all individuals at each locus is high. In addition, a concordance of 100% was observed between the results obtained using MPS and CE for the 23 autosomal markers (see section 5.4.7 below), indicating that overall, the sequencing data are of high quality and the results can be interpreted confidently.

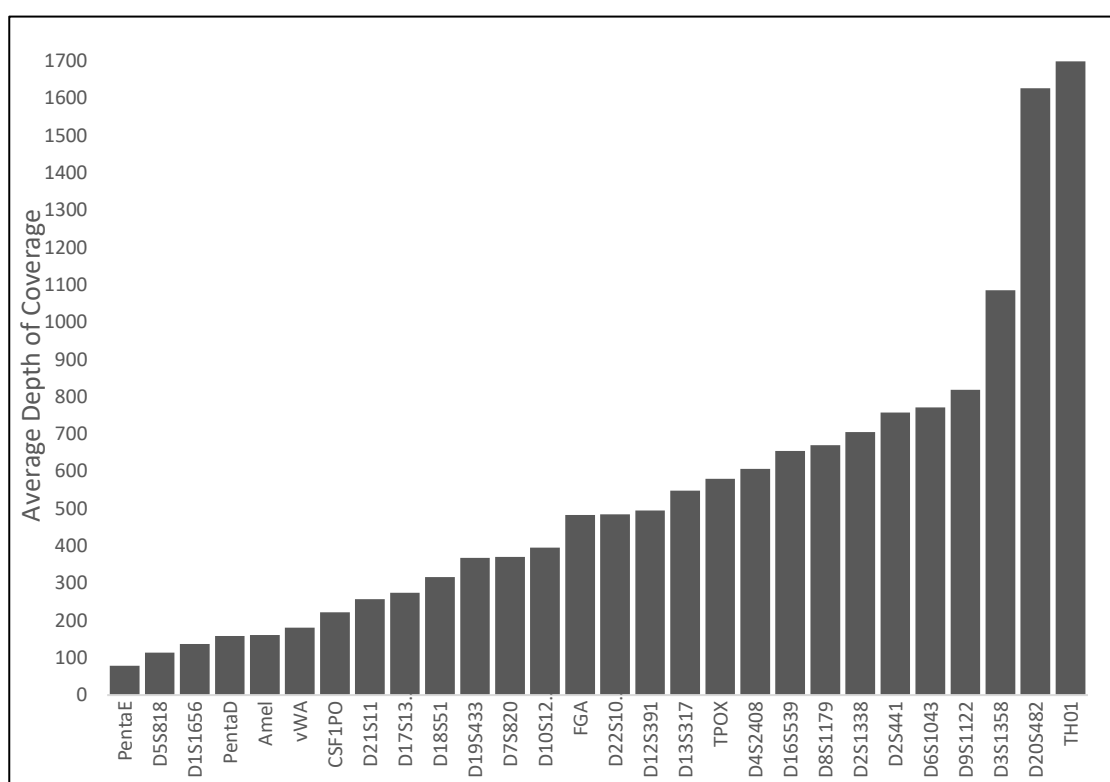


Figure 5.1. Average DoC for the ForenSeq™ STR loci.

Average DoC for autosomal STRs and amelogenin (Amel) across the four runs performed using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® system.

To evaluate the intralocus heterozygous balance, the allele coverage ratio (ACR) was determined for 8227 heterozygote genotypes across the 27 STR loci and amelogenin, by calculating the ratio of the DoC for the two alleles. This was achieved by dividing the number of reads of the lower coverage allele by the number of reads of the higher coverage allele at each heterozygote locus. The average ACRs across markers were in the range 0.6-0.9, except

for D22S1045, which showed an average ACR of 0.36. An ACR of 0.6 is used as the default threshold in the UAS, i.e. an imbalance flag indicator is triggered by the software for any heterozygote genotype with an ACR less than 0.6. It was observed that 336 of the heterozygotes at 15 of the 26 STR loci (excluding D22S1045), had individual ACR values of <0.6. This was reduced from 336 to 128 (a decrease of 62%) when adjusting the ACR threshold from 0.6 to 0.5. Reducing the threshold to 50% is useful, especially for large numbers of reference samples, because for a relatively small increase in the imbalance at a locus, there is a substantial decrease in the number of imbalance flag indicators, reducing time spent interpreting each individual heterozygote balance.

In the case of D22S1045, imbalance (ACR <0.6) was observed at 237 of the 256 heterozygotes (92%). This was expected as Verogen state in the ForenSeq™ Kit manual that imbalance is likely to be detected at this locus. In addition, PentaE, which showed the lowest depth of coverage, had 137 instances of imbalanced alleles (ACR <0.6) in 322 heterozygote (42%) individuals. As shown in **Figure 5.1**, the D5S818 and D1S1656 loci yielded the next lowest DoC values after PentaE, and 44/293 and 39/330 imbalanced heterozygotes were detected at these loci, respectively. In comparison, the markers which yielded high DoC tended also to have no or limited allelic imbalance, suggesting that the imbalance could be related to the DoC for any given sample. This may also be influenced by variation in PCR efficiency, which derives in part from the fact that each amplicon has a different GC content, so some alleles will be amplified more easily than others.

Differential amplification might also be the reason behind the lower DoC at PentaE and D1S1656 where smaller alleles (in terms of the number of repeats) amplified more efficiently, yielding larger number of reads compared to the larger allele. An additional notable observation was that at the D1S1656 locus, 33 of the 39 imbalanced heterozygotes (79% of the total number of observed imbalanced heterozygotes at this marker) had lower coverage at the larger allele, which tended to be a microvariant allele (an allele with a partial repeat unit e.g., 15.3, 16.3, 17.3, 18.3 and 19.3).

The problem associated with this imbalance is that when attempting to interpret mixed DNA profiles (containing DNA from more than one contributor), DoC is a source of information for distinguishing between contributors, so these imbalances may make this interpretation unreliable. However, none of the samples in this study were mixed, and for the purposes of

this study the key requirement was that STR alleles were distinguishable from the noise generated by amplification and sequencing errors.

5.4.3 Sample qualification and genotyping

Each run contained 94 samples, plus a negative control and a positive control (2800M). This resulted in 376 samples, of which 365 were included in this study and 11 were removed due to the presence of extra alleles (likely to be a result of contamination) or low coverage of the samples. The negative controls showed no signs of any product. The positive control was in concordance with the positive control profile provided in the ForenSeq manual (the positive control DNA sample 2800M genotypes are shown in **Figure 5.2**). Full profiles were produced for all but eight of the 365 samples, where missing loci were observed at the PentaE locus.

STRs

● Single Source Sample
● Interlocus Balance
59/59 Loci Typed

Amelogenin	D1S1656	TPOX	D2S441	D2S1338	D3S1358
X Y	12 13	11 11	10 14	22 25	17 18
D4S2408	FGA	D5S818	CSF1PO	D6S1043	D7S820
9 9	20 23	12 12	12 12	12 20	8 11
D8S1179	D9S1122	D10S1248	TH01	vWA	D12S391
14 15	12 12	13 15	6 9.3	16 19	18 23
D13S317	PentaE	D16S539	D17S1301	D18S51	D19S433
9 11	7 14	9 13	11 12	16 18	13 14
D20S482	D21S11	PentaD	D22S1045	DXS10135	DXS8378
14 15	29 31.2	12 13	16 16	28	12
DXS7132	DXS10074	DXS10103	HPRTB	DXS7423	DYS505
13	21	18	12	15	11
DYS570	DYS576	DYS522	DYS481	DYS19	DYS391
17	18	12	22	14	10
DYS635	DYS437	DYS439	DYS389I	DYS389II	DYS438
21	14	12	14	31	9
DYS612	DYS390	DYS643	DYS533	Y-GATA-H4	DYS385a-b
29	24	10	12	11	13 16
DYS460	DYS549	DYS392	DYS448	DYF387S1	
11	13	13	19	37 38	

Figure 5.2. ForenSeq™ positive control profile.

Positive control profile generated from 1 ng of male DNA contained in sample 2800M. The profile shows amelogenin and 58 STRs. The 27 autosomal STRs used in this study are outlined in yellow.

5.4.4 Sequence variation

An overall increase in the number of alleles was observed when sequence variation was taken into account. An increase in allele diversity within the STR repeat region was observed in 20 of the 27 STRs, of which D12S391, D2S1338, and D21S11 had the largest gain of an additional 49, 35 and 30 alleles respectively (**Figure 5.3**). In addition, more allele variants emerged when examining the flanking regions, further increasing the number of alleles at 11 out of the 27 loci. Only three loci (D17S1301, TH01 and TPOX) showed no allele sequence variants. The increase in the number of allele variants in both the STR repeat region and the flanking sequence region will substantially increase the discrimination power of forensic testing in this population. In addition, distinguishing between two individuals with alleles at a locus that are the same in length but with nucleotide sequences that differ will improve mixture interpretation and assist in kinship testing.

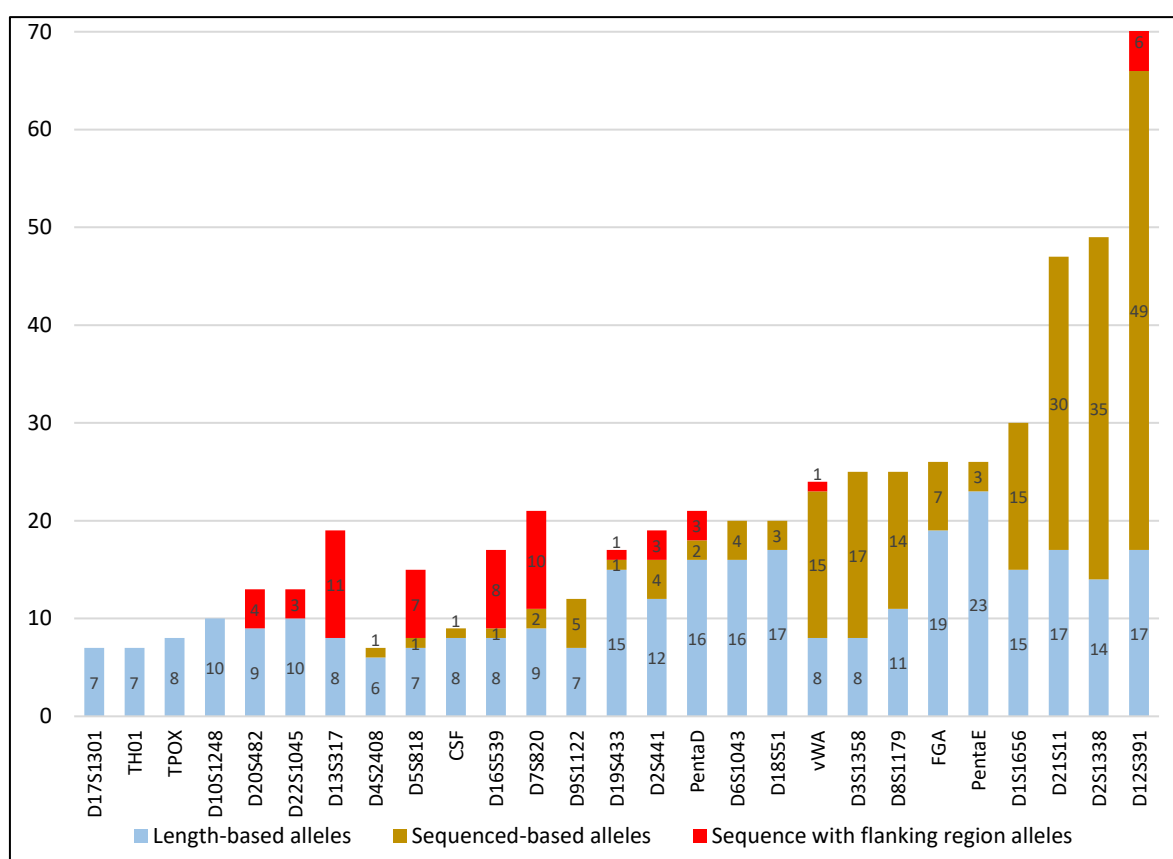


Figure 5.3 Increase in allele number at autosomal STRs as a result of sequence variants.

Number of alleles observed using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® sequencing data in 365 Kuwaiti individuals. The order of loci in the figure is based on the number of additional alleles detected by sequencing the STR repeat region and the flanking region.

5.4.5 Allele frequencies and novelty

Allele frequencies were calculated for the 27 autosomal STR markers in the 365 Kuwaiti individuals, based on three different classifications: alleles based on length variation, alleles based on STR repeat region sequence variation and alleles based on flanking region sequence variation. These frequencies were produced by counting the number times each individual allele is observed at a locus, divided by the number of alleles surveyed at a locus. Frequencies were calculated for length-based alleles and sequence-based alleles obtained from the sequencing data. **Appendix 6** shows frequencies of length-based, repeat sequence-based, and flanking sequence-based alleles in 365 Kuwaiti individuals

A total of 109 unique allele sequences distributed among 22 of the 27 autosomal STRs were identified by the STRait Razor database as novel alleles. However, 66 of these alleles were found in either GenBank STRseq [372], or in the literature [272], and a further six allele sequences were restricted to the Middle East region and observed in one of two previous studies (either [272] or [373]). **Appendix 7** shows the frequencies of all alleles that have previously been observed, and **Table 5.1** shows the 37 novel alleles that have not been reported previously in the literature.

In previous studies the frequencies of variants in the flanking regions of these STR loci have been observed to differ between populations [267, 272]. In this study, rare variants observed at low frequencies in the GnomAD database (The Genome Aggregation Database <https://gnomad.broadinstitute.org>) were found in the flanking regions associated with the PentaD, D19S433, and D22S1045 loci, namely rs927345580 (G>C; frequency of 1 in 30,540), rs900703892 (C>T; 1 in 143,200) and rs903567912 (delTCTC; 1 in 142,420), respectively. In addition, a novel SNP variant was observed in the flanking region of the D13S317 locus. A notable observation was that two individuals with the same surname (male and female) possessed the same variant rs114697632 (T>C) associated with the D16S539 locus. The most likely explanation for this observation is that these two individuals are in fact related, even though the sample was taken from individuals who were thought to be unrelated. However, the fact that two individuals have the same variant suggests that MPS technology may also be useful for kinship analysis in this population in the future.

Table 5.1. Novel autosomal STR sequence variants.

Thirty-seven novel autosomal STR sequence variants observed in the sequence data from 365 Kuwaiti individuals. Sequence nomenclature is according to ISFG recommendations [266]. The column labelled 'Count' indicates the number of times each allele was observed, and the final column indicates whether the sequence variant was observed in either the repeat region sequence [RS] or the flanking region sequence [FS].

Allele nomenclature	Count	Type
D12S391 [CE 18]-GRCh38-Chr12-12296981-12297168 [AGAT]11 [AGAC]6 AGAT 12297179-G (rs138635218)	2	FS
D12S391 [CE 19]-GRCh38-Chr12-12296981-12297168 [AGAT]10 [AGAC]8 AGAT 12297179-G (rs138635218)	1	FS
D12S391 [CE 19.3]-GRCh38-Chr12-12296981-12297168 AGAT GAT [AGAT]9 [AGAC]8 AGAT	1	RS
D12S391 [CE 21]-GRCh38-Chr12-12296981-12297168 [AGAT]10 [AGAC]11	1	RS
D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 [AGAT]14 [AGAC]8 AGAT 12297179-G (rs138635218)	1	FS
D12S391 [CE 26]-GRCh38-Chr12-12296981-12297168 [AGAT]14 [AGAC]12	2	RS
D13S317 [CE 9]-GRCh38-Chr13-82147986-82148107 [TATC]9 82148072-A (rs-N/A)	1	FS
D16S539 [CE 11]-GRCh38-Chr16-86352664-86352781 [GATA]4 GATC [GATA]6 86352761-C (rs11642858)	2	FS
D16S539 [CE 13]-GRCh38-Chr16-86352664-86352781 [GATA]13 86352779-C (rs114697632)	1	FS
D18S51 [CE 15]-GRCh38-Chr18-63281662-63281796 [AGAA]13 ATAA AGAA	1	RS
D19S433 [CE 12]-GRCh38-Chr19-29926205-29926352 [CCTT]10 CCTA CCTT CTTT CCTT 29926216-29926219 DEL (rs903567912)	1	FS
D1S1656 [CE 14]-GRCh38-Chr1-230769555-230769682 TCTA GCTA [TCTA]12	4	RS
D21S11 [CE 30]-GRCh38-Chr21-19181939-19182111 [TCTA]5 [TCTG]6 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]12	1	RS
D21S11 [CE 31.2]-GRCh38-Chr21-19181939-19182111 [TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]11 TA TCTA	1	RS
D21S11 [CE 33]-GRCh38-Chr21-19181939-19182111 [TCTA]11 [TCTG]5 [TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCA TA [TCTA]10	1	RS
D21S11 [CE 34]-GRCh38-Chr21-19181939-19182111 [TCTA]10 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10	1	RS
D21S11 [CE 34.2]-GRCh38-Chr21-19181939-19182111 [TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]13 TA TCTA	1	RS
D22S1045 [CE 11.2]-GRCh38-Chr22-37140181-37140357 [ATT]4 AT [ATT]4 ACT [ATT]2	1	RS
D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 [ATT]12 ACT [ATT]2 37140181-T (rs900703892)	1	FS
D2S1338 [CE 22]-GRCh38-Chr2-218014856-218014964 [GGAA]2 GGAC [GGAA]11 [GGCA]8	1	RS
D2S441 [CE 9.1]-GRCh38-Chr2-68011918-68012017 A [TCTA]9	1	RS
D2S441 [CE 12]-GRCh38-Chr2-68011918-68012017 [TCTA]12 68011922-A (rs74640515)	1	FS
D3S1358 [CE 15]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 TCTC [TCTA]11	1	RS
D3S1358 [CE 19]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 TCTC [TCTA]15	1	RS
D7S820 [CE 9.1]-GRCh38-Chr7-84160191-84160297 C [TATC]9 84160204-A 84160203-84160204 INS [-/A] (rs7789995) (rs540346880)	1	FS
D7S820 [CE 11]-GRCh38-Chr7-84160191-84160297 [TATC]10 TACC 84160204-A (rs7789995)	2	FS
D8S1179 [CE 9]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]7	1	RS
PentaD [CE 8]-GRCh38-Chr21-43636100-43636278 [AAAGA]8 43636172-C (rs927345580)	1	FS
PentaD [CE 10.4]-GRCh38-Chr21-43636100-43636278 AAAAA AAAGA AAGA [AAAGA]8	1	RS
PentaD [CE 12]-GRCh38-Chr21-43636100-43636278 [AAAGA]2 AAACA [AAAGA]9	1	RS
PentaD [CE 18]-GRCh38-Chr21-43636100-43636278 [AAAGA]18	1	RS
PentaE [CE 10.2]-GRCh38-Chr15-96830996-96831114 [TCTTT]9 TT TCTTT	1	RS
PentaE [CE 12]-GRCh38-Chr15-96830996-96831114 CCTTT [TCTTT]11	1	RS
PentaE [CE 14.4]-GRCh38-Chr15-96830996-96831114 [TCTTT]14 CTTT	1	RS
PentaE [CE 17.4]-GRCh38-Chr15-96830996-96831114 [TCTTT]17 TCTT	1	RS
PentaE [CE 18]-GRCh38-Chr15-96830996-96831114 [TATTT]2 [TCTTT]16	3	RS
vWA [CE 17]-GRCh38-Chr12-5983950-5984049 TGGA [TAGA]11 [CAGA]4 TAGA	1	RS

5.4.6 SE33 locus

The SE33 allele sequences for the 365 samples were recovered using STRait Razor with an average DoC of 178 reads per sample. Allele frequencies for these alleles are shown in **Appendix 8**. From the 350 heterozygotes examined, 140 showed ACR values >0.6 , 153 showed ACR values in the range of 0.20-0.59, and 57 were below 0.20. These relatively low ACR values could be attributed to the lower DoC for the larger alleles in heterozygote pairs compared to the higher coverage for the smaller alleles, as shown in **Figure 5.4**.

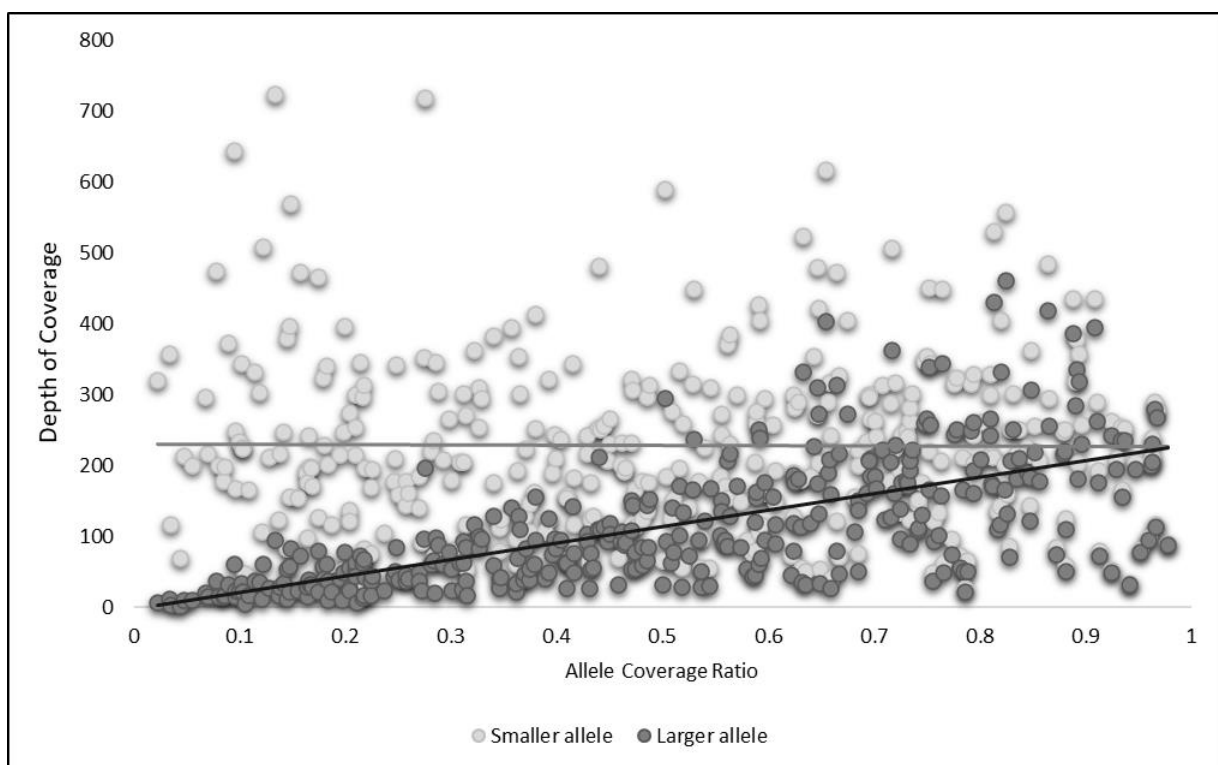


Figure 5.4. A plot of DoC and ACR in heterozygote allele pairs at the SE33 locus.

The DoC in 350 heterozygote allele pairs at the SE33 locus, ordered by the smallest to largest ACR values. The smaller alleles in each pair are shown as pale grey dots and the larger alleles as dark grey dots, indicating that overall, there is higher depth of coverage for the smaller alleles compared to the larger alleles. The trendlines indicate the best fit of the data; the grey trendline shows relatively consistent DoC across the smaller alleles, the black trendline shows that DoC is not consistent across larger alleles, and that higher DoC is associated with higher ACR.

The 365 sequenced samples yielded a total of 41 length-based alleles at the SE33 locus, and this number increased to 147 distinct sequenced-based alleles, of which 18 were novel sequences. Allele frequencies for all 147 alleles are shown in **Appendix 8**. The SE33 sequences were distributed between 17 motif patterns, based on the Borsuk et al. [369] classification. In addition, three new motif patterns were observed that had not previously been reported in

the literature [369, 374]. Three additional IDs were suggested for these motifs: D6, D7 and D8. The SE33 sequence motif patterns are shown in **Table 5.2**.

Table 5.2 SE33 sequence motif patterns observed in 365 Kuwaiti individuals.

The column labelled 'Count' indicates the number of times each allele was observed. The column labelled 'Novel Sequence' shows whether the alleles have been previously reported: *italicised* alleles are novel allele sequences and 'Observed' alleles were reported in [369, 374].

Allele range	Sequence	Count	ID	Novel Sequence
6.3	CT [CTTT]3 [CTTT] _n CT [CTTT]3 CT [CTTT]2	1	C2	Observed
12.2	CT [CTTT]3 C [CTTT] _n [CTTT]3 CT CTTT	1	C0	12.2
15	CT [CTTT]3 T [CTTT] _n CT [CTTT]3 CT [CTTT]2	1	D6*	15
19	CT [CTTT]2 CCTT C [CTTT] _n CT [CTTT]3 CT [CTTT]2	2	A2	Observed
21	CT [CTTT]3 CCTT C [CTTT] _n CT [CTTT]3 CT [CTTT]2	2	A5	Observed
22.1	CT [CTTT]3 C [CTTT] _n T [CTTT] _n CT [CTTT]3 CT [CTTT]2	3	D7*	22.1
22.2	CT [CTTT]3 C [CTTT] _n CT [CTTT] _n CT [CTTT]3 CT [CTTT]2	1	B3	Observed
24.2	CT [CTTT]2 CCTT C [CTTT] _n CTGT [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT [CTTT]2	1	C4	24.2
30.2	CT [CTTT]2 CCTT C [CTTT] _n CT [CTTT] _n CT [CTTT]3 CT CTTT	1	D8*	30.2
31	CT [CTTT]2 CCTT C [CTTT] _n TT [CTTT] _n [CTTT]3 CT [CTTT]2	1	C5	31
11-22	CT [CTTT]3 C [CTTT] _n CT [CTTT]3 CT [CTTT]2	350	A0	Observed
12.2 & 13.2	CT [CTTT]3 C [CTTT] _n [CTTT]3 CT [CTTT]2	2	A6	Observed
13.3-17.3	CT [CTTT]3 C [CTTT] _n C [CTTT] _n [CTTT]3 CT [CTTT]2	4	B2	14.2
15-18	CT [CTTT]2 C [CTTT] _n CT [CTTT]3 CT [CTTT]2	3	B1	15
19.2-34.2	CT [CTTT]2 CCTT C [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT [CTTT]2	311	A1	34.2
25.2-33.2	CT [CTTT]2 CCTT C [CTTT] _n CT [CTTT] _n CT [CTTT]3 CT [CTTT]2	8	A3	31.2, 33.2, 33.2
26.2-31.2	CT [CTTT]2 [CCTT]2 C [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT CTTT	7	B6	26.2, 27.2, 28.2, 31.2
26.2-32.2	CT [CTTT]2 [CCTT]2 C [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT [CTTT]2	9	A4	32.2
27.2-30.2	CT [CTTT]2 CCTT C [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT CTTT	6	A8	Observed
33-38	CT [CTTT]2 CCTT C [CTTT] _n TT [CTTT] _n TT [CTTT] _n CT [CTTT]3 CT [CTTT]2	16	A7	35, 38

* New IDs were given following the work of Borsuk et al. [369].

SE33 concordance testing was performed using the genotypes generated using the CE-based Promega PowerPlex® Fusion 6C kit. A total of three discordant genotypes were found, two of which gave the genotypes (16, 22.2) and (17, 28.2) by CE and (17, 22.2) and (18, 28.2) by sequence length, respectively. By examining the FASTQ files, both discrepancies were due to a TTTT deletion (rs369314007) in the flanking region at chr6:88277355-88277358 (GRCh38.p12) (**Figure 5.5**). Similar discordance has also previously been observed in the literature [369, 374]. The third discordant genotype was (24, 27.2) by CE and (23.2, 27.2) in

the sequence data. The FASTQ file revealed a TC insertion (rs918809513) in the 3' flanking region at chr6:88277300-88277301 (GRCh38.p12) (**Figure 5.5**).

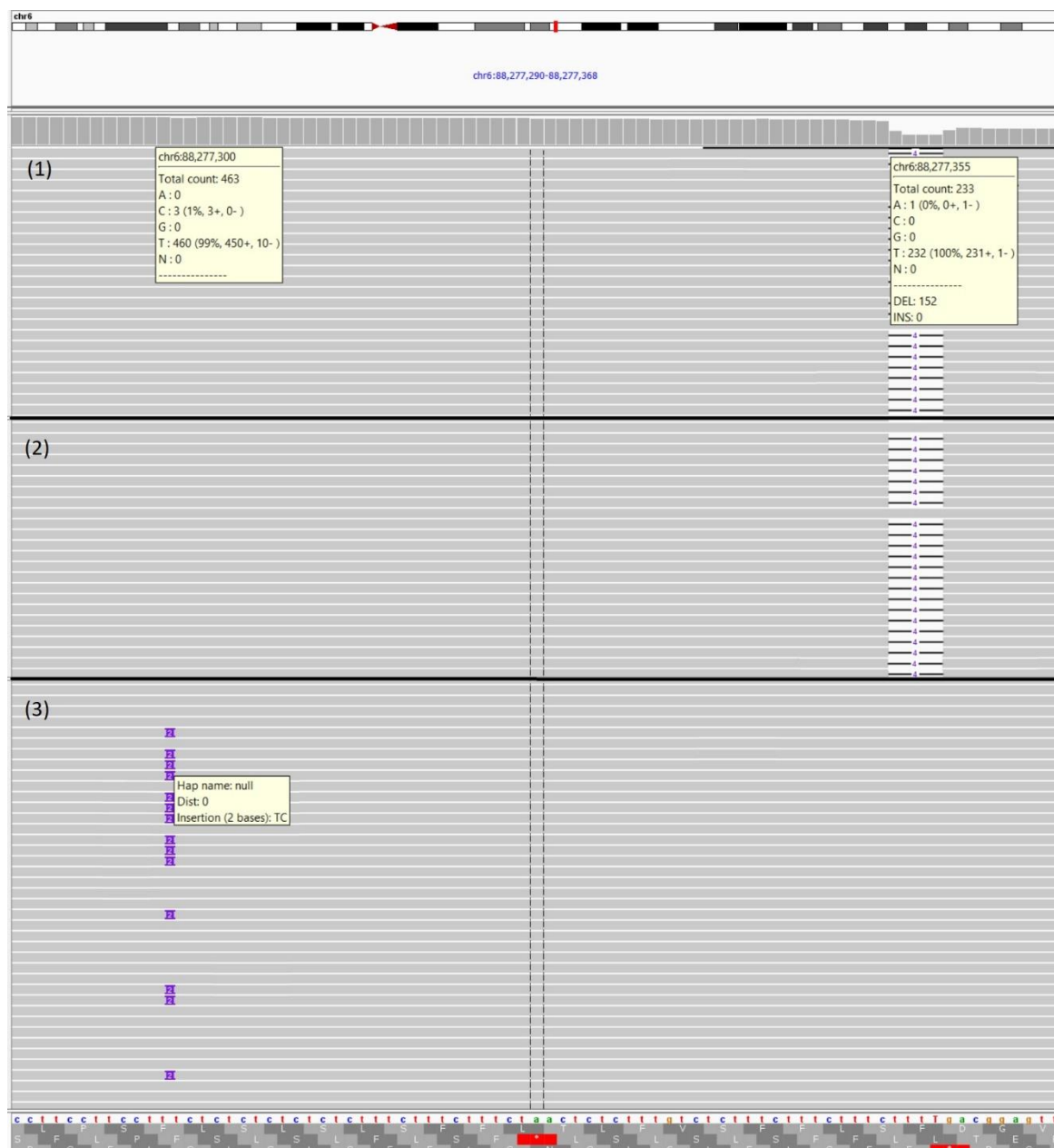


Figure 5.5. IGV display of SE33 read alignment of three samples sequenced using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® system.

The grey tracks represent the alignment reads of three samples to the GRCh38. Sample one and two show a four base pair deletion at 88277355-88277358. The third sample shows a TC insertion at 88277300.

5.4.7 Concordance test

The sequencing results were 100% in concordance with the results obtained using CE for the 23 autosomal loci that are shared between the ForenSeq™ DNA Signature Prep Kit and the Promega PowerPlex® Fusion 6C kit, excluding SE33. One allele drop-out was identified at the D19S433 locus, in which the sample had a heterozygous genotype (12, 14) using CE-based typing, and a homozygous genotype (14, 14) based on allele length taken from the sequencing data. This was confirmed as a drop-out because the sample yielded a full sequencing profile with a high DoC across all loci and no quality control flags triggered. The dropped-out allele was therefore not considered to be a result of discordance between the two kits. Manual examination of the FASTQ file revealed that allele 12 was present, but with coverage of only 4 reads. Further examination of the flanking region did not reveal any variants that might explain why the allele dropped out. The drop-out might therefore be due to a variant at/near the end of the primer binding region, which could disrupt annealing of the primer to its target region.

5.4.8 Forensic statistical analysis

Statistical analysis was carried out on the data divided into three datasets: alleles based on length variation, alleles based on STR repeat region sequence variation and alleles based on flanking region sequence variation. Observed and expected numbers of each genotype at each locus were compared using the Fisher's Exact test in *Arlequin* software, based on the method described by Guo and Thompson [173]. The HWE tests resulted in P-values >0.05 (**Table 5.3**) for all the tests except for the D8S1179 locus, which showed marginal deviation from expectations. After using Bonferroni correction, a new significance value of 0.0018 was determined using the number of tests ($0.05/27=0.0018$), and then none of the loci showed any deviation from HWE. Therefore, all the observed alleles were independent, and can be utilised to calculate match probabilities in the Kuwaiti population. Association between alleles at all possible pairwise combinations of loci was assessed using 378 Linkage Disequilibrium (LD) tests. After Bonferroni correction ($0.05/378= 0.00013$), none of the pairs of loci showed significant LD. Therefore, each locus is inherited independently, and the product rule can be used to multiply match probabilities across loci.

Table 5.3. Hardy-Weinberg Equilibrium (HWE) test

Observed and expected heterozygosity and Fisher's Exact test P-values for length-based, sequence-based and flank-based genotypes. Significant departure from HWE is highlighted in grey.

Locus	Number of samples	Length-based			Repeat-based			Flank-based		
		Observed heterozygosity	Expected heterozygosity	P-value	Observed heterozygosity	Expected heterozygosity	P-value	Observed heterozygosity	Expected heterozygosity	P-value
CSF	365	0.72877	0.7236	0.90238	0.73425	0.72818	0.87809	0.73425	0.72818	0.8532
D10S1248	365	0.76986	0.76564	0.22097	0.76986	0.76564	0.18769	0.76986	0.76564	0.19877
D12S391	365	0.86301	0.884	0.39146	0.92329	0.93986	0.84224	0.92603	0.94154	0.81232
D13S317	365	0.75342	0.76574	0.86857	0.75342	0.76574	0.87127	0.85205	0.87381	0.42999
D16S539	365	0.78904	0.78941	0.1534	0.79178	0.7911	0.14433	0.81918	0.81918	0.1716
D17S1301	365	0.64384	0.67209	0.27911	0.64384	0.67209	0.28912	0.64384	0.67209	0.27496
D18S51	365	0.85205	0.87801	0.10118	0.85205	0.87964	0.13322	0.85205	0.87964	0.13232
D19S433	365	0.83014	0.84469	0.95919	0.83014	0.8453	0.94794	0.83014	0.84556	0.911
D1S1656	365	0.88767	0.87292	0.15721	0.90411	0.8988	0.08845	0.90411	0.8988	0.13322
D20S482	365	0.69863	0.69565	0.3108	0.69863	0.69565	0.30706	0.74521	0.76356	0.27619
D21S11	365	0.86027	0.84212	0.54437	0.91233	0.90651	0.52707	0.91233	0.90651	0.54979
D22S1045	365	0.69863	0.7184	0.46933	0.69863	0.7184	0.49846	0.70137	0.72083	0.33024
D2S1338	365	0.83836	0.88293	0.21412	0.90137	0.92511	0.25605	0.90137	0.92511	0.27594
D2S441	365	0.72055	0.74416	0.4585	0.73425	0.75989	0.57625	0.76164	0.78276	0.6448
D3S1358	365	0.77534	0.76821	0.57638	0.88219	0.89045	0.6348	0.88219	0.89045	0.64508
D4S2408	365	0.76986	0.76077	0.61122	0.7726	0.76458	0.31813	0.7726	0.76458	0.33145
D5S818	365	0.73699	0.75816	0.11462	0.73699	0.7587	0.08127	0.80274	0.83033	0.08192
D6S1043	365	0.82466	0.82679	0.71098	0.82466	0.82712	0.64172	0.82466	0.82712	0.6108
D7S820	365	0.77808	0.79437	0.1623	0.77808	0.79617	0.0502	0.8274	0.84749	0.40971
D8S1179	365	0.8411	0.83595	0.21905	0.89041	0.89744	0.04428	0.89041	0.89744	0.04826
D9S1122	365	0.71233	0.70558	0.12037	0.83562	0.84013	0.34041	0.83562	0.84013	0.36258
FGA	365	0.83288	0.86192	0.29729	0.83836	0.86574	0.2937	0.83836	0.86574	0.31228
PentaD	365	0.83562	0.84994	0.57682	0.83562	0.85057	0.71237	0.83836	0.85131	0.71099
PentaE	357	0.90196	0.9104	0.23311	0.90196	0.91124	0.16412	0.90196	0.91124	0.14968
TH01	365	0.77534	0.79049	0.59955	0.77534	0.79049	0.60198	0.77534	0.79049	0.58854
TPOX	365	0.66575	0.65928	0.51833	0.66575	0.65928	0.51562	0.66575	0.65928	0.48205
vWA	365	0.8137	0.80836	0.99628	0.8411	0.84098	0.32316	0.8411	0.84104	0.27225
SE33	365	0.93973	0.94849	0.78898	0.95616	0.96742	0.93615	0.95616	0.96742	0.26259

Forensically relevant parameters were calculated using STRAF [310] for the three different categories of alleles. The forensic statistics for each marker are detailed in **Table 5.4**, and a summary for each of the three datasets is given in **Table 5.5**. The inclusion of the sequence-based and flank-based alleles improved the values of all the forensic indices compared to the length-based data.

Table 5.4. Forensic statistics in the Kuwaiti population using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® system.

Number of observed alleles, observed heterozygosity and forensic stats for length-based (LB), sequence-based (SB) and flank-based (FB) genotypes.

Locus	Number of sampled alleles	Number of observed alleles			observed heterozygosity			Random Match Probability (RPM)			Power of Discrimination (PD)			Power of Exclusion (PE)			Typical Paternity Index (TPI)			Polymorphism Information Content (PIC)		
		LB	SB	FB	LB	SB	FB	LB	SB	FB	LB	SB	FB	LB	SB	FB	LB	SB	FB	LB	SB	FB
CSF	730	8	9	9	0.729	0.734	0.734	0.129	0.126	0.126	0.871	0.874	0.874	0.474	0.483	0.483	1.843	1.881	1.881	0.672	0.678	0.678
D10S1248	730	10	10	10	0.770	0.770	0.770	0.095	0.095	0.095	0.905	0.905	0.905	0.544	0.544	0.544	2.173	2.173	2.173	0.729	0.729	0.729
D12S391	730	17	66	72	0.863	0.923	0.926	0.027	0.008	0.008	0.973	0.992	0.992	0.721	0.843	0.849	3.650	6.518	6.759	0.871	0.936	0.938
D13S317	730	8	8	19	0.753	0.753	0.852	0.091	0.091	0.031	0.909	0.909	0.969	0.516	0.516	0.699	2.028	2.028	3.380	0.730	0.730	0.859
D16S539	730	8	9	17	0.789	0.792	0.819	0.078	0.077	0.059	0.922	0.923	0.941	0.579	0.584	0.635	2.370	2.401	2.765	0.758	0.760	0.797
D17S1301	730	7	7	7	0.644	0.644	0.644	0.162	0.162	0.162	0.838	0.838	0.838	0.347	0.347	0.347	1.404	1.404	1.404	0.614	0.614	0.614
D18S51	730	17	20	20	0.852	0.852	0.852	0.029	0.029	0.029	0.971	0.971	0.971	0.699	0.699	0.699	3.380	3.380	3.380	0.864	0.866	0.866
D19S433	730	15	16	17	0.830	0.830	0.830	0.042	0.042	0.042	0.958	0.958	0.958	0.656	0.656	0.656	2.944	2.944	2.944	0.826	0.827	0.827
D1S1656	730	15	30	30	0.888	0.904	0.904	0.031	0.020	0.020	0.969	0.980	0.980	0.770	0.804	0.804	4.451	5.214	5.214	0.859	0.890	0.890
D20S482	730	9	9	13	0.699	0.699	0.745	0.138	0.138	0.086	0.862	0.862	0.914	0.426	0.426	0.502	1.659	1.659	1.962	0.653	0.653	0.736
D21S11	730	17	47	47	0.860	0.912	0.912	0.048	0.019	0.019	0.952	0.981	0.981	0.715	0.821	0.821	3.578	5.703	5.703	0.822	0.898	0.898
D22S1045	730	10	10	13	0.699	0.699	0.701	0.120	0.120	0.118	0.880	0.880	0.882	0.426	0.426	0.430	1.659	1.659	1.674	0.677	0.677	0.680
D2S1338	730	14	49	49	0.838	0.901	0.901	0.027	0.013	0.013	0.973	0.987	0.987	0.672	0.798	0.798	3.093	5.069	5.069	0.870	0.919	0.919
D2S441	730	12	16	19	0.721	0.734	0.762	0.105	0.092	0.075	0.895	0.908	0.925	0.461	0.483	0.530	1.789	1.881	2.098	0.706	0.727	0.755
D3S1358	730	8	25	25	0.775	0.882	0.882	0.098	0.024	0.024	0.902	0.976	0.976	0.554	0.759	0.759	2.226	4.244	4.244	0.729	0.879	0.879
D4S2408	730	6	7	7	0.770	0.773	0.773	0.100	0.097	0.097	0.900	0.903	0.903	0.544	0.549	0.549	2.173	2.199	2.199	0.719	0.725	0.725
D5S818	730	7	8	15	0.737	0.737	0.803	0.099	0.099	0.050	0.901	0.901	0.950	0.488	0.488	0.604	1.901	1.901	2.535	0.718	0.719	0.809
D6S1043	730	16	20	20	0.825	0.825	0.825	0.053	0.053	0.053	0.947	0.947	0.947	0.646	0.646	0.646	2.852	2.852	2.852	0.806	0.806	0.806
D7S820	730	9	11	21	0.778	0.778	0.827	0.075	0.074	0.042	0.925	0.926	0.958	0.559	0.559	0.651	2.253	2.253	2.897	0.763	0.765	0.829
D8S1179	730	11	25	25	0.841	0.890	0.890	0.054	0.024	0.024	0.946	0.976	0.976	0.677	0.776	0.776	3.147	4.563	4.563	0.814	0.888	0.888
D9S1122	730	7	12	12	0.712	0.836	0.836	0.146	0.047	0.047	0.854	0.953	0.953	0.448	0.667	0.667	1.738	3.042	3.042	0.653	0.819	0.819
FGA	730	19	26	26	0.833	0.838	0.838	0.036	0.034	0.034	0.964	0.966	0.966	0.661	0.672	0.672	2.992	3.093	3.093	0.845	0.850	0.850
PentaD	730	16	18	21	0.836	0.836	0.838	0.041	0.041	0.040	0.959	0.959	0.960	0.667	0.667	0.672	3.042	3.042	3.093	0.831	0.832	0.833
PentaE	714	23	26	26	0.902	0.902	0.902	0.017	0.017	0.017	0.983	0.983	0.983	0.799	0.799	0.799	5.100	5.100	5.100	0.902	0.903	0.903
SE33	730	41	147	147	0.940	0.956	0.956	0.008	0.002	0.002	0.992	0.995	0.995	0.877	0.911	0.911	8.295	11.405	11.405	0.945	0.965	0.965
TH01	730	7	7	7	0.775	0.775	0.775	0.077	0.077	0.077	0.923	0.923	0.923	0.554	0.554	0.554	2.226	2.226	2.226	0.757	0.757	0.757
TPOX	730	8	8	8	0.666	0.666	0.666	0.162	0.162	0.162	0.838	0.838	0.838	0.377	0.377	0.377	1.496	1.496	1.496	0.614	0.614	0.614
vWA	730	8	23	24	0.814	0.841	0.841	0.067	0.047	0.047	0.933	0.953	0.953	0.625	0.677	0.677	2.684	3.147	3.147	0.780	0.821	0.821

Table 5.5. Summary statistics for forensic parameters in the Kuwaiti population.

Forensic statistics summary across 28 STR loci (including SE33) calculated using length-based, sequence-based and flank-based alleles obtained by sequencing 365 Kuwaiti individuals.

Forensic parameter	Length-based	Sequence-based	Flank-based
Average Heterozygosity	0.7906	0.8101	0.8216
Average Polymorphism Information Content (PIC)	0.7688	0.7945	0.8101
Average Typical Paternity Index (TPI)	2.7909	3.3741	3.5106
Combined paternity index (CPI)	2.65×10^{10}	1.28×10^{13}	5.92×10^{13}
Combined Match Probability (CMP)	1.54×10^{-34}	3.21×10^{-38}	1.17×10^{-39}
Combined Power of Exclusion (CPE)	> 0.99999999	> 0.99999999	> 0.99999999

The combined match probability (CMP) yielded by the 28 autosomal loci (including SE33) decreased from 1.54×10^{-34} for the length-based alleles, to 3.21×10^{-38} when considering the STR repeat region sequence variation. In addition, the inclusion of the flanking region sequence variant alleles further decreased the CMP to 1.17×10^{-39} . That is, the probability of finding two unrelated individuals with identical profiles in the Kuwaiti population is one in

8.55×10^{38} . The use of MPS technology surpasses conventional CE-based technology by providing considerably more power to DNA profile evidence, decreasing the match probabilities that can be generated using a single assay (**Table 5.6**).

Table 5.6. Comparison of combined match probabilities (CMP) in the Kuwaiti population.

Comparison of CMP values calculated using different marker types surveyed in this study and previous studies. In the column labelled 'marker type', (LB) is length-based, (SB) is sequence-based and (FB) is flank-based.

Multiplex kit	Marker type	Number of loci	CMP	Reference
ForenSeq™ DNA Signature Prep (without SE33)	STR (LB)	27	2.03×10^{-32}	This study
	STR (SB)	27	1.37×10^{-35}	This study
	STR (FB)	27	4.99×10^{-37}	This study
ForenSeq™ DNA Signature Prep (with SE33)	STR (LB)	28	1.54×10^{-34}	This study
	STR (SB)	28	3.21×10^{-38}	This study
	STR (FB)	28	1.17×10^{-39}	This study
PowerPlex® Fusion 6C	STR	23	7.37×10^{-30}	Chapter 4
CS7+ PP21 (combined)	STR	25	3.08×10^{-31}	[300]
Identifiler	STR	15	2.62×10^{-18}	[299]
DIPplex	Indel	30	2.73×10^{-13}	Chapter 3

The typical paternity index (TPI) is used in paternity testing and it is the likelihood ratio of the probability of the alleged man being the biological father of the questioned offspring, compared to a random individual in the population. The higher the value is, the greater the power of a set of markers to determine whether the alleged father is the real one. In this study, the combined paternity index (CPI) for the length-based results was calculated as 2.65×10^{10} , which increased to $\sim 10^{13}$ for the sequence-based results, for the 28 STRs in the Kuwaiti samples. This high CPI yielded when sequence variation is incorporated increases the informativeness of the system for paternity testing.

The current problem associated with using CE-based STR analysis in paternity testing is the potential genotype mismatches caused by mutations such as indels or SNPs within the STR locus, or due to drop-out of alleles, which is occasionally seen in paternity cases involving father-mother-child trios or father-child/mother-child duos. In such cases, the STRs used would be inconclusive, causing erroneous exclusion of paternity. The current solution to provide more evidence in such cases is to use additional STR markers or to use Sanger sequencing to identify the mutations. The use of the ForenSeq™ DNA Signature Prep kit in

this context can be useful to solve these difficult cases by detecting the actual sequence and revealing the existence of mutations.

5.5 Conclusion

This study was undertaken to evaluate the forensic utility of the 27 autosomal STR markers contained within the Verogen ForenSeq™ DNA Signature Prep Kit in the Kuwaiti population and to generate allele frequency data for a population where none has previously been generated for this kit as well as to determine whether data for the SE33 locus could be recovered from the sequencing results. All markers met the statistical requirements of showing no significant deviation from the expectations of the Hardy Weinberg Equilibrium, and no significant linkage disequilibrium, confirming that the system can be utilised for forensic testing in the Kuwaiti population. Allele variants were categorised based on length alone, based on the sequence of the STR repeat region, and based on the sequence of the STR flanking region. A number of novel alleles were observed in this study that have not been previously published in any population. The increased number of alleles generated by sequencing both the STR repeat region and the flanking region highlights the potential value of this sequencing technology. The technology is rapidly improving, and it is anticipated that the methodology involved will become increasingly straightforward and affordable as the cost of sequencing has dramatically dropped over the last two decades. Further studies are therefore required to facilitate a large-scale comparison of populations, particularly in the Middle East region, which is relatively underrepresented in terms of population genetic data.

Chapter 6 Massively parallel sequencing of 94 identity-informative SNPs in Kuwaiti populations

6.1 Introduction

Single-nucleotide polymorphisms (SNPs) are single-base substitutions and represent the most abundant type of polymorphism found in the human genome. SNP identification can be utilised to supplement the standard short-tandem repeat (STR) typing system, especially for the analysis of forensic samples that may contain DNA at low quantities or with a high degree of degradation, which may be more amenable to analysis using small-sized amplicons. SNPs are characterised by a lower mutation rate (10^{-8} per locus per generation [375]) compared with that of STRs (10^{-3} per locus per generation [376]), which means SNPs are inherited in stable patterns across generations and makes them suitable markers for studying ancestral relationships and for use in phenotypic prediction studies [65]. In addition, large SNP panels can be as informative as the smaller, 15-locus commercial STR kits (e.g. Identifiler and PowerPlex 16) [67, 77]. However, to generate match probabilities similar to those produced by recent large STRs kits (e.g. Fusion 6C), additional SNP markers are necessary, requiring the acquisition of more template DNA from the often limited quantity associated with forensic samples. This requirement for a larger quantity of DNA, combined with the intensive laboratory work involved in SNP typing using Sanger Sequencing and capillary electrophoresis (CE), means that SNP markers are not generally considered to be suitable for forensic analysts.

With the introduction of MPS techniques into forensic practice, hundreds of SNPs, in addition to other classes of genetic markers, can be co-amplified and individually sequenced in parallel. A major disadvantage of SNP markers has been the complexity of deciphering mixed DNA samples using SNPs, because they tend to be diallelic whereas STRs have large numbers of alleles at each locus. Although most known SNPs are diallelic, tri- and tetra-allelic [6, 7], and haplotype-SNPs may serve as an alternative approach for mixed DNA sample cases because they generate multiple alleles, which is more comparable to the traditional STR typing methods [8].

Microhaplotypes (microhaps) have recently received a great deal of attention within the forensics community as promising alternatives to independent SNPs and STRs for the

purposes of human identification [377, 378]. A microhap is defined as a marker that comprises two or more closely linked SNPs (a haplotype), which are separated by fewer than 200 bp (ensuring a low recombination frequency between haplotypes) and are capable of generating three or more allelic combinations [379]. Microhaplotypes are therefore multi-SNP haplotypes, and can be more informative than independent SNP markers on a per locus basis.

Microhaplotypes are distributed throughout the genome and are advantageous compared with STR markers. One advantage is the absence of stutter artefacts, which can complicate the distinction of a minor contributor from a major contributor in mixed-source samples. Another advantage is that Microhaplotypes possess similar allele lengths at each locus, whereas STR alleles can differ in length by up to 100 nucleotides [378]. During the analysis of STRs, especially in samples with degraded, low-quality, or low-quantity nuclear material, smaller alleles (in terms of repeats) are preferentially amplified over larger alleles, which can complicate the interpretation of results in both single- and mixed-source samples.

One limitation of Microhaplotypes is the small number of alleles they feature, which can subsequently make it difficult to produce acceptable values of match probabilities, unlike conventional STRs, which are more discriminative due to their large and multi-allelic nature. However, to reach a similar level of discrimination as STR analysis, large numbers of SNPs and multi-allelic Microhaplotypes can be combined into a single panel. Given that allele frequencies differ between populations, for such large panels to be implemented in forensic laboratories, population-specific allele frequencies must be established to determine match probabilities and likelihood ratios for DNA evidence. This chapter represents the first study performed to assess the population genetics of identity-informative SNPs (iiSNPs) for use as forensic markers in the Kuwaiti population. These markers are all contained in the ForenSeq™ DNA Signature Prep Kit.

6.2 Aims and objectives

The primary aim of this chapter was to evaluate the forensic utility of the 94 iiSNP markers included in the ForenSeq™ DNA Signature Prep Kit in the Kuwaiti population and to contribute to the population genetic data available for this under-represented region. This study involves:

- Calculating the allele frequencies for, and assess the forensic parameters calculated using the sequenced iiSNP makers.
- Identifying the genetic variants within the entire sequence of 94 iiSNP-containing-amplicons and assess the level of diversity associated with these variants among the Kuwaiti population.
- Examining the influence of including these additional variants on the performance of forensic parameters.
- Examining the genetic distance between the Kuwaiti population and other populations, at the global and regional levels.

6.3 Materials and methods

The iiSNP dataset used this chapter was obtained from sequencing data that was generated from 376 samples using the FGx® system and the ForenSeq™ DNA Signature Prep Kit, as described in Section 2.8.

6.3.1 Data analysis

Sequencing data were initially analysed using the graphical interface program, ForenSeq™ Universal Analysis software (UAS), version 1.3. The default analysis settings of 1.5% for the analytical threshold (AT) and 4.5% for the interpretation threshold (IT) were used for data interpretation (see Section 5.3.5 for illustrations of AT and IT).

Positive and negative controls for SNP markers were all verified after the completion of each run. SNP genotype calls were viewed through the UAS graphical interface, which also includes quality control flags (indicators) for each locus. Manual interpretation/editing was performed when, for example, a flag showed a low-coverage allele at a locus or imbalance, which required a user action to define alleles. A genotype report was generated for each sample, which included genotype calls, the number of reads, and the quality control indicators for all markers. All reports were manually reviewed and then exported to a single Excel file prior to performing statistical analyses. In addition, the flanking sequence report option, implemented in the UAS, was used to generate the flanking region sequences for all samples.

6.3.2 Bioinformatics analysis

FASTQ files were further analysed using STRait Razor, version 3.0 [306], and were mapped to the GRCh38 reference genome. This was performed to examine the concordance of the allele designations between UAS, which uses an alignment-based method, and STRait Razor, which uses the haplotype capture method, to assign the correct nomenclature to each SNP allele, according to ISFG recommendations [266]. Where no SNP/variant was identified in the STRait Razor database, the allele was designated as 'NovelSeq'. Two loci, rs891700 and rs914165, were analysed manually because the interpretation was complicated by the presence of homopolymers (repeated bases) in the small flanking region captured by STRait Razor [380].

The FASTQ files for the samples containing novel alleles were manually aligned to the GRCh38 genome using BWA (Burrows-Wheeler Aligner) [307], version 0.7.15-r1140, and the resulting bam files were visualised using IGV (Integrative Genomics Viewer), version 2.8.0 [308]. This provided a view of the variant locations, which were then identified using Variation Viewer, 2.1.4 [371], and labelled with the appropriate variant code (rs-number), if available.

6.3.3 Forensic statistical analysis

Arlequin version 3.5 was used to calculate a number of population genetic parameters, including allele frequencies and observed and expected heterozygosities, to test for departure from Hardy-Weinberg Equilibrium (HWE), and identify evidence of linkage disequilibrium (LD) [309]. The effective number of alleles (A_e) was calculated using the formula $1/\sum p_i^2$, where p_i is the frequency of each allele, summed across all loci [381].

Forensic parameters, including the random match probability (RMP), discrimination power (PD), power of exclusion (PE), typical paternity index (TPI), and the polymorphic information content (PIC), were calculated using STRAF, an online tool for STR data analysis [310].

6.3.4 Genetic distance

Arlequin version 3.5 was used to estimate pairwise F_{ST} genetic distances between 51 global populations based on genotypic data for 87 iiSNP loci. These populations were grouped into seven geographical regions: Africa (AFR; $N = 113$), America (AMR; $N = 63$), Central South Asia (CSA; $N = 199$), the Middle East (ME; $N = 161$), Europe (EUR; $N = 156$), East Asia (EAS; $N = 229$), and Oceania (OCE; $N = 30$). The genotypic data were called from the Human Genome Diversity

Project (HGDP) - the CEPH (Centre d'Etude du Polymorphisme Humain) [382], by the work done by Khubrani et al. [373]. Data from 89 Saudi Arabian (KSA) individuals were also included in the analysis [373]. F_{ST} -based genetic distance heatmaps were generated using the R function 'pairFstMatrix.r', which was implemented in Arlequin software.

STRUCTURE analysis was performed as described in section 2.11.2 using the following parameters: 'admixture' and 'correlated allele frequencies' models; 100,000 Markov chain Monte Carlo (MCMC) steps for each run, with the first 10,000 steps discarded as burn-in; and the inferred number of K set from 1 to 10. At each K , the analysis was repeated five times to test the results for consistency. The results were visualised using CLUMPAK (Clustering Markov Packager Across K [327], and the best K was determined using Structure Harvester [328].

6.4 Results and discussion

6.4.1 Sequencing performance

The ForenSeq™ DNA Signature Prep Kit comprises 94 primer pairs designed to amplify identity-informative SNPs (iiSNPs) that are forensically relevant. The depth of coverage (DoC) per SNP locus was calculated to evaluate the accuracy of MPS for the genotyping the 94 iiSNPs. Read coverage per locus has been associated with allele dropout, with the risk of dropout increasing with reduced coverage. In this study, variation in the DoC was observed across the 94 SNP loci, as shown in **Figure 6.1**. Several factors might affect SNP sequencing coverage, such as PCR amplification efficiency, the accuracy of purification steps, or unsuccessful bead-based library normalisation [258]. The initial quantity of DNA input and the fact that a large number of markers are multiplexed into one sequencing run could also result in DoC differences among SNP loci [258].

The six loci with the lowest DoC were observed to feature either missing genotypes or low-coverage homozygotes (allele reads < 30). These low-performing loci were removed from the analysis due to the high possibility of allele dropout and to provide an accurate analysis of 88 high-performing loci which are included in this study. The removed loci were: rs1736442 (missing genotypes: 42; homozygotes <30 reads: 153), rs7041158 (missing genotypes: 10; homozygotes <30 reads: 98), rs1031825 (missing genotypes: 11; homozygotes <30 reads: 82), rs2920816 (missing genotypes: 19; homozygotes <30 reads: 74), rs1357617 (missing

genotypes: 6; homozygotes <30 reads: 46), and rs1493232 (homozygotes <30 reads: 33). These loci have also previously been reported in the literature as being associated with low DoC [258, 269, 373, 383].

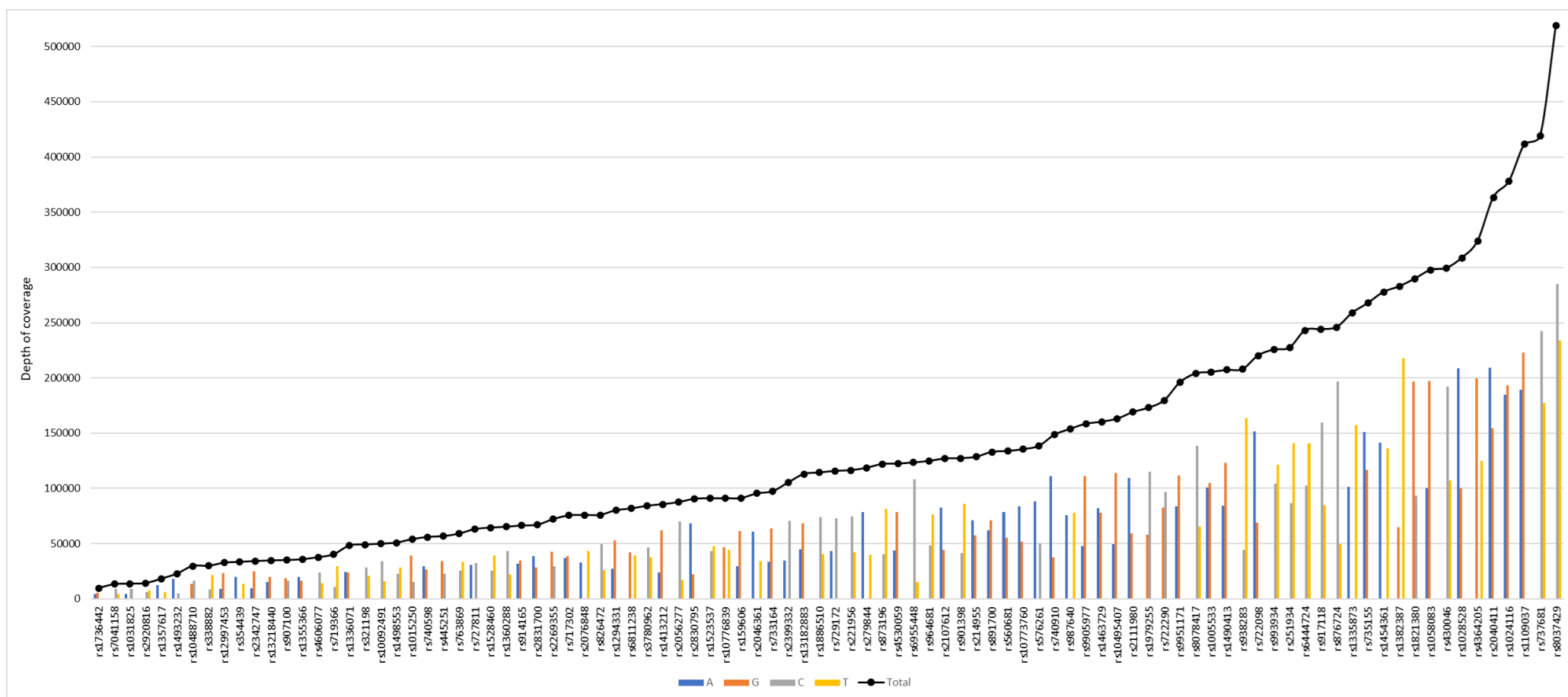


Figure 6.1. Depth of Coverage for ForenSeq™ DNA Signature Prep Kit iiSNPs.

The 94 iiSNPs are arranged in order of increasing total DoC per locus. The bars show the number of reads per allele at each locus. The black line indicates the total number of reads generated per locus.

To evaluate the intralocus heterozygous balance, the allele coverage ratio (ACR) was determined for each heterozygote among the remaining 88 iiSNPs by calculating the ratio of the number of reads for each of the two alleles. Five loci: rs6955448 (ACR range: 0.2-0.5), rs338882 (ACR range: 0.2-0.6), rs4530059 (ACR range: 0.1-0.5), rs2111980 (ACR range: 0.3-0.5), and rs1005533 (ACR range: 0.3-0.5) in 147, 117, 102, 52, and 27 heterozygotes, respectively. ACR values < 0.6 have previously been reported in the literature for rs6955448 [258, 269], rs338882 [258, 269, 383], rs4530059 [258, 383], rs2111980 [383], and rs1005533 [383]. These loci should be interpreted with caution when evaluating SNP profiles typed using the MiSeq FGx[®] system.

6.4.2 Sample qualification and genotyping

As described in section 5.3, 376 samples were sequenced using four sequencing runs and 369 out of 376 samples were included in this study. Two of the samples were excluded due to low coverage, whereas the remaining five samples were excluded due to the presence of extra alleles, likely due to contamination (which could be clearly observed in the STR analysis). The positive control aligned completely with the positive control profile provided in the ForenSeq manual. The positive control DNA (2800M) profile is shown in **Figure 6.2**. Full iiSNP profiles were produced for all 369 samples, except for two, which featured two missing loci, at rs10488710 (two individuals) and rs2076848 (one individual).

IDENTITY SNPs					
Single Source Sample 94/94 Loci Typed					
rs1490413	rs560681	rs1294331	rs10495407	rs891700	rs1413212
AA	AG	GA	GG	AG	GG
rs876724	rs1109037	rs993934	rs12997453	rs907100 II	rs1357617 II
CC	GG	CC	AA	GC	TA
rs4364205	rs2399332	rs1355366 II	rs6444724	rs2046361	rs279844
GG	AC	AG	TT	AA	AT
rs6811238	rs1979255	rs717302	rs159606	rs13182883	rs251934
GG	GG	GG	AA	GA	TT
rs338882	rs13218440 II	rs1336071	rs214955	rs727811	rs6955448 II
CC	GA	GG	GG	AA	CT
rs917118	rs321198	rs737681	rs763869	rs10092491	rs2056277
CC	TT	TT	CT	TC	CC
rs4606077	rs1015250	rs7041158 II	rs1463729	rs1360288	rs10776839
TC	GG	CC	GA	CC	GG
rs826472	rs735155	rs3780962	rs740598	rs964681	rs1498553
TT	AA	TT	GA	TC	CT
rs901398	rs10488710 II	rs2076848	rs2107612	rs2269355	rs2920816 II
TT	GC	TA	GA	CC	TT
rs2111980	rs10773760	rs1335873	rs1886510	rs1058083	rs354439
GG	AG	TT	CT	AG	TT
rs1454361	rs722290	rs873196	rs4530059	rs1821380	rs8037429
AT	GG	CT	GG	GG	TT
rs1528460	rs729172	rs2342747 II	rs430046	rs1382387	rs9905977
TT	CC	AG	CC	GT	GG
rs740910	rs938283	rs8078417	rs1493232 II	rs9951171	rs1736442 II
AA	TT	CT	AA	GG	GG
rs1024116	rs719366	rs576261	rs1031825 II	rs445251	rs1005533
AA	TT	AC	CC	CG	AA
rs1523537	rs722098	rs2830795	rs2831700	rs914165	rs221956
CC	AG	AA	AA	GA	CC
rs733164	rs987640	rs2040411	rs1028528		
GA	TA	AA	AG		

Figure 6.2. ForenSeq™ DNA Signature Prep Kit positive control profile for 94 loci SNP.

Positive control profile generated from 0.2 ng/μL of the male DNA sample 2800M. The profile shows all 94 iiSNPs. The UAS highlights in orange those loci that require manual interpretation/investigation of quality-control flags.

6.4.3 Sequence variation

Sequence variants within the flanking region of each SNP amplicon were analysed using the Flanking Region Report generated by the UAS. Further analyses and the assignment of nomenclature were performed in STRait Razor. Of the 88 iiSNP amplicons analysed, 45 (51%) had variants within the flanking regions. These multi-SNP loci (microhaps) produced a total of 172 distinct haplotypes, ranging from three to seven haplotypes, as shown in **Figure 6.3**. Increased discriminatory power was observed when comparing the microhaps at each locus with the power for the diallelic, independent target SNP, as shown in **Figure 6.3**. Of the distinct haplotypes, 55 (32%) had rare haplotype frequencies of less than 0.01, whereas the majority of haplotypes were found to be common across the Kuwaiti population, with frequencies above 0.01. **Appendix 9** shows the allele frequencies of all 88 analysed iiSNPs in the 369 Kuwaiti individuals based on two classification criteria: targeted-iiSNP, and iiSNP-flanking region sequence variants (referred to as iiSNP-flank), using nomenclature based on ISFG recommendations [266].

A well-known limitation of SNP markers is the low discrimination power they confer due to their diallelic nature. However, these results show that the inclusion of sequencing data from a higher level of sequencing technology increases the informative value of single-base-SNPs, especially those associated with variants in their flanking regions, which can provide additional power for the evaluation of evidence. As SNPs are amplified in smaller amplicons, with no stutter production compared with traditional STR markers, the SNP-microhaplotype system may serve as an efficient tool for discrimination between individuals, in addition to serving as a useful typing tool for the analysis of degraded or low-quantity DNA samples.

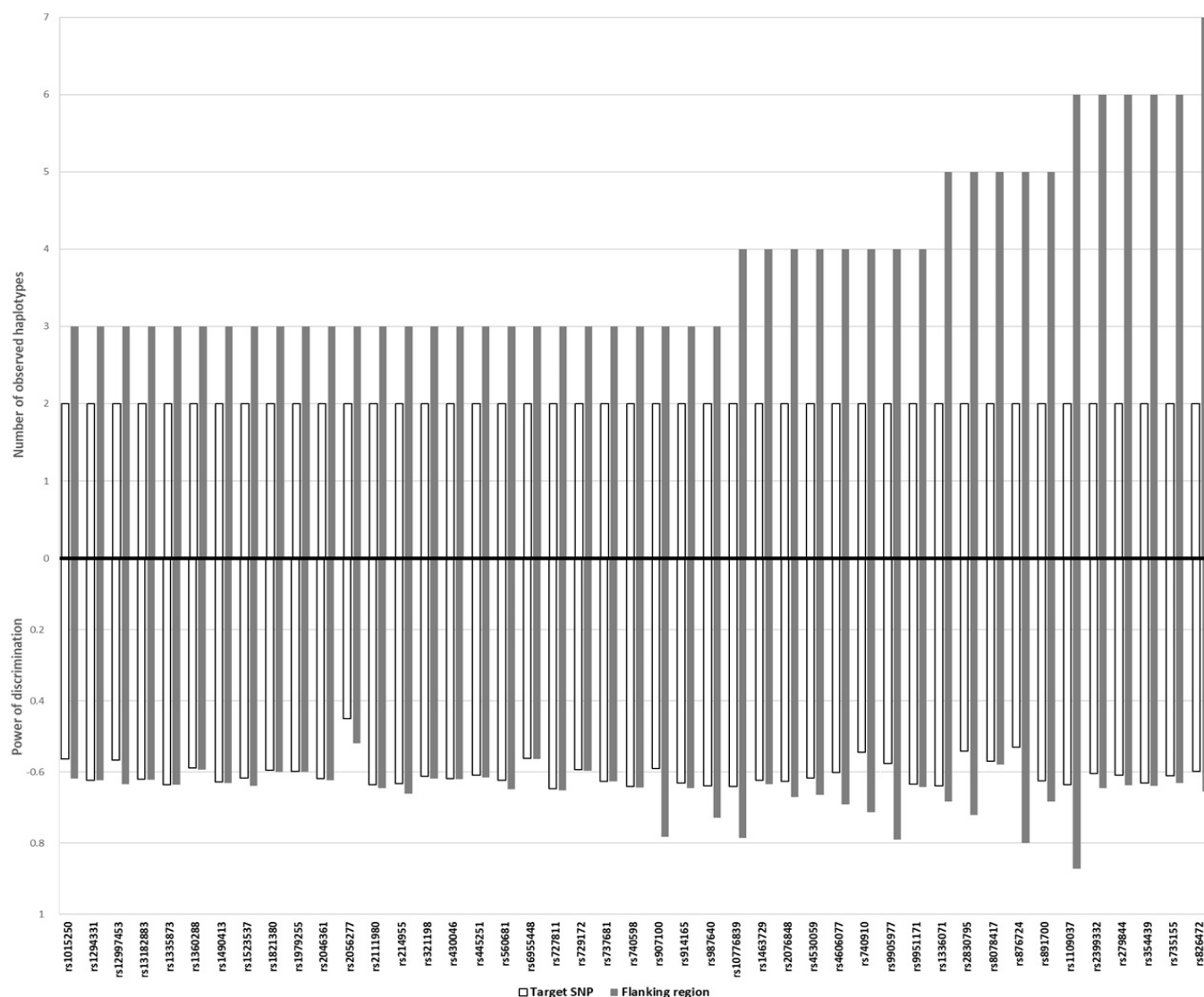


Figure 6.3. Number of observed haplotypes at each SNP locus and the increased discriminatory power associated with sequence variants.

Above the x-axis, the white bars represent the number of di-allelic targeted SNP alone, and the grey bars represent the total number of haplotypes per multi-allelic microhaplotype, based on iiSNPs with flanking region variants. Below the x-axis, the influence of flanking variants is demonstrated by the increased discriminatory power of each microhaplotype (grey bars), compared with the targeted SNP alone (white bars).

6.4.4 Novel haplotypes

A total of 27 distinct haplotypes, distributed among 17 iiSNPs, were identified by STRait Razor as novel alleles due to the presence of a variant proximate to the targeted SNP. These novel alleles are summarised in **Table 6.1**, and their sequences and nomenclature are highlighted in green in **Appendix 9**. Of these 27 distinct haplotypes, three (highlighted **Table 6.1**) were reported in Saudi Arabia [373]: rs999755320 (frequency 1/178), which is associated with the

rs1109037 amplicon [373]; rs922804818 (frequency 2/178), which is associated with the **rs1294331** amplicon; and rs543563536 (frequency 1/178), which is associated with the **rs891700** [373]. Generally, rare variants tend to result from recent mutations and are more likely to be associated with common ancestry between individuals who share the variant [384]. Therefore, rare variants are expected to cluster geographically or appear in isolated/specific populations [385]. The identification of shared variants between Saudi Arabian individuals and Kuwaiti individuals, who originate from populations that inhabit areas with geographic proximity, suggests that these rare variants in these two populations are identical by descent (IBD; sharing similar alleles inherited from a common ancestor) but the populations as a whole are not identical.

Table 6.1. Novel alleles associated with iiSNP in the Kuwaiti population.

Twenty-seven variants were identified as novel alleles by STRait razor among the sequencing data from 369 Kuwaiti individuals. The column labelled 'Rs Identifier' shows the identifier label for each variant. N/A indicates novel variants with no rs-number association. The highlighted SNPs were previously reported in a Saudi Arabian population [373]. The column labelled 'Count' indicates the number of times each variant was observed.

iiSNP amplicon	Variant motif	Rs Identifier	Frequency	Position (GRCh38)	Count
rs354439	CC insertion	N/A	Novel	chr13:106286115-106286116	1
	T>C	N/A	Novel	chr13:106285999	1
	T>A	rs894755414 (T>C)	Novel	chr13:106286068	1
	A>C	rs574448517	102/152234 (GnomAD)	chr13:106286032	1
rs826472	T>G	rs950907051-G	7/152214 (GnomAD)	chr10:2364375	1
	C>A	rs1588389080 (C>T)	Novel	chr10:2364426	1
	G>C	N/A	Novel	Chr10:2364376	2
	CTAT deletion	N/A	Novel	Chr10:2364427-2364430	2
rs891700	T insertion	rs543563536	53/29530 (GnomAD)	Chr1:239718608-239718609	6
	T deletion	rs543563536	11/29530 (GnomAD)	Chr1:239718609	1
	G>A	rs12047255	32/31344 (GnomAD)	Chr1:239718578	
rs1109037	A>G	rs1572313857	1/2 (PRJNA586841*)	chr2:9945623	1
	C>T	rs999755320	6/152132 (GnomAD)	chr2:9945592	2
rs279844	C>T	N/A	Novel	chr4:4632763846327691	1
	T>C	N/A	Novel	chr4:4632763846327708	3
rs735155	CT deletion	N/A	Novel	chr:3331999-3332000	2
rs1294331	A>G	rs922804818	4/152214 (GnomAD)	chr1:233312663	1
rs13182883	G>A	rs1336721677	4/125568 (PRJNA400167*)	chr5:137297590	1
rs1336071	CT deletion	rs540177118	26/152118 (GnomAD)	chr6:93827490-93827491	2
rs1979255	G>A	N/A	Novel	chr4:189396896	1
rs430046	T>G	rs190617353	178/152044(GnomAD)	chr16:77983173	1
rs4530059	G>A	rs999246876	3/152246 (GnomAD)	chr14:104302820	1
rs4606077	G>C	rs975955864	1/125568 (PRJNA400167*)	chr8:143574584	1
rs729172	C>T	rs562809565	1/31412 (GnomAD)	chr16:5556195	1
rs737681	C>G	N/A	Novel	chr7:156198071	1
rs740910	T>C	N/A	Novel	chr17:5803273	1
rs9951171	C>T	N/A	Novel	Chr18:9749874	1

* GenBank accession number.

As seen in **Table 6.1**, two targeted iiSNPs (**rs354439** and **rs826472**) were each associated with four variants identified as novel alleles by STRait Razor. The iiSNP **rs354439** (reverse-strand reported by UAS) was associated with: (1) a novel [CC] insertion at GRCh38-Chr13:106286115-106286116, (2) a novel [T>C] variant at GRCh38-Chr13:106285999, (3) a novel [T>A] variant at GRCh38-Chr13:106286068, identified as rs894755414 (dbSNP reports the T>C variant), and (4) an A>C variant reported as rs574448517 (GnomAD frequency 102/152,234). The iiSNP **rs826472** was associated with: (1) a T>G variant, reported as rs950907051 (GnomAD frequency 7/152,214); (2) a novel [C>A] variant, reported as rs1588389080 (dbSNP reports C>T with frequency 1/2,922 in a Korean population); (3) a novel [G>C] variant at GRCh38-Chr10:2364376; and (4) a novel [CTAT] deletion at GRCh38-Chr10:2364427-2364430. These additional variants informativeness to the targeted iiSNP, which can be useful in forensic identification and mixture deconvolution [386]. It must be noted that the sequencing data will be submitted to dbSNP if it meets the requirements.

6.4.5 Statistical analysis

Statistical analysis was performed after dividing the data into two datasets: alleles based on the targeted-iiSNP only and alleles based on the targeted-iiSNPs with flanking region variants. The observed and expected numbers of each genotype at each locus were compared, and *P*-values were obtained using a Fisher's Exact test in *Arlequin* software. Observed and expected levels of heterozygosity and the associated *P*-values are shown in **Appendix 10**. These HWE tests resulted in *P*-values >0.05 for all markers, except for six: rs1382387, rs1335873, and rs2107612 for both datasets, and rs907100 and rs727811 in the iiSNP dataset, and rs2111980 in the iiSNP-flanking dataset. When Bonferroni corrections were applied, the majority were found to conform with HWE expectations ($p > 0.00057$). However, a significant departure from HWE continued to be observed for rs1382387 ($P < 0.00057$). Deviation from HWE can be explained by an excess of homozygotes at this locus which could be associated with non-random mating due to consanguineous marriage, which is frequent among Kuwaitis, occurring at a rate of 54.3% [341]. Increased homozygosity could also indicate the existence of an ethnic/racial group that is associated with a specific genotype, population stratification, admixture, or genotyping errors [387]. A similar departure from equilibrium at the same locus was also reported in an Iranian population sample [388].

Analysis of molecular variance (AMOVA) was performed in Arlequin to estimate the level of genetic structure both among and within individuals, and to calculate F-statistics. The test showed that 97% of genetic variations occurred within Kuwaiti individuals, whereas the remaining 3% occurred among them. In addition, the average inbreeding coefficient between individuals was 0.02865, which indicated low levels of inbreeding. Associations between alleles across all possible pairwise combinations of loci were assessed using 3,828 linkage disequilibrium (LD) calculations, and no pair of loci showed significant LD after the application of the Bonferroni correction.

Among the 45 microhaps observed in this study, two (rs1109037; observed heterozygosity = 0.688 and rs907100; observed heterozygosity = 0.677) exceeded the heterozygosity of two autosomal STRs (TPOX and D17S1301; see **Chapter 4**) and exhibited only marginally smaller levels than D20S482 and D22S1045 (each with $H_o = 0.698$). These four STRs show very little to no sequence variation. In addition, sixteen microhaps were associated with heterozygosity of more than 0.5 compared to the heterozygosity when only considering the target SNP, whereas the remaining microhaps had heterozygosity below 0.5, even when more than two haplotypes were present. This result is due to the presence of rare haplotypes with low frequencies that do not contribute to the average heterozygosity of a locus. For example, rs430046 (expected heterozygosity = 0.47125) possess three haplotypes with frequencies of 0.37, 0.62, and 0.001.

To investigate the level of diversity within microhaplotypes, the effective number of alleles (A_e) was calculated for each microhaplotype relative to their expected heterozygosity. A_e represents the number of equally distributed allele frequencies that would generate the same expected heterozygosity (diversity). High A_e values are observed for loci with more similar frequencies between their alleles, which contribute to increased heterozygosity, compared with those featuring wide variation in the frequency of their alleles, such as those containing variants with low frequencies. This statistic reflects genetic variation within a locus; therefore, higher A_e values indicate microhaps that are more informative for differentiation between individuals, deconvolution of mixtures, and determination of the number of contributors in mixtures [378]. Twenty-two of the identified 45 microhaps had A_e values ≥ 2.00 . In addition, three (rs8037429, rs6811238, and rs1024116) of the target SNPs had A_e values = 2.00, which

is the maximum value for a diallelic locus. Overall, the microhap performance was associated with a noticeable elevation in Ae values, as shown in **Figure 6.4**.

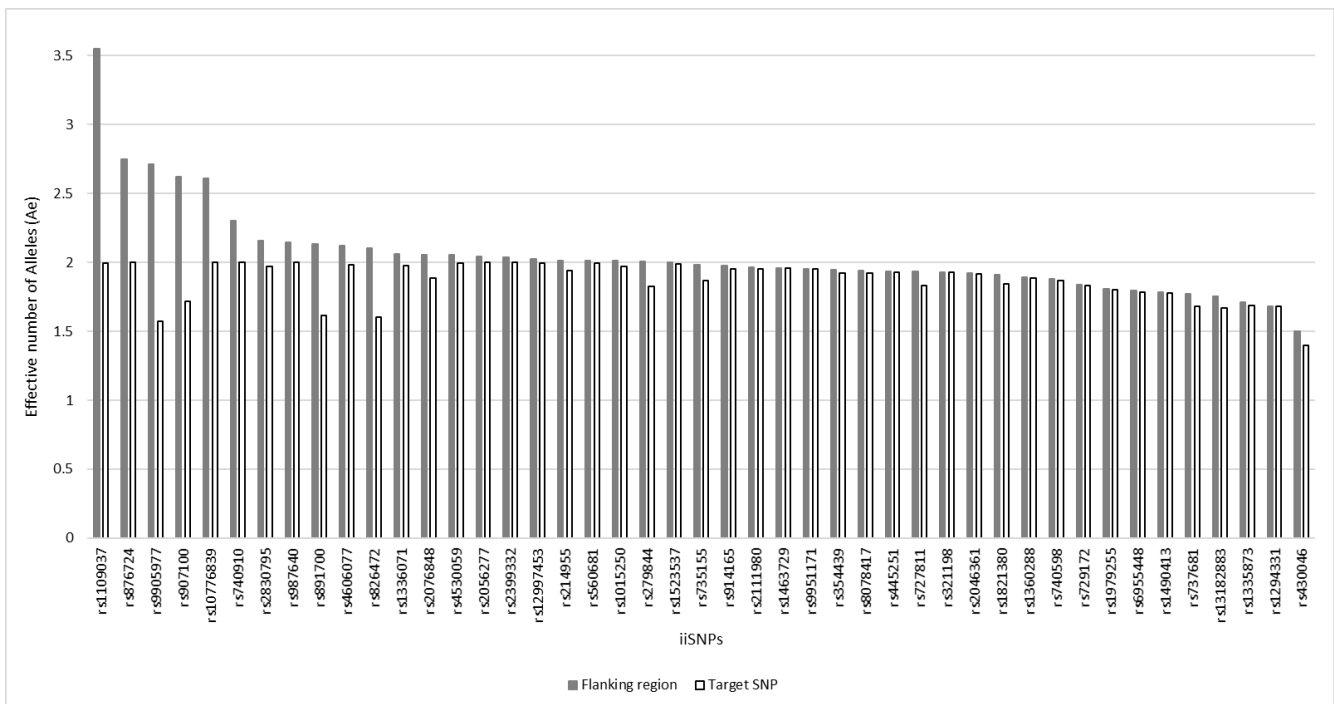


Figure 6.4. Effective number of alleles (Ae) for 45 iiSNP microhaplotypes.

The increase in Ae value observed for each of the 45 identified microhaplotypes, based on the inclusion of the flanking region (grey bars), compared with the values for the associated target iiSNP (white bars). The loci are presented in the order of the largest AE gained.

Forensically relevant parameters were calculated using STRAF for the two datasets. The forensic statistics for each marker are detailed in **Appendix 11**, and a summary for each dataset is provided in **Table 6.2**. As expected, the addition of the flanking sequence variants improved the values of all forensic indices compared with the target iiSNP alone, as shown in **Table 6.2**. The combined match probability for the 88 included iiSNPs was 2.736×10^{-36} , indicating that the likelihood of finding two identical iiSNPs profile was 1 in 4.20×10^{35} . When considering the variants observed in the flanking region, the CMP decreased to 1.87×10^{-39} . The dataset including the 88 iiSNPs and the flanking region variants was able to perform discrimination at a similar order of magnitude as the 28 autosomal ForenSeq™ DNA Signature Prep Kit flank-based (including flanking region variants) STRs (including SE33). In addition, the 45 microhaps alone yielded a CMP of 4.704×10^{-22} , which exceeded the CMP yielded by the typing of 15 STRs (Identifiler kit) among the Kuwaiti population (see **Table 5.6**).

Table 6.2. Summary statistics for forensic parameters among the Kuwaiti population.

Summary of the forensic statistics calculated for 88 iiSNPs and for the iiSNPs with flanking region variants among 369 Kuwaiti individuals.

Forensic parameter	iiSNPs	iiSNPs with flanking region
Average Heterozygosity	0.4495	0.4702
Average Polymorphism Information Content (PIC)	0.3544	0.3813
Average Typical Paternity Index (TPI)	0.9151	0.9588
Combined Match Probability (CMP)	2.38×10^{-36}	1.87×10^{-39}
Combined Power of Exclusion (PE)	0.99999943	0.999999926

Previous studies performed in several populations have shown differences in the match probabilities between estimates using target iiSNPs and those using iiSNPs and flanking region variants. For instance, King et al. reported the respective CPM values for 94 iiSNPs and the same iiSNPs with flanking region variants among Africans as 1.11×10^{-35} and 5.13×10^{-41} , among Hispanics as 1.98×10^{-37} and 5.42×10^{-40} , among Asians as 8.24×10^{-36} and 7.12×10^{-38} , and among Caucasians as 4.55×10^{-38} and 1.72×10^{-40} [265]. In a population regionally close to the Kuwaiti population, Al Khubrani et al. reported CMP values of 9.97×10^{-37} (iiSNPs) and 8.88×10^{-40} (iiSNPs with flanking region) for 91 iiSNPs in the Saudi Arabian population [373]. In this study, the Kuwaiti population yielded comparable match probabilities (2.38×10^{-36} and 1.87×10^{-39}) for only 88 iiSNPs highlighting that these markers can be successfully used in forensic applications in Kuwait.

The ForenSeq™ DNA Signature Prep Kit contains both STRs and iiSNPs, and combining both sets of markers is likely to substantially increase the overall informativeness of this analysis, decreasing the match probabilities by several orders of magnitude. Prior to calculating these statistics for the combined set of SNP and STR markers using the product rule, the associations between all possible pairwise combinations of alleles were examined across the different loci. No pairs of loci showed significant LD after applying the Bonferroni correction ($p > 0.0001$), which indicated that all loci are statistically independent and can, therefore, be used to calculate the genotype frequencies among the Kuwaiti population.

The CMP obtained for the combined 28 autosomal STRs (including SE33) and 88 independent iiSNPs was estimated as 3.6675×10^{-70} . Although this represents a significant increase in discrimination power, the discrimination of good-quality or single-source samples can be performed adequately using PCR-CE typing systems. Therefore, increasing the number of

analysed markers is not a useful approach for the analysis of such samples. However, when evaluating degraded samples of low quality or low quantity, STR loci may not be fully typed, resulting in partial/no STR profiles for both the current PCR-CE and the improved MPS systems. Therefore, the inclusion of SNPs and their additional variants can be beneficial for increasing the overall discrimination power for degraded/low-quality DNA samples. In the present study, the CMP was estimated at 1.87×10^{-39} for the 88 iiSNPs (43 iiSNPs with 3.975×10^{-18} and 45 microhaps with 4.704×10^{-22}), and so even a partial iiSNP profile is likely to be able to generate significant discrimination power.

6.4.6 Population differentiation

To illustrate the genetic relationship between the Kuwaiti population and a number of previously reported global populations, pairwise F_{ST} genetic and STRUCTURE analyses were performed, based on genotypic data for the 87 iiSNPs that are shared between the Kuwaitis surveyed in this study, and seven major population groups: Africa (AFR; N = 113), America (AMR; N = 63), Central South Asia (CSA; N = 199), the Middle East (ME; N = 161), Europe (EUR; N = 156), East Asia (EAS; N = 229), Oceania (OCE; N = 30) and Saudi Arabia (KSA; N = 89).

STRUCTURE analysis was performed to investigate the distribution patterns of genetic variants among populations. Although nine populations/population groups were involved in the analysis, the results indicated that the most likely number of distinct populations (K) was three, based on the method described in Evanno et al. [316]. As shown in **Figure 6.5**, at $K = 3$, Kuwait showed a similar genetic distribution pattern as those observed for Europe, the Middle East and Saudi Arabia. However, at $K = 4$, Kuwait appeared slightly differentiated from Europe, while continuing to share distribution patterns with the Middle East and Saudi Arabia. In addition, Kuwait was distinguishable from Africa, Central South Asia, East Asia, Oceania, and America.

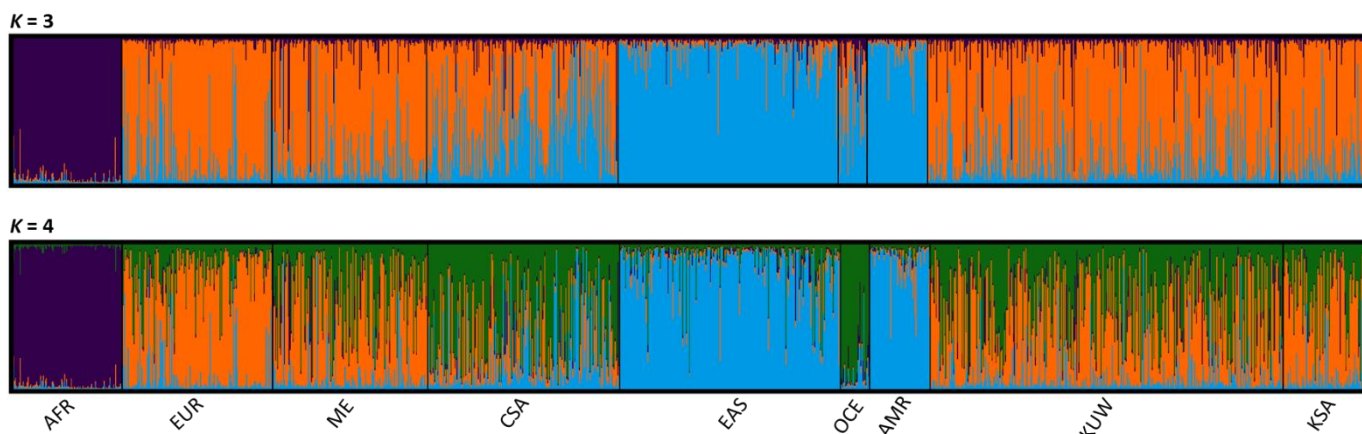


Figure 6.5. Analysis of the structures of nine population groups using the 87 iiSNP loci.

Assessment of population differentiation using the Bayesian-based approach implemented in *STRUCTURE* software. The number of colours in each bar represents the number of assumed populations ($K = 3$ and $K = 4$). Each vertical line represents an individual, and the amount of each colour in each line represents the proportion of that individual's genetic ancestry that has been assigned to each of the K populations.

The pairwise F_{ST} heatmap matrix between the Kuwaiti population and the other global populations for the 87 iiSNPs is shown in **Figure 6.6**, and the corresponding values are provided in **Appendix 12A**. The results revealed that Kuwaitis were genetically more closely related to individuals from the Middle East and Saudi Arabia, with pairwise F_{ST} values of 0.00183 and 0.00348, respectively. The close genetic relationship between Kuwaitis and Saudi Arabians is expected because these populations are geographically, culturally, and religiously connected, which is likely to accelerate gene flow between them. However, the Middle Eastern group, which contains Mozabite, Druze, Palestinian, and Bedouin groups, was even more closely genetically related to the Kuwaiti population than the Saudi Arabian population, indicating that they share similar allele frequencies, as supported by their low pairwise F_{ST} .

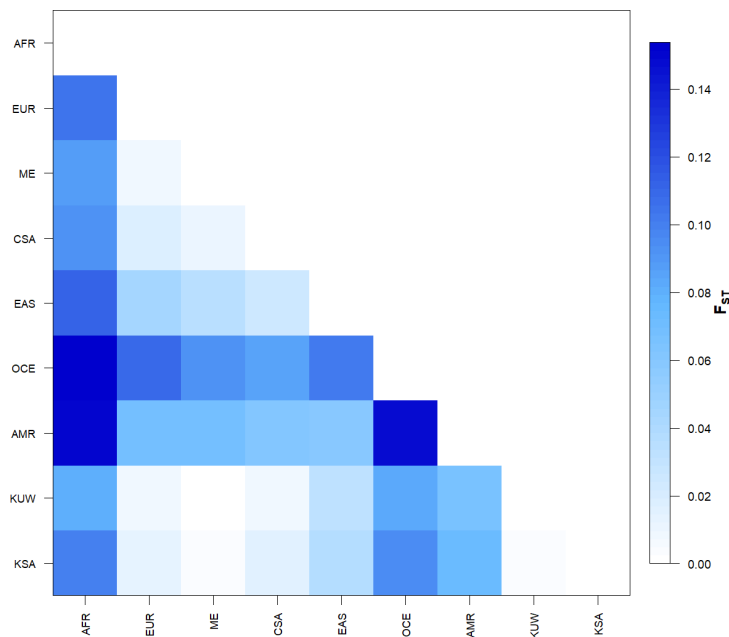


Figure 6.6. F_{ST} -based genetic distance matrix among global population groups.

Heatmap depicting pairwise F_{ST} distances based on the genotypes of 87 iiSNPs, among individuals from Kuwait (KUW), Saudi Arabia (KSA) and seven global population groups: Africa (AFR), America (AMR), Central South Asia (CSA), the Middle East (ME), Europe (EUR), East Asia (EAS), and Oceania (OCE), from a 51 HGDP-CEPH population panel. Colour intensities correlate with the F_{ST} values, which are tabled in **Appendix 12A**.

To further investigate the genetic distance between the Kuwaiti and Middle East populations, pairwise F_{ST} evaluation was performed, based on the genotypes of 87 iiSNP loci, between five Middle East populations: Palestinians ($N = 46$), Druze ($N = 42$), Bedouin ($N = 46$), Mozabite ($N = 27$), and Saudi Arabia ($N = 89$). Kuwait showed a lower genetic distance from the Palestinian ($F_{ST} = 0.00391$), Druze ($F_{ST} = 0.00442$), and Bedouin ($F_{ST} = 0.00403$) populations than from the Mozabite ($N = 27$, $F_{ST} = 0.00992$) population. The pairwise F_{ST} heatmap matrix between the Kuwaiti population and these Middle East populations for the 87 iiSNPs are shown in **Figure 6.7**, and corresponding values are provided in **Appendix 12B**. The small F_{ST} values between Kuwaitis and these examined populations agrees with the results of a previous study reported by Alsmadi et al. [289], further supporting the likelihood that these populations share a common ancestry.

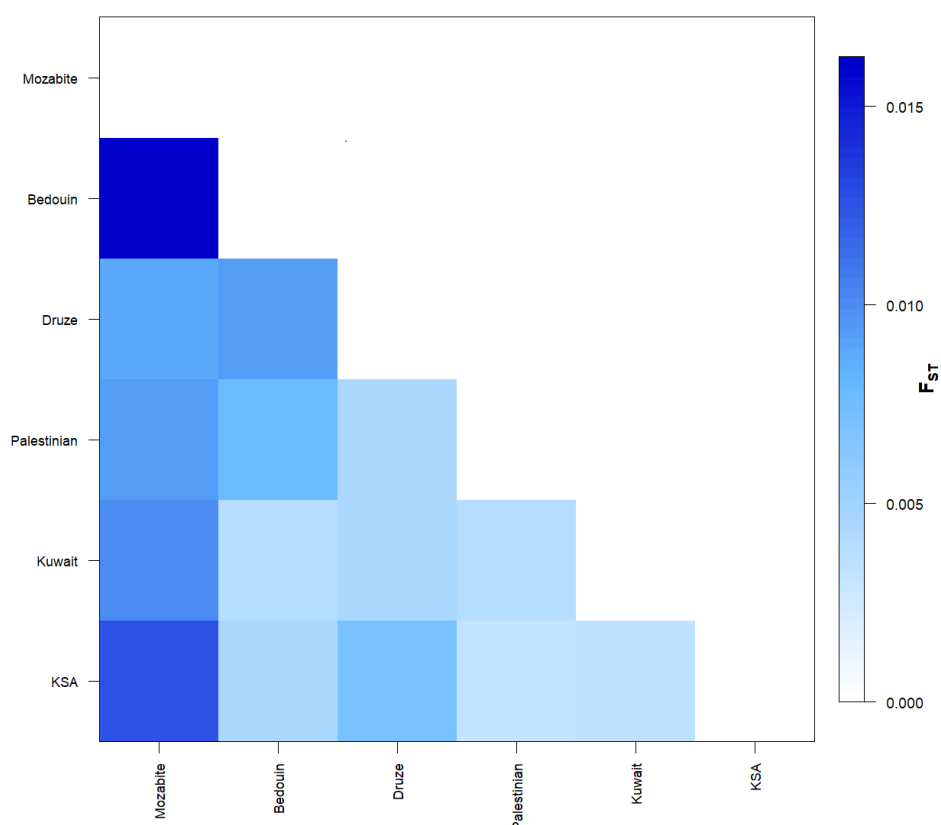


Figure 6.7. F_{ST} -based genetic distance matrix among Middle East populations.

Heatmap depicting pairwise F_{ST} distances, based on genotypes at 87 iiSNPs, among individuals from Kuwaiti (KUW), Saudi Arabian (KSA), Palestinian, Druze, Bedouin, and Mozabite populations. Colour intensities correlate with F_{ST} values, which are tabled in **Appendix 12B**.

6.5 Conclusion

This chapter represents the first study to evaluate the performance of forensically relevant SNP markers, the 94 iiSNP markers included in the ForenSeq™ DNA Signature Prep Kit, in the Kuwaiti population. The results showed that some loci had better performance than others in terms of Depth of Coverage and Allele Coverage Ratio. Only one marker deviated from the expectations of HWE after Bonferroni correction, due to an observed excess of homozygotes. The results also showed no significant LD across all possible pairwise locus combinations, indicating that the product rule can be used to combine allele frequencies among the Kuwaiti population. The sequencing uncovered variants within the flanking regions of the iiSNPs, which increased the informativeness of the single base target SNP. As a result, the SNP-microhaplotype system is likely to be an efficient tool for discriminating between individuals, and for the analysis of low-quality, low-quantity DNA samples. Although the set were not

markers that were designed to be Ancestry Informative Markers, the SNPs studied were able to differentiate the Kuwaiti population from populations with distant geographical locations, and demonstrated a close genetic relationship between Kuwait and other Middle Eastern populations, which is likely to be due to common shared ancestry.

Chapter 7 Analysis of sex chromosome STRs in the Kuwaiti population.

7.1 Introduction

One of the foremost potential values in MPS technology is the wide range of genetic markers that can be combined in a single multiplex, allowing the analysis of multiple marker types in a single sample. Along with the autosomal SNPs and STRs that were discussed in the previous chapters, the ForenSeq™ DNA Signature Prep Kit also includes 24 Y-STRs and 7 X-STRs. The addition of the sex markers can be potentially useful in forensic cases where autosomal analysis is inconclusive. For example, the Y-chromosome can identify the male contributor in sexual assault cases, and this can be especially useful in cases involving a male/female mixed DNA sample where the male autosomal profile is often masked by the female due to the high proportion of female DNA, which can cause preferential amplification [389], or to the quantity of male DNA being low. In addition, paternal transmission of the Y chromosome allows identification of the patrilineality of family lineages, which can be applied to specific scenarios in paternity investigations such as cases where the father/mother is deceased/not available for testing. In addition, Y-STRs can be informative in tracking the genetic relationship between two or more males. Y-STR haplotypes can also be used to predict haplogroups. Haplogroups represent lineages defined by Y-SNPs, which differ significantly in frequencies between geographic regions [157]. This allows a strong association of a region with the Y-SNP haplogroup that defines it, and this can also be inferred from its associated Y-STR haplotype [157]. Therefore, the prediction of haplogroups from Y-STR haplotypes can be informative for the inference of the biogeographical area of origin of individuals.

The X-chromosome can also be detected in male/female mixed stains, however in this scenario, the analysis of the X-chromosome is more informative when the male DNA is in excess of the female, causing the loss of the female autosomal profile [198]. Analysis of the X chromosome can also be beneficial in complex paternity cases such as parent/child relationship investigations that involve at least one female individual, such as mother/son, mother/daughter, and father/daughter [197].

7.2 Aims and objectives

The main aim of this chapter is to evaluate the forensic performance of the 7 X-STRs contained within the Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics system, in the Kuwaiti population. The X-STR markers, along with their sequence variation, will be reported for the first time in the Kuwaiti population. A second aim is examining the concordance of the Y-STR markers collected using different technologies (MPS and PCR-CE), and to predict the haplogroups from the Y-STR haplotypes. This study involves:

- Generating allele frequency data for the X-STR makers and their haplotypes to increase the amount of genetic data available for the Kuwaiti population.
- Cataloguing the underlying genetic variation in the sequenced X-STR markers in the Kuwaiti population.
- Generating Y-STR profiles using the Promega PowerPlex® Y23 System for 248 Kuwaiti males.
- Comparing the Y-STR profiles generated by CE-based genotyping (Promega PowerPlex® Y23) with those generated by sequencing (ForenSeq™ DNA Signature Prep Kit), to examine concordance between the two systems.
- Identifying the main Y-chromosome haplogroups in the Kuwaiti population, using the Y-STR haplotypes.

7.3 Materials and methods

The X-STR and the Y-STR sequencing datasets analysed in this chapter were obtained from the data generated from 376 samples using the Verogen MiSeq FGx® Forensic Genomics system with the ForenSeq™ DNA Signature Prep Kit.

7.3.1 Y-STR PCR-CE genotyping

The Promega PowerPlex® Y23 PCR amplification kit was used to generate Y-chromosome haplotypes from 248 unrelated Kuwaiti males for the following 23 STRs: DYS576, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385a/b, DYS456, and Y-GATA-H4. DNA was amplified directly from a 1.2 mm FTA® card punch, according to the instructions given in the PowerPlex® Y23 manual, using a SureCycler 8800 thermal cycler (Agilent Technologies, CA, USA) as described in section 2.7.1. Detection and separation of the

DNA fragments were carried out using an Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific) with the internal lane standard WEN ILS 500 Y23 and allelic ladder provided with the PowerPlex® Y23 kit as described in section 2.7.2. New POP-4™ polymer, anode/cathode buffers and a 36 cm capillary array were installed. Genotype determination and allele calling for the 23 Y-STR loci were carried out using *GeneMapper™ ID-X* software version 1.4 (Thermo Fisher Scientific).

7.4 Data analysis

Sequencing data were initially analysed using the graphical interface program, ForenSeq™ Universal Analysis software (UAS), version 1.3. The default analysis settings for the analytical and interpretation threshold were used as recommended by the software recommendation (**Appendix 13**, see 5.3.5 for more details). Positive and negative controls for Y- and X-STRs were all verified after the completion of each run. The flanking sequence report option, implemented in the UAS, was used to generate the flanking region sequences for all samples.

The Y-chromosome STR genotypes called from the UAS were independently analysed for concordance with the genotypes generated using the PowerPlex® Y23 kit. Haplogroup prediction was carried out using NevGen Haplogroup predictor (desktop version) implemented in NevGen Genealogy Tools version 1.1 [390], which is based on a Bayesian approach [190].

The program StatsX (version 2.0) was used to calculate forensic statistics for the X-STRs including the power of discrimination in males (PDM) and females (PDF), and polymorphism information content (PIC). In addition, mean exclusion chance (MEC) in standard trios involving daughters was calculated according to Desmarais et al. [391] (MEC_Desmarais) and Kishida et al. [392] (MEC_Kishida), in father/daughter or mother/son duos according to (MEC_Desmarais_Duo), and in deficiency cases involving mother, daughter, paternal grandmother according to Krüger et al. [393] (MEC_Krüger). An Exact test for population differentiation was carried out in Arlequin, using the allele frequencies of males and females per locus.

7.4.1 Bioinformatics analysis

FASTQ files were further analysed using STRait Razor, version 3.0 [306], and were mapped to the GRCh38 reference genome. This was performed to examine the concordance of the allele designations between UAS, which uses an alignment-based method, and STRait Razor, which uses the haplotype capture method, to assign the correct nomenclature to each SNP allele found in the STR flanking region, according to ISFG recommendations [266]. Where no corresponding SNP/variant was identified in the STRait Razor database, the allele was designated as 'NovelSeq'. The FASTQ files for the samples containing novel alleles were manually aligned to the GRCh38 genome using BWA (Burrows-Wheeler Aligner) [307], version 0.7.15-r1140, and the resulting bam files were visualised using IGV (Integrative Genomics Viewer), version 2.8.0 [308]. This provided a view of the variant locations, which were then identified using Variation Viewer, 2.1.4 [371], and labelled with the appropriate variant code (rs-number), if available.

7.5 Results and discussion

7.5.1 PowerPlex® Y23 kit haplotype genotyping

The Y23 kit uses primers designed to co-amplify each of the 23 markers labelled with one of four dyes, as illustrated in **Table 7.1**. A total of 248 male samples were analysed using the Applied Biosystems 3500 Genetic Analyzer, which separates amplicons according to their size and arranges them based on the dye colour emitted. Peak size is determined by a calibration curve, which is produced from internal size standard fragments of known size; the internal size standard used here was WEN 500 Y23, which contains 21 fragments (**Figure 7.1**). Alleles are determined by comparing peaks with an allelic ladder, which contains all the possible alleles for each locus (**Figure 7.2**). No sign of contamination was detected in any of the negative controls. The positive control DNA 9948 profile (**Figure 7.3**) was concordant with the positive control given in the manufacturers protocol. The genotyping resulted in full profiles for all 248 Kuwaiti males using the Y23 kit.

Table 7.1. The PowerPlex® Y23 System dye-labels and allelic Ladder Information.

STR Locus	Dye-Label	Size Range of Allelic Ladder Components (bases)	Repeat Numbers of Allelic Ladder Components	Repeat Sequence 5'→ 3'
DYS576	Fluorescein	97–145	11–23	AAAG
DYS448	Fluorescein	196–256	14–24	AGAGAT
DYS389I	Fluorescein	147–179	9–17	(TCTG)
DYS389II	Fluorescein	259–303	24–35	(TCTA)
DYS19	Fluorescein	312–352	9–19	TAGA
DYS391	JOE	86–130	5–16	TCTA
DYS481	JOE	139–184	17–32	CTT
DYS549	JOE	198–238	7–17	GATA
DYS533	JOE	245–285	7–17	ATCT
DYS438	JOE	293–343	6–16	TTTTC
DYS437	JOE	344–380	11–18	TCTA
DYS570	TMR-ET	90–150	10–25	TTTC
DYS635	TMR-ET	150–202	15–28	TSTA compound
DYS390	TMR-ET	207–255	17–29	(TCTA) (TCTG)
DYS439	TMR-ET	263–307	6–17	AGAT
DYS392	TMR-ET	314–362	4–20	TAT
DYS643	TMR-ET	368–423	6–17	CTTTT
DYS393	CXR-ET	101–145	7–18	AGAT
DYS458	CXR-ET	159–215	10–24	GAAA
DYS385a/b	CXR-ET	223–307	7–28	GAAA
DYS456	CXR-ET	316–364	11–23	AGAT
Y-GATA-H4	CXR-ET	374–414	8–18	TAGA

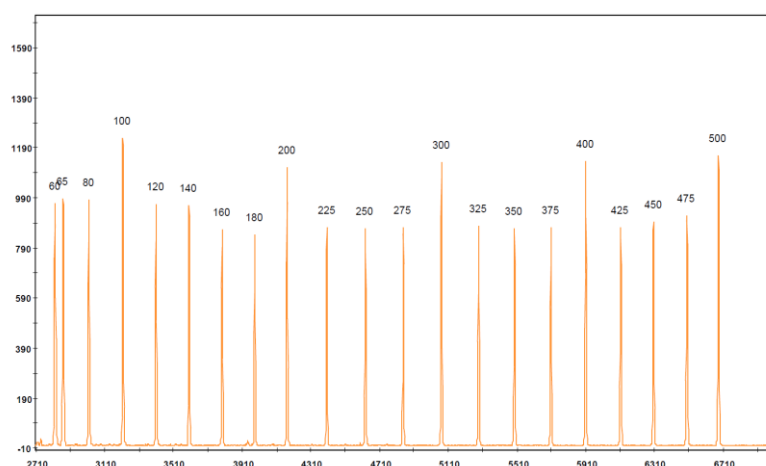


Figure 7.1 Electropherogram of the internal size standard (WEN 500 Y23).

Twenty-one fragments ranging from 60 bp to 500 bp are separated according to their length.

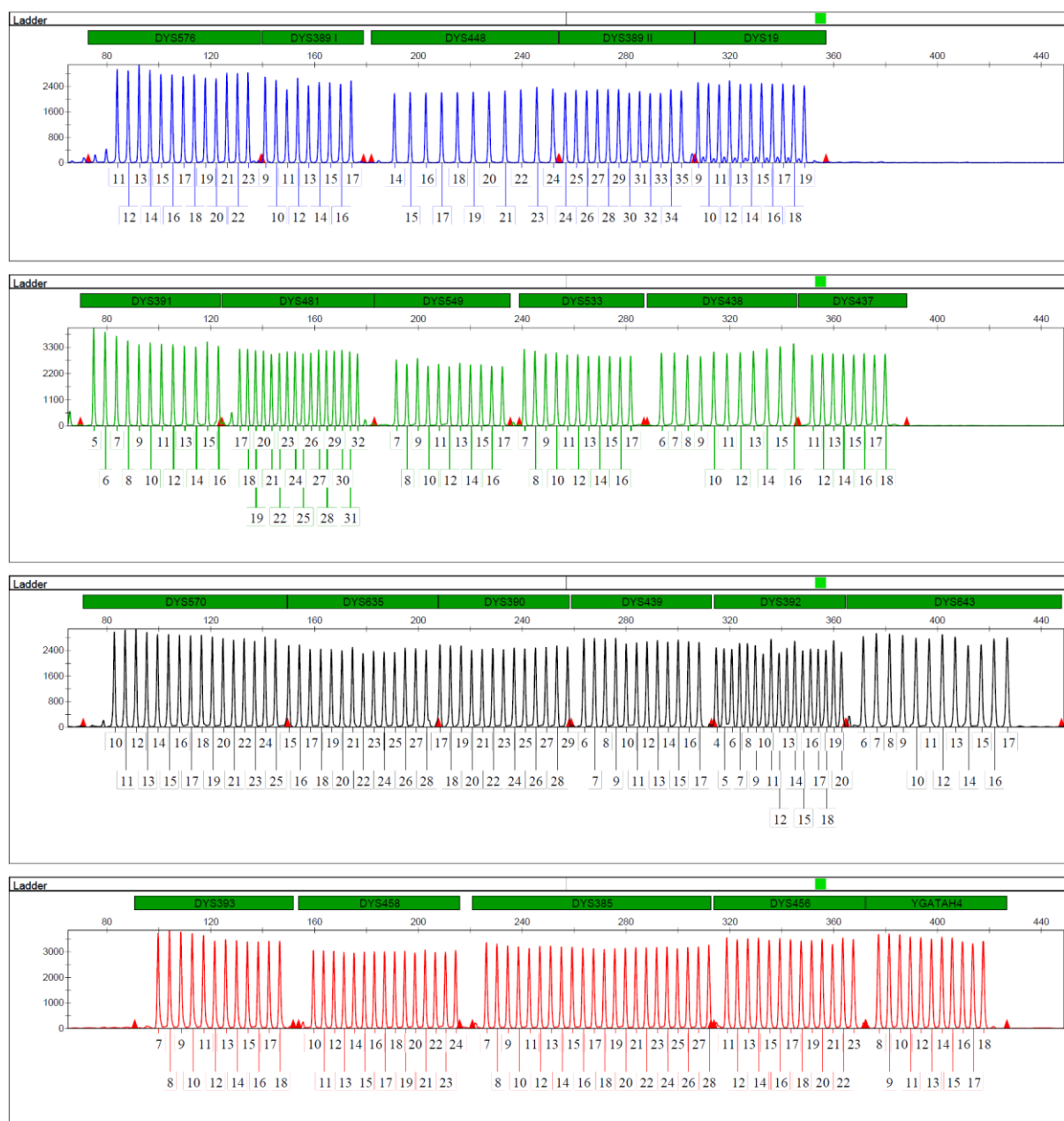


Figure 7.2. Electropherogram of the Y23 allelic ladder.

The allelic ladder contains all possible alleles for each locus of the Y23 kit.

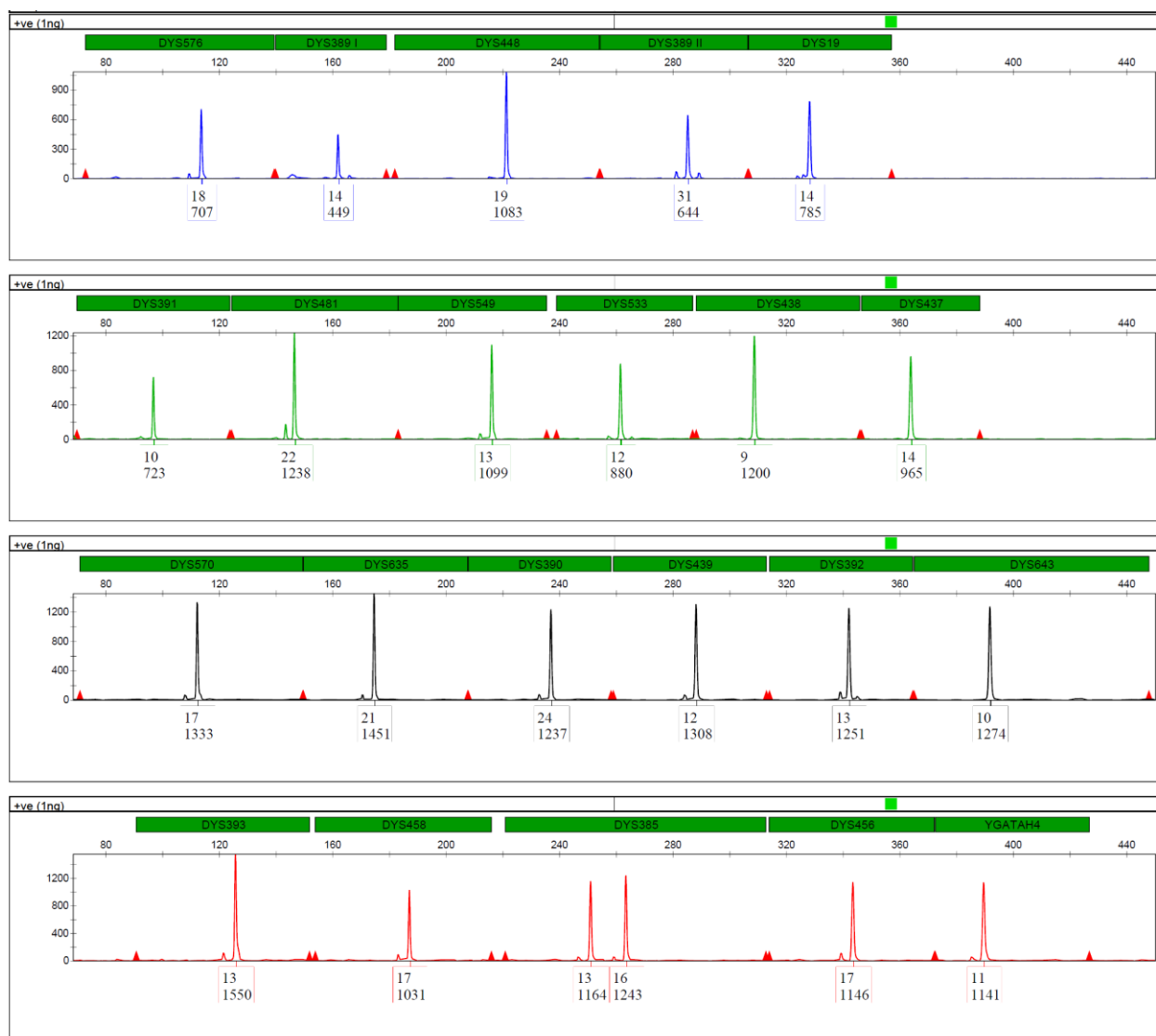


Figure 7.3. Electropherogram of the Y23 positive control.

The positive control profile generated from 1 ng of the male DNA 2800 sample.

7.5.2 Sequencing performance of X and Y-STRs

Sequencing depth, or depth of coverage (DoC), is the number of high-quality aligned sequencing reads generated from a target genomic region. The average DoC for individual Y-STR markers was within the range ~42X - ~860X and for the X-STR within the range of ~20X - ~6000X. As shown in **Figure 7.4**, The X-STR DXS10103 and the Y-STR Y-GATA-H showed the lowest coverage among all the sex-chromosomal loci, while the Y-STRs DYS438 and DYS389I exhibited the highest coverage.

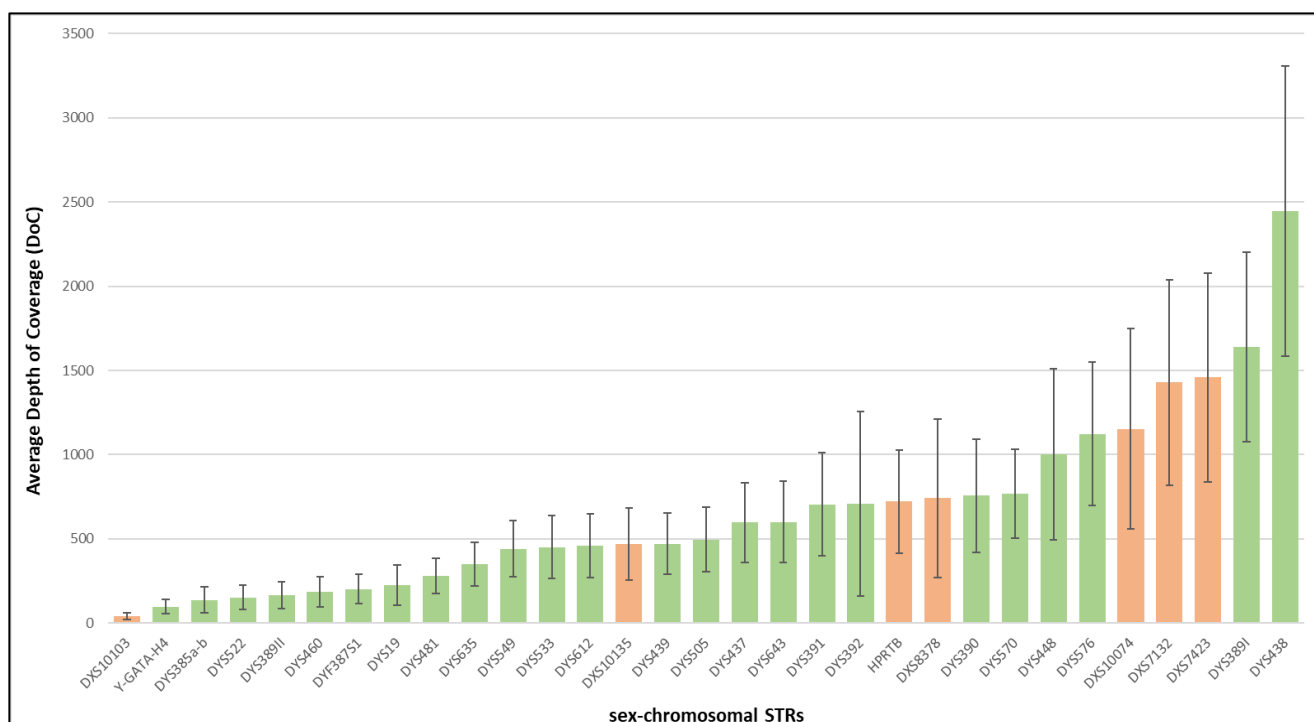


Figure 7.4. Average DoC for the ForenSeq™ DNA Signature Prep kit sex-chromosome STR loci.

Average DoC for sex-chromosome STRs across the four runs performed using the ForenSeq™ DNA Signature Prep Kit. The green bars represent Y-STR loci and the orange bars represent X-STR loci.

7.5.3 X-STR allele frequencies and variation.

Each sequencing run contained 94 samples, plus a negative control and a positive control (2800M). This resulted in 376 samples, of which 371 (248 males and 123 males) were included in this study and five were removed due to the presence of extra alleles, likely to be a result of contamination or low coverage. The negative controls did not show any contamination. The positive control was in concordance with the positive control profile provided in the ForenSeq™ DNA Signature Prep Kit manual (the positive control DNA sample 2800M genotypes are shown in **Figure 5.2**). Allele frequencies were calculated for the seven X-STR markers in the 371 Kuwaiti individuals, based on three different classifications: alleles based on length variation, alleles based on STR repeat region sequence variation and alleles based on flanking region sequence variation. These frequencies were produced by counting the number times each individual allele was observed at a locus, divided by the number of alleles surveyed at a locus. **Appendix 14** shows the allele frequencies of the seven X-STRs in the 371

Kuwaiti individuals based on the three classifications described above, and nomenclature based on ISFG recommendations [266].

An overall increase in the number of alleles was observed at six of the seven X-STR loci when sequence variation was considered (**Figure 7.5**). A total of 74 additional alleles were observed in the sequencing data, consisting of 67 repeat sequence variant alleles and seven flanking region sequence variant alleles, among the six X-STRs, whereas no gain of additional alleles was observed at DXS7423. The largest gain was observed at DXS10135, for which 45 additional repeat sequence variants were observed (**Figure 7.5**). The increase in the number of allele variants in both the STR repeat sequence region and the flanking sequence region will substantially increase the discrimination power of these markers for forensic testing in this population.

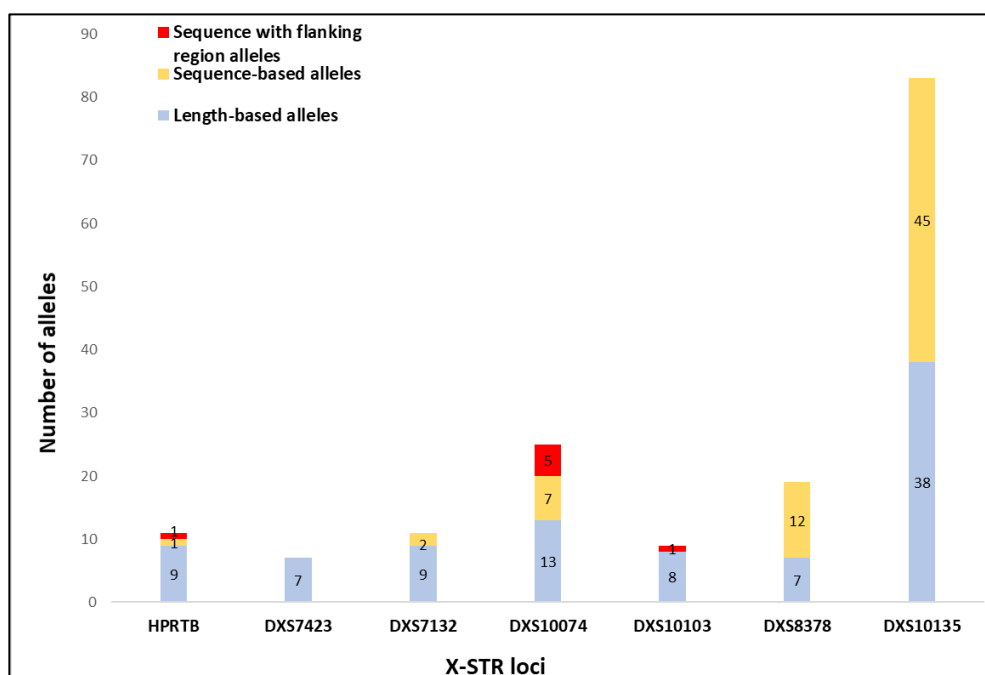


Figure 7.5. Increase in allele number at autosomal STRs as a result of sequence variants. Number of alleles observed using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx® sequencing data in 371 Kuwaiti individuals.

A total of 36 unique allele sequences distributed among six of the seven X-STRs were identified by STRait Razor as novel alleles. However, 12 of these alleles were found in either GenBank STRseq [372], or in the literature [272]. Novel alleles that have not been reported previously are summarised in **Table 7.2**, and their sequences and nomenclature are highlighted in green in **Appendix 14**.

Table 7.2. Novel X-STR sequence variants.

Twenty-four novel X-STR sequence variants observed in 371 Kuwaiti individuals. Sequence nomenclature is according to ISFG recommendations [266]. The column labelled 'Count' indicates the number of times each allele was observed, and the final column indicates whether the sequence variant was observed in either the repeat region sequence [RS] or the flanking region sequence [FS].

Allele nomenclature	Count	Type
HPRTB [CE 11]-GRCh38-ChrX-134481429-134481588 [ATCT]11 rs761894372 G>A; 134481502-134481513 DEL	1	FS
DXS7132 [CE 12]-GRCh38-ChrX-65435623-65435778 [TAGA]2 TAAA [TAGA]9	1	RS
DXS7132 [CE 13.3]-GRCh38-ChrX-65435623-65435778 [TAGA]9 TGA [TAGA]4	1	RS
DXS7132 [CE 14]-GRCh38-ChrX-65435623-65435778 [TAGA]12 TAAA TAGA	1	RS
DXS7132 [CE 19]-GRCh38-ChrX-65435623-65435778 [TAGA]19	1	RS
DXS10074 [CE 8]-GRCh38-ChrX-67757300-67757464 [AAGA]8 67757401-67757404 DEL	1	FS
DXS10074 [CE 14]-GRCh38-ChrX-67757300-67757464 [AAGA]13 AAGG	1	RS
DXS10074 [CE 15]-GRCh38-ChrX-67757300-67757464 [AAGA]12 AAGG [AAGA]2 rs774577988 T>A	1	FS
DXS10103 [CE 18]-GRCh38-ChrX-134284939-134285040 TAGA TTGA CAGA [TAGA]9 [CAGA]6 TAGA	1	RS
DXS8378 [CE 13]-GRCh38-ChrX-9402182-9402373 [ATAG]13 9402309 C>T	2	FS
DXS10135 [CE 16.2]-GRCh38-ChrX-9338302-9338453 [AAGA]16 AA GAAA rs201630737	1	FS
DXS10135 [CE 20]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]13 [AAGG]2 AAGA AAAG	1	RS
DXS10135 [CE 21.1]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]9 A [AAGA]6 AAGG AAGA AAAG	1	RS
DXS10135 [CE 21.1]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]18 A AGGA rs201630737	2	FS
DXS10135 [CE 22.2]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]15 AA [AAGA]3 AAAG	1	RS
DXS10135 [CE 24.2]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]5 GA [AAGA]13 AAGG AAGA AAAG	1	RS
DXS10135 [CE 24.3]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GA [AAGA]21 AAAG	1	RS
DXS10135 [CE 26]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]18 AGGG [AAGG]2 AAGA AAAG	1	RS
DXS10135 [CE 26]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]15 [AAGG]2 [AAGA]3 AAGG AAGA AAAG	1	RS
DXS10135 [CE 27]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]13 AAGG [AAGA]3 AAGG [AAGA]3 AAGG AAGA AAAG	1	RS
DXS10135 [CE 27.1]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]24 AAAG rs201630737	1	FS
DXS10135 [CE 30]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]13 [AAGG]12 AAGA AAAG	1	RS
DXS10135 [CE 34]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]17 [AAGG]12 AAGA AAAG	1	RS
DXS10135 [CE 37]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]13 [AAGG]19 AAGA AAAG	1	RS

7.5.4 Null allele at the HPRTB X-STR

It was observed that UAS version 1.3 called one allele at the HPRTB locus in one female individual, whereas STRait Razor revealed two alleles. When visualising the sequenced sample, the IGV showed deletion and SNP variants, as illustrated in **Figure 7.6**. A frameshift mutation altering the way the sequence read might be the reason behind the miscalled allele by the UAS. In addition, STRait razor designated the allele as 10, however, after counting the tetramer repeats (ATCT), 11 were seen and therefore it was designated as an allele 11. The G > A SNP at ChrX:134481499 (GRCh38.p12) observed in this allele was identified as rs761894372, which is rare according to the GnomAD database (frequency 1/18470) and the 1000 Genomes database (frequency 12/3775) and has mostly been reported in African

individuals from these databases. It is worth mentioning that the individual possessing the SNP in this study is an African-Kuwaiti female.

[illegible]

Figure 7.6. Illustration of a sequence variant at HPRTB locus.

A visualisation of 12 bp upstream of six tetramer repeats (ATCT) in part of an HPRTB allele in a Kuwaiti female, sequenced using the ForenSeq™ DNA Signature Prep Kit/MiSeq FGx®. Highlighted in black are the sequence variants observed in corresponding to the reference sequence.

7.5.5 X-STR population genetic and forensic statistics.

Statistical analysis was performed after dividing the data into two datasets containing male and female genotypes. For the female dataset, observed and expected numbers of each genotype at each locus were compared, and P-values were obtained using a Fisher's Exact test in Arlequin software. Observed and expected levels of heterozygosity and the associated p-values are shown in **Appendix 15B**. No statistically significant deviation from Hardy Weinberg Equilibrium was observed at any of the loci (p-values > 0.05). Association between alleles at all possible pairwise combinations of loci was evaluated using the linkage disequilibrium test for both males and females. Significant linkage disequilibrium was detected between one (of a total of 21) pairs in females and five pairs of loci in males (p < 0.05). However, after Bonferroni correction of the significance level using the number of tests ($0.05/21 = 0.0238$), significant linkage disequilibrium was observed only for three pairs of loci in males: (DXS7423, DXS10135), (HPRTB, DXS10103) and (DXS7423, DXS10135).

Prior to pooling the X-STR allele frequencies for male and female individuals in order to calculate forensic parameters, an Exact test for population differentiation was carried out in ARLEQUIN using the allele frequencies in males and females per locus. No significant difference was found between the frequencies in males and females ($P > 0.05$; **Appendix 15A**).

Forensic parameters calculated using pooled allele frequencies for the seven X-STRs in the Kuwaiti population are shown in **Table 7.3**. The most polymorphic locus was DXS10135 (PIC= 0.9441) followed by DXS10074 (PIC= 0.8357) while DXS7423 showed the lowest level of polymorphism (0.6257). The combined power of discrimination was observed to be higher in

females (PDF) than males (PDM), due to the di-allelic nature of X-STRs in females, meaning that the probability of discriminating two unrelated females is higher than for two unrelated males, using these seven X-STRs. The probabilities of excluding a non-true trio or duo relationship were calculated using mean exclusion chance (MEC), which yielded high values, as shown in **Table 7.3**.

Table 7.3. Forensic parameters calculated using pooled male and female allele frequencies for the Kuwaiti population.

GD: genetic diversity (expected heterozygosity); PIC: polymorphism information content; PDM: power of discrimination in males; PDF: power of discrimination in females; MEC_Krüger: mean exclusion chance according to Krüger et al.; MEC_Kishida: mean exclusion chance according to Kishida et al, al. MEC_Desmarais: mean exclusion chance in standard trios involving daughters according to Desmarais et al.; MEC_Desmarais_Duo: mean exclusion chance in father/daughter or mother/son duos according to Desmarais et al.

Forensic Parameters	HPRTB	DXS8378	DXS7423	DXS7132	DXS10135	DXS10074	DXS10103	Combined
Alleles	9	8	7	9	38	13	7	-
GD	0.7518	0.6883	0.6848	0.7537	0.9487	0.8543	0.7209	-
PIC	0.7103	0.6257	0.6276	0.7114	0.9441	0.8357	0.6833	-
PDM	0.7502	0.6868	0.6834	0.7521	0.9467	0.8525	0.7193	0.9999864653
PDF	0.8977	0.8408	0.8440	0.8978	0.9946	0.9614	0.8851	0.9999999937
MEC_Kruger	0.5265	0.4174	0.4298	0.5250	0.8927	0.7060	0.5008	0.9988231049
MEC_Kishida	0.7104	0.6257	0.6276	0.7112	0.9447	0.8358	0.6834	0.9999664578
MEC_Desmarais	0.7103	0.6257	0.6276	0.7114	0.9441	0.8357	0.6833	0.9999661040
MEC_Desmarais_duo	0.5731	0.4796	0.4829	0.5739	0.8969	0.7310	0.5417	0.9993778208

The seven X-STR loci were categorised into four linkage groups (LG) based on their chromosomal localisation: LG1: DXS8378-DXS10135, LG2: DXS7132-DXS10074, LG3: DXS10103-HPRTB, and LG4: DXS7423. Each pair of two loci is considered as one haplotype for the genotyping of males. The observed number of unique haplotypes at each LG in the Kuwaiti population were: 79, 49, 31 and 6 for LG1, LG2, LG3, and LG4, respectively. The most informative linkage group in the Kuwaiti samples was LG1 (PIC = 0.9782) with 79 unique haplotypes observed. Allele frequencies for the LG haplotypes are shown in **Appendix 14** and their forensic parameters are shown in **Table 7.4**.

Table 7.4. Forensic parameters calculated using male haplotype frequencies for the Kuwaiti population.

GD: genetic diversity (expected heterozygosity); PIC: polymorphism information content; PDM: power of discrimination in males; PDF: power of discrimination in females; MEC_Krüger: mean exclusion chance according to Krüger et al.; MEC_Kishida: mean exclusion chance according to Kishida et al, al. MEC_Desmarais: mean exclusion chance in standard trios involving daughters according to Desmarais et al.; MEC_Desmarais_Duo: mean exclusion chance in father/daughter or mother/son duos according to Desmarais et al.

Forensic Parameters	LG1	LG2	LG3	LG4	Combined
haplotypes	79	49	31	6	-
GD	0.9828	0.9634	0.9161	0.6860	-
PIC	0.9782	0.9579	0.9069	0.6270	-
PDM	0.9787	0.9594	0.9123	0.6831	0.99997593
PDF	0.9991	0.9968	0.9870	0.8435	0.999999994
MEC_Kruger	0.9548	0.9171	0.8282	0.4287	0.999632368
MEC_Kishida	0.9759	0.9568	0.9062	0.6269	0.999963592
MEC_Desmarais	0.9782	0.9579	0.9069	0.6270	0.999968196
MEC_Desmarais_duo	0.9580	0.9210	0.8376	0.4823	0.999720678

7.5.6 Concordance study and CE artefact issues

A direct comparison was carried out for the 19 Y-STR loci (DYS576, DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS437, DYS438, DYS439, DYS448, DYS481, DYS533, DYS549, DYS570, DYS635, DYS643, YGATAH4 and DYS385a/b) that are shared between the ForenSeq™ DNA Signature Prep Kit and the Promega PowerPlex® Y23 kit. It must be noted that the DYS385 duplicated Y-STR locus is highly polymorphic, producing two amplicons referred to as DYS385a and DYS385b and identifying two alleles with a single PCR primer pair.

In the ForenSeq™ DNA Signature Prep Kit data, two dropout haplotypes were observed at the DYS385 locus, in which the two samples had the haplotypes (14, 21) and (14,16) using CE-based typing, and (14, 14; with 62 reads) and (14,14; 17 reads) respectively, based on allele length taken from the sequencing data. It is likely that the first dropout was due to preferential amplification, where the smaller allele was more efficiently amplified, and the second dropout due to the low DoC in that sample at the locus. Considering the Y23 kit, six dropout alleles were observed as follows: three at DYS576, two at DYS392 and one at DYS389II. By manually examining the FASTQ files, the three individuals with dropout alleles at DYS576 showed a rare SNP (C>T) in the flanking region at chrY:7185404 (GRCh38.p12) identified as rs757752030 (1000 Genomes frequency 7/1233), which might be the reason behind the artefact. The dropout at DYS389II might also be due to two variants found in the

flanking region: a C>T variant at chrY:12500614 (GRCh38.p12) identified as rs757752030 (accession number: PRJNA586841) and a T>C unreported variant at chrY:12500620 (GRCh38.p12). Further examination of the flanking region for the two dropouts at DYS392 did not reveal any variants that might explain the artefact. Therefore, the dropout might be due to a variant at/near the end of the primer binding region, that is not covered by the FASTQ files, which could disrupt annealing of the primer to its target region.

An observation in one individual was an off-ladder (OL) peak of 361 bp within the size range of the locus DYS439 (**Figure 7.7**). The individual also lacked a peak within the DYS643 locus, which shares the same dye panel with the DYS439. The UAS identified the OL allele as 4, while the allele was flagged as a novel sequence by STRait Razor. Manual examination of the FASTQ file revealed no deletions with respect to the reference genome within the repeated region. However, a gap with an unmapped read was observed (**Figure 7.8**), splitting the read and producing a read which included four pentamer repeats (CTTTT) 5'→3', which was identified by the UAS. A gap within a read (split-read) might suggest a deletion variant that was not aligned with the reference genome. Therefore, the OL allele might be due to the deletion, which produces four CTTT repeats and an allele length of 360 bp, which is outside the DYS643 size range on GeneMapper (368-423 bp covering 6-17 alleles; see **Table 7.1**). A similar artefact has been reported previously for DYS448 [394]. Overall, the dropout and OL alleles were not considered to be a result of discordance between the two kits, therefore it was concluded that there was 100% concordance for the results obtained using CE and MPS for the 19 Y-STRs studied.

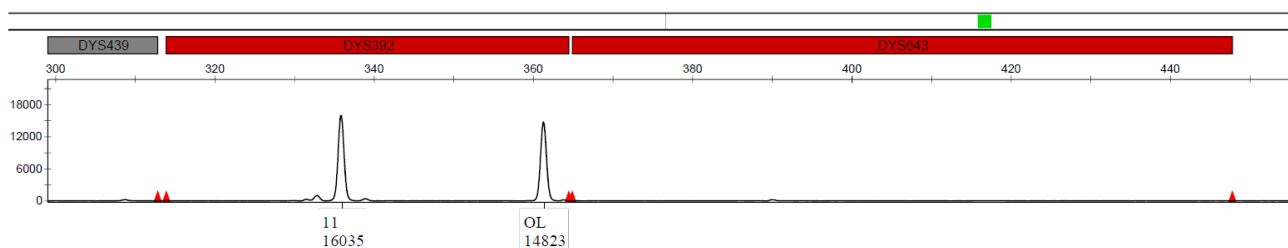


Figure 7.7. Appearance of an allele peak classified as a DYS392 off-ladder allele and a DYS643 null allele.

Off-ladder (OL) peak of 361 bp observed at the DYS392, found to be associated with DYS643 with four pentamer repeats of CTTT.



Figure 7.8. IGV display of DYS643 read alignment of a sample sequenced using the ForenSeq™ DNA Signature Prep Kit.

The grey tracks represent the alignment reads to the GRCh38. The gap observed separated the complete read in a Kuwaiti male, producing a rare haplotype with four CTTT repeats at the DYS643 locus.

7.5.7 Haplotype diversity and haplogroup prediction

NevGen Haplogroup predictor (desktop version) implemented in NevGen Genealogy Tools version 1.1 was used to predict haplogroups based on the 248 full Y-STR profiles (samples with null/OL alleles were included after identification of allele designations by the sequencing data as described above) obtained from the Y23 kit. The major haplogroups present in Arabian Peninsula populations were previously found to be accurately predicted using the NevGen predictor [188]. Of the 248 predictions, 16 were removed from the analysis because they did not match with the software prediction model and scored more than 97.5% of misassignment likelihood along with fitness values of less than < 25, resulting in 232 haplogroup predictions, which were included in the analysis. Haplotypes and predictions are shown in **Appendix 16**.

The largest predicted haplogroup was haplogroup J, comprising 99 (43%) of the total of 232 males studied. Within the J haplogroup, two subclades were suggested by NevGen, including J1 (N=65) and J2 (N=34). All individuals predicted to be in haplogroup J1 were individuals who had a microvariant allele (17.2, 18.2, 19.2, and 20.2) at DYS458, which agrees with previous studies describing the association between microvariant alleles at this locus (0.2) and the haplogroup [188, 394, 395]. Major Haplogroup E was the second most highly predicted,

associated with 47 (20%) Kuwaiti males, of which 15 and 28 individuals were found belonging to the subclades E1b1a and E1b1b, respectively, and four to the haplogroup E2-M75, which is one of the major branches of haplogroup E. Haplogroup R was predicted for 38 (16%) males, associated with the two subclades R1a (N=23) and R1b (N=13), and one major branch R2-M479 (N=2). In addition, other predicted groups were assigned to haplogroups G (subclade G1; N=2, and subclade G2; N=13), T (N=10), L (N=10), Q (N=6), B (N=5), C (N=1) and A1b1b (N=1).

Two additional intermediate alleles, 10.2 and 15.2, were observed at the DYS385 locus belonging to haplogroups B and E1b1b, respectively, and have not been recorded before for these haplogroups in literature. In addition, at DYS385, a tri-allelic pattern was seen in three instances. The duplications in these samples were found to be associated with the R1 haplogroup. Duplication at DY19 produced two individuals with a tri-allelic pattern which was associated with haplogroup G2. The association of a duplication with G2 was confirmed previously by Capelli et al [396]. Duplication was also observed at DYS387S1, where four samples yielded tri allelic patterns associated with haplogroups J2 (N=2), T and R1.

7.6 Conclusion

This chapter evaluates the forensic performance of the seven X-STR contained within the ForenSeq™ DNA Signature Prep Kit, and presents a concordance study of the Y-STR markers collected using different methodologies. X-STR allele variants were categorised based on length alone, based on the sequence of the STR repeat region, and based on the sequence of the STR flanking region. A number of novel alleles were observed in this study that have not been previously published in any population. The increased number of alleles generated by sequencing both the STR repeat region and the flanking region highlights the potential value of this sequencing technology and opens up the possibility for a larger population scale study to characterise these sequence variants and to improve the genotyping approach. In addition, the allele frequencies for the seven X-STRs as well as the frequencies of their four linkage groups were reported for the first time for the Kuwaiti population. All the X-STRs were highly polymorphic and informative indicating their usefulness for forensic applications in the Kuwaiti population. For the 19 Y-STR markers included in both the PCR-CE-based PowerPlex® Y23 kit and the MPS-based ForenSeq™ DNA Signature Prep Kit, a concordance study was carried out and this resulted in 100% concordance between the two methods. several unusual

alleles such as intermediate alleles (0.2) and di/tri allelic patterns were observed. Some of these alleles have been reported in literature with well-established association with their haplogroups. In addition, some artefacts such as null and OL alleles were observed in the profiles generated using CE, and the true underlying alleles were identified using the sequencing data. This also highlights the potential of the MPS technology to overcome some of the limitations of the CE technology. In addition, ten haplogroups were assigned to the male participants in this study, with the haplogroup J being the major haplogroup predicted, which is in concordance with its high frequency in the Arabian Peninsula.

Chapter 8 General discussion

Since the early 1990s, and with the invention of PCR, forensic geneticists have relied almost entirely on short tandem repeat (STR) polymorphisms for the identification of individuals in forensic investigations. Capillary electrophoresis is the primary method that has been used to detect and separate STR fragments. Although single nucleotide polymorphism (SNP) markers can be amplified in small amplicons, and panels of these markers are available for genotyping, they tend not to be practical as they require intensive laboratory work; this is not feasible in forensic laboratories because of the large number of evidentiary samples they analyse. Given that the majority of forensic samples contain low-quality DNA, often in low quantities, the demand for markers that can combine the forensic characteristics of SNPs but use simple STR typing techniques is increasing. Therefore, insertion–deletion (indel) markers have recently been developed to deliver these advantages to forensic analysis. Another milestone is the introduction of massively parallel sequencing (MPS) for forensic use. MPS has the ability to co-amplify hundreds of different types of markers in a single run, signifying a new era of high-resolution sequencing in forensic applications.

Before implementing any typing system in laboratories, population studies must be carried out to produce allele frequencies that enable calculation of statistics to estimate the strength of DNA evidence. Population studies are also necessary for assessing the performance, in that specific population, of each marker contained in a forensic genetic kit to provide guidelines and information for data interpretation. This thesis was undertaken to evaluate the forensic performance of indel markers typed by traditional CE and forensic markers both typed by CE and sequenced by MPS in the Kuwaiti population.

In chapter 3, the Qiagen Investigator® DIPplex kit was used to generate indel profiles of 150 unrelated Kuwaiti individuals, aiming to evaluate the forensic performance of the 30 indel markers contained in the kit and to increase the amount of population genetic data for the Kuwaiti population for which no data are available for this kit.

The results of this study show that all the indel markers are polymorphic for the population, and yielded a combined match probability (CMP) of 2.736×10^{-13} . This means that the likelihood of finding two identical profiles in the Kuwaiti population is 1 in 3.70×10^{12} . This is

an acceptable value given that 15 STRs typed in the Kuwaiti population yielded a CMP of 1.37×10^{-16} [298]. The low CMP obtained for these markers supported by a high combined power of discrimination (CPD) exceeding 99.999% indicates the usefulness of these markers for identification of individuals in the Kuwaiti population. The typical paternity index (TPI) ranged between 0.750 and 1.136 (average = 0.936) for the 30 indels. In contrast, STR markers show a high TPI of 2.475 in the Kuwaiti population [298]. Therefore, indel markers are not informative for paternity testing as a stand-alone system. However, the advantages of their low mutational rate and simple genotyping technique allow them to be useful in difficult paternity testing scenarios, such as when STR markers are inconclusive (for example if mutations exist in families) or in cases where additional markers are needed to provide supporting evidence.

The population differentiation test using allele frequencies of the 30 indel markers from 28 populations (including Kuwaiti) showed satisfactory clusters which indicate that the geographical distance between locations restricts gene flow between populations. Thus, there is more gene flow between geographically close countries, which consequently causes more similar distributions of allele frequencies among them. The allele frequencies of the indel markers in the DIPplex kit can be used to successfully separate geographically distant global populations on the basis of their genetic distance, making them a good choice for determining genetic differentiation between populations. However, the DIPplex kit is not well suited to studying geographically close populations because it was developed for forensic human identification purposes.

Overall, this study shows that the biallelic indel typing system can be used as a stand-alone or as a supplementary marker to STRs in cases in which partial profiles are generated. It can also serve as a potential alternative to SNP typing for identification of individuals in the Kuwaiti population. That said, it has great potential for linking different parts of evidence or for association of body fragments in human identification cases. This report on the allele frequencies of INDELs should serve as a reference database for individual identification in the Kuwaiti population in the future.

In chapter 4, the Promega PowerPlex® Fusion 6C kit was used to generate STR profiles in 400 unrelated Kuwaiti individuals to evaluate the forensic performance of the 23 autosomal STR markers contained in the kit and to generate data comparable to those generated by MPS

technology to examine the concordance of the markers collected using these different methodologies. This chapter also reported four loci (D10S1248, D22S1045, D2S441 and SE33) for the first time in the Kuwaiti population. The statistical analysis showed no significant deviation from the expectations of the Hardy–Weinberg equilibrium, meaning that the PowerPlex® Fusion 6C autosomal STR alleles are independent and can be used to estimate allele frequencies from their genotype frequencies in the Kuwaiti population. In addition, none of the pairs of loci showed significant linkage disequilibrium, indicating that all loci are statistically independent, and that their allele frequencies can be multiplied together to estimate match probabilities in the Kuwaiti population. Therefore, statistical analysis of the forensic parameters show that they all meet the requirements for use in forensic DNA analysis in Kuwait.

The genetic data indicate that these 23 autosomal STRs are highly polymorphic in the Kuwaiti population. The calculated CMP was 7.37×10^{-30} , meaning that the probability of observing two identical profiles for the 23 autosomal loci in the Kuwaiti population was 1 in 1.36×10^{29} . The TPI ranged between 1.439 (TPOX) and 8.333 (SE33), and the combined PE was > 99.9999%. These high values indicate the usefulness of the PowerPlex® Fusion 6C kit for both human identification and paternity testing in the Kuwaiti population. Several off-ladder (OL) artefacts generated by CE were identified using MPS sequencing data, highlighting the importance of MPS technology in addressing some of the limitations of CE. Generally, adoption of the PowerPlex® Fusion 6C kit as a standard analytical kit is highly encouraged for the Kuwaiti Identification DNA Laboratory (KIDL). The allele frequencies of the autosomal STRs provided by this study can be used as a reference database for the Kuwaiti population in the future.

Chapters 5, 6, and 7 discuss the outcome of MPS of 376 Kuwaiti individuals using the Verogen ForenSeq™ DNA Signature Prep Kit, part of the MiSeq FGx® Forensic Genomics system. For this system to be implemented in forensic laboratories, population studies must be carried out to produce allele frequencies that allow the calculation of statistics to estimate the strength of DNA profile evidence, as well as to check for population-specific variants and null alleles.

In chapter 5, the 27 ForenSeq™ DNA Signature Prep Kit autosomal STR markers were studied in terms of their sequence variation and forensic performance in the Kuwaiti population. The

results showed that sequencing data increased the number of length-based alleles by 89%, which highlights the informativeness of the high-resolution technology for forensic use. This study demonstrated that sequence variants detected within both STRs and their flanking regions provide additional allelic diversity compared to CE-based methods. This increased number of genetic variants detected within forensic marker regions using MPS raises the power of this kit as a forensic DNA test. The CMP yielded by the sequenced autosomal loci decreased from $\sim 10^{-34}$ to 10^{-39} when including the sequence variants. MPS technology surpasses conventional CE-based technology by providing considerably greater power to DNA profile evidence, decreasing the match probabilities that can be generated using a single assay.

Concordance studies, in which the same DNA samples are tested with different STR kits, are important to locate potential primer binding site mutations that could lead to allele dropout and discordance of results. In this study, the sequencing results were in 100% concordance with the results obtained using CE for the 23 autosomal loci that are shared between the ForenSeq™ DNA Signature Prep Kit and the Promega PowerPlex® Fusion 6C kit, excluding SE33. The SE33 concordance testing revealed three discordant genotypes which were shown to be due to variants in the flanking region. Although SE33 was included in the primer cocktail of the ForenSeq kit, it has not been reported by the Universal Analysis Software (UAS). It may be difficult for the locus to be included in the sequencing data due to the large size of its repetitive region, restrictions on primer binding sites due to adjacent flank variants, or systematic errors due to the presence of homopolymers in the flanking region.

In chapter 6, the forensic performance of the 94 identity-informative SNP (iiSNP) markers in the ForenSeq™ DNA Signature Prep Kit were evaluated, and genetic variants within the entire sequence of the 94 iiSNP-containing amplicons were identified to assess the level of diversity associated with these variants in the Kuwaiti population. During the evaluation of the 94 iiSNPs, 88 high-performing loci included in this study and six loci were removed due to their lower coverage which is linked to a high risk of allelic dropout. Several factors may affect SNP sequencing coverage, such as PCR amplification efficiency, the accuracy of purification steps, or unsuccessful bead-based library normalisation [258]. The initial quantity of DNA input and the fact that a large number of markers are multiplexed into one sequencing run could also result in depth of coverage (DoC) differences between SNP loci [258]. Of the 88 iiSNP

amplicons analysed, 45 (51%) had variants within the flanking regions. These multi-SNP loci (microhaps) produced a total of 172 distinct haplotypes, ranging from three to seven haplotypes. The inclusion of sequencing data from higher-level sequencing technology increases the informative value of single-base SNPs, especially those associated with variants in their flanking regions, which can provide additional power for the evaluation of evidence.

As SNPs are amplified in smaller amplicons, with no stutter production compared to traditional STR markers, the SNP-microhaplotype system may serve as an efficient tool for discrimination between individuals, in addition to being a useful typing tool for the analysis of degraded or low-quantity DNA samples. The CMP for the 88 iiSNPs was 2.736×10^{-36} , indicating that the likelihood of finding two identical iiSNP profiles was 1 in 4.20×10^{35} . When considering the variants observed in the flanking region, the CMP decreased to 1.87×10^{-39} . The data set including the 88 iiSNPs and the flanking region variants was able to discriminate at a similar order of magnitude than the 28 autosomal ForenSeq™ DNA Signature Prep Kit flank-based (including flanking region variants) STRs (including SE33). In addition, the 45 microhaps alone yielded a CMP of 4.704×10^{-22} , which exceeded the CMP resulting from the typing of 15 STRs (Identifiler kit) in the Kuwaiti population.

The ForenSeq™ DNA Signature Prep Kit contains both STRs and iiSNPs, and combining both sets of markers is likely to substantially increase the overall informativeness of this analysis, decreasing the match probabilities by several orders of magnitude. Prior to calculating these statistics for the combined set of SNP and STR markers using the product rule, the associations between all possible pairwise combinations of alleles were examined across the different loci and showed no LD between them. This indicates that all loci are statistically independent and can, therefore, be used to calculate the genotype frequencies of the Kuwaiti population.

The CMP obtained for the combined 28 autosomal STRs (including SE33) and 88 independent iiSNPs was 3.6675×10^{-70} . Although this represents a significant increase in the discrimination power, the discrimination of good-quality or single-source samples can be performed adequately using PCR-CE typing systems. Therefore, increasing the number of analysed markers is not a useful approach for the analysis of such samples. However, when evaluating degraded samples of low quality or low quantity, STR loci may not be fully typed, resulting in partial/no STR profiles for both the current PCR-CE and the improved MPS systems. Therefore, the inclusion of SNPs and their additional variants can be beneficial for increasing the overall

discrimination power for degraded/low-quality DNA samples. In this study, the CMP was 1.87×10^{-39} for the 88 iiSNPs (43 iiSNPs with 3.975×10^{-18} and 45 microhaps with 4.704×10^{-22}); therefore, even a partial iiSNP profile is likely to be able to generate significant discrimination power.

In chapter 7, the forensic performance of the seven X-STRs in the ForenSeq™ DNA Signature Prep Kit and their sequence variation were evaluated for the first time in the Kuwaiti population. In addition, the Promega PowerPlex® Y23 System was used to generate 248 Y-STR profiles for a concordance study with those generated using sequencing. The MPS results demonstrated a significant increase in allele diversity of 91% in six of the seven X-STR loci. All the X-STRs were highly polymorphic and informative, indicating their usefulness for forensic applications in the Kuwaiti population. The increased number of alleles generated by sequencing both the STR repeat region and the flanking region highlights the potential value of this sequencing technology and opens up the possibility of a larger population-scale study to characterise these sequence variants and to improve the genotyping approach. The concordance study of the Y-STR genotypes generated using the CE and MPS systems showed 100% concordance between the two, indicating that the profiles generated by the two technologies can be compared and used with confidence. The OL and null alleles observed in the CE data were identified using their sequencing data. Ten haplogroups were assigned to the male participants in this study, with haplogroup J being the major haplogroup predicted, which is in accordance with its high frequency in the Arabian Peninsula.

Validation of performance must be carried out before implementation of any DNA method/instrument in forensic laboratories. Validation is a fundamental process that fosters confidence in the results generated. Today, judges and defence lawyers infrequently question the DNA typing science, but they question the process by which the analyst carries out DNA analysis in the laboratory. Therefore, the validation of new techniques and methods must be carefully documented to ensure that the method used accurately reflects the examined samples. For this reason, laboratories must also carefully document the technical procedure for data interpretation and follow it to ensure that each sample is handled and processed properly. Generally, there are two levels of validation: developmental validation and internal validation. Developmental validation is carried out by large laboratories and is supported by

the manufactures of the new DNA typing kits/instruments to verify and establish procedures, whereas internal validation is performed by individual laboratories.

The Verogen ForenSeq™ DNA Signature Prep Kit and the MiSeq FGx® Forensic Genomics system are planned to be implemented in the KIDL soon. As discussed in section 1.11.4, the system was fully validated by Jager et al. who reported developmental validation of the MiSeq FGx®/ForenSeq DNA Signature Prep Kit system based on SWGDAM (www.swgdam.org) guidelines [247]. They validated the kit by performing several experiments, including assessment of species specificity, DNA sensitivity, degraded samples, the impact of PCR inhibitors, DNA mixtures and concordance. However, internal validation will be carried in the KIDL and will follow the recommendations in the SWGDAM guidelines which include establishing analytical and interpretational thresholds, sequencing run evaluation, criteria for locus and allele designation, etc. In addition, population studies must be performed to provide statistical power for the DNA evidence as well as guidelines for the analyst by evaluating the performance of each locus in the kit. This thesis shows that the Verogen ForenSeq™ DNA Signature Prep Kit meets forensic requirements and will be a potential tool to deliver a new level of high power to DNA analysis.

Currently, one of the main challenges facing forensic laboratories wishing to implement NGS is the laborious workflow involved, which is not feasible for many. It is anticipated that this methodology will increasingly become straightforward in the coming years, as the technology is rapidly improving; this will have a positive impact on forensic genomic applications. Moreover, the high cost of sequencing (instrument and kits) hinders its implementation in many forensic laboratories. Again, it is anticipated that the system will become increasingly affordable as the cost of sequencing has dramatically dropped over the last two decades. It should be noted that NGS technologies/platforms are not expected to replace the traditional PCR-CE-based typing system, but will solve its limitations and deliver high-resolution genotyping in instances when standard testing is not successful.

One of the big challenges encountered by the KIDL is identifying the victims of the First Gulf War (1990-1991) during which 602 Kuwaitis went missing and were later confirmed dead. The victims had been killed and buried for approximately 13 years before their skeletal remains were recovered in 2003 from grave sites in Iraq and Kuwait. To date, 273 bodies have been recovered. The process is still ongoing to identify all the skeletal remains brought from Iraq.

Such mass-disaster remains are exposed to harsh environmental conditions, which cause degradation of the DNA molecule. STR profiling has some limitations when dealing with degraded samples with a low quantity of DNA or damaged DNA, which often result in a partial or null STR profile. The use of STRs and SNPs in the ForenSeq™ kit will be beneficial because even a partial iiSNP will generate significant discrimination power, as discussed above.

The current problem associated with using CE-based STR analysis in paternity testing is the potential genotype mismatches caused by single mutations such as indels or SNPs within the STR locus, or due to dropout of alleles which is occasionally seen in paternity cases involving father-mother-child trios or father-child/mother-child duos. In such cases, the STRs could be inconclusive, causing erroneous exclusion of paternity. The current solution for providing more evidence in such cases is to use additional STR markers or Sanger sequencing to identify the mutations which is not feasible for the analyst. The ForenSeq™ DNA Signature Prep kit in this context can be useful to solve these difficult cases by detecting the actual sequence and revealing the existence of mutations.

8.1 Future work

8.1.1 New indel markers

Although this thesis demonstrated the usefulness of the 30 indel markers in the Qiagen Investigator® DIPplex kit, they remain less informative than STRs due to their diallelic state; therefore, more indel markers are needed in order to increase the power of DNA evidence. This may be accomplished using small-length indels (2-5 bp) with a multi-dye system that allows six fluorescent dyes to be used, which is a novel approach to incorporating more indels into one multiplex in a single PCR reaction tube.

An innovative approach to finding new indel markers is to mine the 1000 Genomes Project variants to discover indels that meet the criteria required for forensic markers. To date, no study has examined phase 3 raw data from the 1000 Genomes Project with the sole purpose of using indels as genetic markers for human identification. The Kaviar (Known variants) Project is another repository of variants from 35 projects and contains ~77,800 individuals (13.2K whole genomes and 64.6K exomes), including dbSNP (build 146) and the 1000 Genomes Project (phase 3) [397]. Moreover, the recent build of this project contains more than 50 million small indels from different sources. This large database could be used for new

indels that can be applied for human identification. However, the absence of Middle Eastern populations from such large projects limits our understanding of global genomic diversity and, therefore, has an impact on the choice of globally informative markers for forensic analysis.

Qatar launched the largest national genome project in 2013 with the aim of sequencing the whole Qatari population. This is feasible, considering the small population size (2.7 million) compared to that of many other countries. The MPS data for 1100 Qatari genomes are available in the public domain (1000 whole exomes and 100 whole genomes) [398]. The genetic studies of Qatari and Kuwaiti individuals have revealed three main ancestry groups in both populations descended from the tribes of the Arabian Peninsula that migrated and settled in both countries – Bedouin nomads that settled when they found sites appropriate for their needs and Persians who were coastal neighbours and migrated from Iran to work as merchants and craftsmen [289, 399]. These subpopulations are also present across Arabian Peninsula countries, which means that the results of ancestry studies in each country are representative of the whole region [291, 292, 399, 400]. Therefore, mining genomic data available for these populations could be useful for the development of new markers for forensic human identification, in addition to the latest 1000 Genomes Project release (23 populations). Besides indel markers, mining other types of markers such as SNPs can also be carried out using online genomic databases; however, typing indel markers is more feasible for forensic DNA analysts due to the fact that they employ the simple techniques used in STR typing, as well as sharing the same forensic characteristics of SNPs, i.e. they can be amplified in smaller amplicons with no stutter production.

To date, no published work has considered the application of Ancestry Informative Markers (AIMs) in the Middle East, particularly in the Arabian Peninsula. The importance of the countries in the Arabian Peninsula is due to their position as the largest oil and gas producers, as well as their growing political struggles. Therefore, AIM markers to estimate the proportion of individual ancestry in different populations would have great potential in forensics which can use them to predict the biographical origin of unknown samples to lead investigations. This can be evaluated using the available genomic data. However, more whole-genome sequencing data from Arab countries are still required to better understand the variation between them.

8.1.2 Y-chromosome data

Additional analysis could be carried out for this data set. Therefore, future work includes cataloguing underlying genetic variants in the flanking regions of these STR loci and studying their impact on diversity, as well as investigating whether the surnames of individual donors in the sample, or their social groupings, are linked by common patrilineal descent.

References

1. Gill, P., A.J. Jeffreys, and D.J. Werrett, *Forensic application of DNA /'fingerprints/'*. *Nature*, 1985. **318**(6046): p. 577-579.
2. Morling, N., *Forensic genetics*. *The Lancet*, 2004. **364**: p. 10-11.
3. Butler, J.M., *Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing*. *Journal of Forensic Sciences*, 2006. **51**(2): p. 253-265.
4. Clayton, T.M., J.P. Whitaker, and C.N. Maguire, *Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci*. *Forensic Science International*, 1995. **76**(1): p. 7-15.
5. Underhill, P.A. and T. Kivisild, *Use of y chromosome and mitochondrial DNA population structure in tracing human migrations*. *Annu Rev Genet*, 2007. **41**: p. 539-64.
6. Salzano, F.M., *Reflections of Our Past: How Human History Is Revealed in Our Genes*. *American Journal of Human Genetics*, 2003. **73**(3): p. 706-706.
7. Hammer, M.F., et al., *Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases*. *Forensic Science International*, 2006. **164**(1): p. 45-55.
8. Oki, T., et al., *Development of multiplex assay with 16 SNPs on X chromosome for degraded samples*. *Legal Medicine*, 2012. **14**(1): p. 11-16.
9. Foran, D.R., *Relative degradation of nuclear and mitochondrial DNA: an experimental approach*. *J Forensic Sci*, 2006. **51**(4): p. 766-70.
10. Kayser, M., *Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes*. *Forensic Science International: Genetics*, 2015. **18**: p. 33-48.
11. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome*. *Nature*, 2004. **431**(7011): p. 931-945.
12. *The Genome Reference Consortium*. Available from: <https://www.ncbi.nlm.nih.gov/grc>.
13. Butler, J.M., *Advanced Topics in Forensic DNA Typing: Interpretation*. 2015, San Diego: Academic Press.
14. Gibbs, R.A., et al., *The International HapMap Project*. *Nature*, 2003. **426**(6968): p. 789-796.
15. McVean, G.A., et al., *An integrated map of genetic variation from 1,092 human genomes*. *Nature*, 2012. **491**(7422): p. 56-65.
16. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing*. *Nature*, 2010. **467**(7319): p. 1061-1073.
17. Auton, A., et al., *A global reference for human genetic variation*. *Nature*, 2015. **526**(7571): p. 68-74.
18. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome*. *Nat Rev Genet*, 2006. **7**(2): p. 85-97.
19. Jeffreys, A.J., V. Wilson, and S.L. Thein, *Individual-specific 'fingerprints' of human DNA*. *Nature*, 1985. **316**(6023): p. 76-79.
20. Jeffreys, A.J., J.F. Brookfield, and R. Semeonoff, *Positive identification of an immigration test-case using human DNA fingerprints*. *Nature*, 1985. **317**(6040): p. 818-9.
21. Cobain, I. *Killer breakthrough – the day DNA evidence first nailed a murderer*.
22. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase*. *Science*, 1988. **239**(4839): p. 487-91.
23. Edwards, A., et al., *DNA typing and genetic mapping with trimeric and tetrameric tandem repeats*. *American Journal of Human Genetics*, 1991. **49**(4): p. 746-756.

24. Edwards, A., et al., *Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups*. Genomics, 1992. **12**(2): p. 241-253.
25. Patrinos, G.P., W.J. Ansorge, and P.B. Danielson, *Molecular Diagnostics*. Third ed. 2017: Academic Press.
26. Butler, J.M., *Fundamentals of Forensic DNA Typing*. 2010, San Diego: Academic Press.
27. Wren, J.D., et al., *Repeat polymorphisms within gene regions: phenotypic and evolutionary implications*. American journal of human genetics, 2000. **67**(2): p. 345-356.
28. Willems, T., et al., *The landscape of human STR variation*. Genome research, 2014. **24**(11): p. 1894-1904.
29. Sun, J.X., et al., *A direct characterization of human mutation based on microsatellites*. Nat Genet, 2012. **44**(10): p. 1161-1165.
30. Urquhart, A., et al., *Variation in Short Tandem Repeat sequences —a survey of twelve microsatellite loci for use as forensic identification markers*. International Journal of Legal Medicine, 1994. **107**(1): p. 13-20.
31. Kimpton, C., et al., *Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci*. Int J Legal Med, 1994. **106**(6): p. 302-11.
32. Kimpton, C.P., et al., *Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification*. Electrophoresis, 1996. **17**(8): p. 1283-93.
33. Butler, J.M., *Advanced Topics in Forensic DNA Typing: Methodology*. 2012, San Diego: Academic Press.
34. Cotton, E.A., et al., *Validation of the AMPFISTR® SGM Plus™ system for use in forensic casework*. Forensic Science International, 2000. **112**(2–3): p. 151-161.
35. (CPS), T.C.P.S.; Available from: http://www.cps.gov.uk/legal/d_to_g/dna-17_profiling/.
36. (SPA), S.P.A.; Available from: <http://www.spa.police.uk/news/322981/296781/>.
37. Budowle, B., Moretti, T.R., Niezgoda, S.J. and Brown, B.L. , *Proceedings of the Second European Symposium on Human Identification*. Madison, Wisconsin: Promega, 1998: p. 73–88.
38. Greenspoon, S.A., et al., *Validation and implementation of the PowerPlex 16 BIO System STR multiplex for forensic casework*. J Forensic Sci, 2004. **49**(1): p. 71-80.
39. Wang, D.Y., et al., *Developmental validation of the AmpFISTR(R) Identifiler(R) Plus PCR Amplification Kit: an established multiplex assay with improved performance*. J Forensic Sci, 2012. **57**(2): p. 453-65.
40. Moretti, T.R., et al., *Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States*. Forensic Science International: Genetics, 2016. **25**: p. 175-181.
41. Arnaud, C.H., *Thirty years of DNA forensics: How DNA has revolutionized criminal investigations*, in *Chemical & Engineering News (C&EN)*. 2017, American Chemical Society. p. 16-20.
42. Lin, S.-w., C. Li, and S.C.Y. Ip, *A selection guide for the new generation 6-dye DNA profiling systems*. Forensic Science International: Genetics, 2017. **30**: p. 34-42.
43. Wiegand, P. and M. Kleiber, *Less is more--length reduction of STR amplicons using redesigned primers*. Int J Legal Med, 2001. **114**(4-5): p. 285-7.
44. Tsukada, K., et al., *Multiplex short tandem repeat typing in degraded samples using newly designed primers for the TH01, TPOX, CSF1PO, and vWA loci*. Leg Med (Tokyo), 2002. **4**(4): p. 239-45.
45. Whitaker, J.P., et al., *Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples*. Biotechniques, 1995. **18**(4): p. 670-7.
46. Butler, J.M., Y. Shen, and B.R. McCord, *The development of reduced size STR amplicons as tools for analysis of degraded DNA*. J Forensic Sci, 2003. **48**(5): p. 1054-64.

47. Holland, M.M., et al., *Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the World Trade Center attacks*. Croat Med J, 2003. **44**(3): p. 264-72.
48. Muller, K., et al., *Improved STR typing of telogen hair root and hair shaft DNA*. Electrophoresis, 2007. **28**(16): p. 2835-42.
49. Chung, D.T., et al., *A study on the effects of degradation and template concentration on the amplification efficiency of the STR Miniplex primer sets*. J Forensic Sci, 2004. **49**(4): p. 733-40.
50. Schumm, J.W., R.S. Wingrove, and E.K. Douglas, *Robust STR multiplexes for challenging casework samples*. International Congress Series, 2004. **1261**: p. 547-549.
51. Pizzamiglio, M., et al., *The use of mini-STRs on degraded DNA samples*. International Congress Series, 2006. **1288**: p. 498-500.
52. Martin, P., et al., *Application of mini-STR loci to severely degraded casework samples*. International Congress Series, 2006. **1288**: p. 522-525.
53. Mulero, J.J., et al., *Development and validation of the AmpFISTR MiniFiler PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA*. J Forensic Sci, 2008. **53**(4): p. 838-52.
54. Primorac, D., M.S. Schanfield, and D. Marjanovic, *Basic Genetics and Human Genetic Variation, in Forensic DNA Applications: An Interdisciplinary Perspective*, ed. D. Primorac and M. Shanfield. 2014: CRC Press Taylor & Francis Group.
55. Goodwin, W., A. Linacre, and S. Hadi, *An Introduction to Forensic Genetics*. John Wiley and Sons.
56. *DNA Fragment Analysis by Capillary Electrophoresis*. Available from: <https://www.thermofisher.com/content/dam/LifeTech/global/Forms/PDF/fragment-analysis-chemistry-guide.pdf>.
57. Hauge, X.Y. and M. Litt, *A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR*. Human Molecular Genetics, 1993. **2**(4): p. 411-415.
58. Walsh, P.S., N.J. Fildes, and R. Reynolds, *Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA*. Nucleic Acids Research, 1996. **24**(14): p. 2807-2812.
59. Walsh, P.S., H.A. Erlich, and R. Higuchi, *Preferential PCR amplification of alleles: mechanisms and solutions*. PCR Methods Appl, 1992. **1**(4): p. 241-50.
60. Ogden, S., K. Lamerton, and P. Tatnell, *Evaluation of the AGCU Expressmarker 16 and 22 PCR Amplification Kits Using Biological Samples Applied to FTA Micro Cards in Reduced Volume Direct PCR Amplification Reactions*. Journal of Forensic Science and Medicine, 2015. **1**(1): p. 3-7.
61. Collins, F.S., L.D. Brooks, and A. Chakravarti, *A DNA polymorphism discovery resource for research on human genetic variation*. Genome Res, 1998. **8**(12): p. 1229-31.
62. Brookes, A.J., *The essence of SNPs*. Gene, 1999. **234**(2): p. 177-186.
63. Reich, D.E., et al., *Human genome sequence variation and the influence of gene history, mutation and recombination*. Nat Genet, 2002. **32**(1): p. 135-42.
64. Weber, J.L. and C. Wong, *Mutation of human short tandem repeats*. Hum Mol Genet, 1993. **2**(8): p. 1123-8.
65. Budowle, B. and A. van Daal, *Forensically relevant SNP classes*. Biotechniques, 2008. **44**(5): p. 603-8, 610.
66. Budowle, B., *SNP typing strategies*. Forensic Sci Int, 2004. **146 Suppl**: p. S139-42.
67. Sanchez, J.J., et al., *A multiplex assay with 52 single nucleotide polymorphisms for human identification*. Electrophoresis, 2006. **27**(9): p. 1713-24.
68. Biesecker, L.G., et al., *Epidemiology. DNA identifications after the 9/11 World Trade Center attack*. Science, 2005. **310**(5751): p. 1122-3.
69. Butler, J.M., M.D. Coble, and P.M. Vallone, *STRs vs. SNPs: thoughts on the future of forensic DNA testing*. Forensic Sci Med Pathol, 2007. **3**(3): p. 200-5.

70. Gill, P., *An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes*. Int J Legal Med, 2001. **114**(4-5): p. 204-10.
71. Pakstis, A.J., et al., *SNPs for a universal individual identification panel*. Hum Genet, 2010. **127**(3): p. 315-24.
72. Børsting, C., M. Mikkelsen, and N. Morling, *Kinship Analysis with Diallelic SNPs - Experiences with the SNPforID Multiplex in an ISO17025 Accredited Laboratory*. Transfusion Medicine and Hemotherapy, 2012. **39**(3): p. 195-201.
73. Børsting, C., E. Rockenbauer, and N. Morling, *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, 2009. **4**(1): p. 34-42.
74. Butler, J.M., et al., *Report on ISFG SNP Panel Discussion*. Forensic Science International: Genetics Supplement Series, 2008. **1**(1): p. 471-472.
75. Kidd, K.K., et al., *Developing a SNP panel for forensic identification of individuals*. Forensic Science International, 2006. **164**(1): p. 20-32.
76. Phillips, C., et al., *Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel*. Forensic Science International: Genetics, 2007. **1**(2): p. 180-185.
77. Butler, J.M., Hill, C.R. and Coble, M.D. , *Variability of New STR Loci and Kits in US Population Groups*. Available from: <http://www.promega.co.uk/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/> 2012.
78. Phillips, C., et al., *Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs*. Forensic Science International: Genetics, 2007. **1**(3-4): p. 273-280.
79. Kidd, K.K., et al., *Progress toward an efficient panel of SNPs for ancestry inference*. Forensic Science International: Genetics, 2014. **10**: p. 23-32.
80. Phillips, C., et al., *Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation*. PLoS ONE, 2009. **4**(8): p. e6583.
81. Phillips, C., *Forensic genetic analysis of bio-geographical ancestry*. Forensic Science International: Genetics, 2015. **18**: p. 49-65.
82. Prestes, P.R., et al., *The SNPforID 34-plex—Its ability to infer level of admixture in individuals*. Forensic Science International: Genetics Supplement Series, 2013. **4**(1): p. e13-e14.
83. Valentine, T. and R.J. Fitzgerald, *Identifying the Culprit: An International Perspective on the National Academy of Sciences Report on Eyewitness Identification Evidence*. Applied Cognitive Psychology, 2016. **30**(1): p. 135-138.
84. Frudakis, T.N., *Chapter 9 - Direct Method of Phenotype Inference*, in *Molecular Photofitting*. 2008, Academic Press: Burlington. p. 497-598.
85. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height*. Nature, 2010. **467**(7317): p. 832-8.
86. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height*. Nat Genet, 2014. **46**(11): p. 1173-86.
87. Aulchenko, Y.S., et al., *Predicting human height by Victorian and genomic methods*. Eur J Hum Genet, 2009. **17**(8): p. 1070-1075.
88. Liu, F., et al., *Common DNA variants predict tall stature in Europeans*. Hum Genet, 2014. **133**(5): p. 587-97.
89. Richards, J.B., et al., *Male-pattern baldness susceptibility locus at 20p11*. Nat Genet, 2008. **40**(11): p. 1282-4.
90. Redler, S., et al., *Investigation of the male pattern baldness major genetic susceptibility loci AR/EDA2R and 20p11 in female pattern hair loss*. Br J Dermatol, 2012. **166**(6): p. 1314-8.
91. Fujimoto, A., et al., *A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness*. Hum Mol Genet, 2008. **17**(6): p. 835-43.
92. Fujimoto, A., et al., *FGFR2 is associated with hair thickness in Asian populations*. J Hum Genet, 2009. **54**(8): p. 461-5.

93. Medland, S.E., et al., *Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans*. American Journal of Human Genetics, 2009. **85**(5): p. 750-755.
94. Medland, S.E., G. Zhu, and N.G. Martin, *Estimating the heritability of hair curliness in twins of European ancestry*. Twin Res Hum Genet, 2009. **12**(5): p. 514-8.
95. Paternoster, L., et al., *Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position*. Am J Hum Genet, 2012. **90**(3): p. 478-85.
96. Liu, F., et al., *A genome-wide association study identifies five loci influencing facial morphology in Europeans*. PLoS Genet, 2012. **8**(9): p. e1002932.
97. Boehringer, S., et al., *Genetic determination of human facial morphology: links between cleft-lips and normal variation*. Eur J Hum Genet, 2011. **19**(11): p. 1192-7.
98. Sheehan, M.J. and M.W. Nachman, *Morphological and population genomic evidence that human faces have evolved to signal individual identity*. Nat Commun, 2014. **5**: p. 4800.
99. Claes, P., et al., *Modeling 3D Facial Shape from DNA*. PLoS Genetics, 2014. **10**(3): p. e1004224.
100. Kayser, M. and P. de Knijff, *Improving human forensics through advances in genetics, genomics and molecular biology*. Nat Rev Genet, 2011. **12**(3): p. 179-92.
101. Walsh, S., et al., *IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information*. Forensic Science International: Genetics, 2011. **5**(3): p. 170-180.
102. Walsh, S., et al., *DNA-based eye colour prediction across Europe with the IrisPlex system*. Forensic Sci Int Genet, 2012. **6**(3): p. 330-40.
103. Walsh, S., et al., *Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence*. Forensic Sci Int Genet, 2011. **5**(5): p. 464-71.
104. Dembinski, G.M. and C.J. Picard, *Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population*. Forensic Science International: Genetics, 2014. **9**: p. 111-117.
105. Walsh, S., et al., *The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA*. Forensic Science International: Genetics, 2013. **7**(1): p. 98-115.
106. Draus-Barini, J., et al., *Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains*. Investigative Genetics, 2013. **4**: p. 3-3.
107. Haeusler, M., et al., *Multidisciplinary Identification of the Controversial Freedom Fighter Jörg Jenatsch, Assassinated 1639 in Chur, Switzerland*. PLoS ONE, 2016. **11**(12): p. e0168014.
108. King, T.E., et al., *Identification of the remains of King Richard III*. Nat Commun, 2014. **5**: p. 5631.
109. Chaitanya, L., et al., *Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system*. Forensic Sci Int Genet, 2017. **26**: p. 48-57.
110. Sobrino, B., M. Brión, and A. Carracedo, *SNPs in forensic genetics: a review on SNP typing methodologies*. Forensic Science International, 2005. **154**(2-3): p. 181-194.
111. Daniel, R., et al., *A SNaPshot of next generation sequencing for forensic SNP analysis*. Forensic Science International: Genetics, 2015. **14**: p. 50-60.
112. Mills, R.E., et al., *An initial map of insertion and deletion (INDEL) variation in the human genome*. Genome Research, 2006. **16**(9): p. 1182-1190.
113. Nachman, M.W. and S.L. Crowell, *Estimate of the mutation rate per nucleotide in humans*. Genetics, 2000. **156**(1): p. 297-304.
114. Weber, J.L., et al., *Human Diallelic Insertion/Deletion Polymorphisms*. The American Journal of Human Genetics, 2002. **71**(4): p. 854-862.
115. Pereira, R., et al., *A new multiplex for human identification using insertion/deletion polymorphisms*. ELECTROPHORESIS, 2009. **30**(21): p. 3682-3690.
116. Mullikin, J.C., et al., *An SNP map of human chromosome 22*. Nature, 2000. **407**(6803): p. 516-520.

117. Dawson, E., et al., *A SNP Resource for Human Chromosome 22: Extracting Dense Clusters of SNPs From the Genomic Sequence*. Genome Research, 2001. **11**(1): p. 170-178.
118. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes*. Nature, 2008. **453**(7191): p. 56-64.
119. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes*. Human Molecular Genetics, 2010. **19**(R2): p. R131-R136.
120. The Genomes Project, C., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
121. Braganholi, D.F., et al., *Insertion of amelogenin marker in the 38 Hid-Indels multiplex reaction*. Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e165-e166.
122. Manta, F., et al., *Indel markers: Genetic diversity of 38 polymorphisms in Brazilian populations and application in a paternity investigation with post mortem material*. Forensic Science International: Genetics, 2012. **6**(5): p. 658-661.
123. Ferragut, J.F., et al., *Genetic diversity of 38 insertion–deletion polymorphisms in Jewish populations*. Forensic Science International: Genetics, 2016. **21**: p. 1-4.
124. Friis, S.L., et al., *Typing of 30 insertion/deletions in Danes using the first commercial indel kit—Mentype® DIPplex*. Forensic Science International: Genetics, 2012. **6**(2): p. e72-e74.
125. Turrina, S., G. Filippini, and D. De Leo, *Forensic evaluation of the Investigator DIPplex typing system*. Forensic Science International: Genetics Supplement Series, 2011. **3**(1): p. e331-e332.
126. Kis, Z., et al., *Genome deletion and insertion polymorphisms (DIPs) in the Hungarian population*. Forensic Science International: Genetics, 2012. **6**(5): p. e125-e126.
127. da Silva, C.V., et al., *Genetic portrait of south Portugal population with InDel markers*. Forensic Science International: Genetics, 2013. **7**(4): p. e101-e103.
128. Neuvonen, A.M., et al., *Discrimination power of Investigator DIPplex loci in Finnish and Somali populations*. Forensic Sci Int Genet, 2012. **6**(4): p. e99-102.
129. Shi, M., et al., *Population data of 30 insertion-deletion markers in four Chinese populations*. Int J Legal Med, 2015. **129**(1): p. 53-6.
130. Ferreira Palha, T.d.J.B., et al., *Population genetic analysis of insertion–deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit*. Forensic Science International: Genetics, 2015. **19**: p. 10-14.
131. Reis, F., et al., *Genetic characterization of the Brazilian immigrant population in Lisboa with InDel genetic markers (P)*. Forensic Science International: Genetics Supplement Series, 2017.
132. Hefke, G., S. Davison, and M.E. D'Amato, *Forensic performance of Investigator DIPplex indels genotyping kit in native, immigrant, and admixed populations in South Africa*. ELECTROPHORESIS, 2015. **36**(24): p. 3018-3025.
133. Pepinski, W., et al., *Population genetics of 30 INDELs in populations of Poland and Taiwan*. Molecular Biology Reports, 2013. **40**(7): p. 4333-4338.
134. Kim, E.H., et al., *Population data for 30 insertion–deletion markers in a Korean population*. International Journal of Legal Medicine, 2014. **128**(1): p. 51-52.
135. Torres, S.R.R., et al., *Population genetic data and forensic parameters of 30 autosomal InDel markers in Santa Catarina State population, Southern Brazil*. Molecular Biology Reports, 2014. **41**(8): p. 5429-5433.
136. D' Amato, M.E., G. Hefke, and S. Davison, *Evaluation of DIPplex investigator kit in European, Asian and African populations*. Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e470-e471.
137. Tomas, C., et al., *Thirty autosomal insertion-deletion polymorphisms analyzed using the Investigator® DIPplex Kit in populations from Iraq, Lithuania, Slovenia, and Turkey*. Forensic Science International: Genetics, 2016. **25**(Supplement C): p. 142-144.
138. Martinez-Cortes, G., et al., *Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations*. Int J Legal Med, 2016. **130**(3): p. 683-5.

139. Wei, Y.-L., et al., *A validation study of a multiplex INDEL assay for forensic use in four Chinese populations*. Forensic Science International: Genetics, 2014. **9**: p. e22-e25.
140. Nunotani, M., et al., *Analysis of 30 insertion-deletion polymorphisms in the Japanese population using the Investigator DIPplex(R) kit*. Leg Med (Tokyo), 2015. **17**(6): p. 467-70.
141. Seong, K.M., et al., *Population genetics of insertion–deletion polymorphisms in South Koreans using Investigator DIPplex kit*. Forensic Science International: Genetics, 2014. **8**(1): p. 80-83.
142. Martín, P., et al., *Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations*. Forensic Science International: Genetics, 2013. **7**(2): p. e27-e30.
143. Fondevila, M., et al., *Forensic performance of two insertion-deletion marker assays*. Int J Legal Med, 2012. **126**(5): p. 725-37.
144. Fondevila, M., et al., *Forensic performance of insertion–deletion marker systems*. Forensic Science International: Genetics Supplement Series, 2011. **3**(1): p. e443-e444.
145. LaRue, B.L., et al., *A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification*. International Journal of Legal Medicine, 2012. **126**(4): p. 533-540.
146. Zidkova, A., et al., *Application of the new insertion–deletion polymorphism kit for forensic identification and parentage testing on the Czech population*. International Journal of Legal Medicine, 2013. **127**(1): p. 7-10.
147. Prokop, J.W. and C.F. Deschepper, *Chromosome Y genetic variants: impact in animal models and on human disease*. Physiological genomics, 2015. **47**(11): p. 525-537.
148. Colaco, S. and D. Modi, *Genetics of the human Y chromosome and its association with male infertility*. Reprod Biol Endocrinol, 2018. **16**(1): p. 14.
149. Veerappa, A.M., P. Padakannaya, and N.B. Ramachandra, *Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome*. Functional & Integrative Genomics, 2013. **13**(3): p. 285-293.
150. Petr, M., et al., *The evolutionary history of Neandertal and Denisovan Y chromosomes*. bioRxiv, 2020: p. 2020.03.09.983445.
151. Bachrog, D., *Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration*. Nature Reviews Genetics, 2013. **14**(2): p. 113-124.
152. Jobling, M.A., *In the name of the father: surnames and genetics*. Trends Genet, 2001. **17**(6): p. 353-7.
153. Zeng, Z., et al., *Inferring the history of surname Ye based on Y chromosome high-resolution genotyping and sequencing data*. Journal of Human Genetics, 2019. **64**(8): p. 703-709.
154. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview*. Hum Genet, 2017. **136**(5): p. 621-635.
155. Roewer, L., et al., *Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts*. Human Genetics, 1992. **89**(4): p. 389-394.
156. Roewer, L. and J.T. Epplen, *Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work*. Forensic Science International, 1992. **53**(2): p. 163-171.
157. Kayser, M., *Forensic use of Y-chromosome DNA: a general overview*. Human Genetics, 2017. **136**(5): p. 621-635.
158. Sarno, S., et al., *An Ancient Mediterranean Melting Pot: Investigating the Uniparental Genetic Structure and Population History of Sicily and Southern Italy*. PLOS ONE, 2014. **9**(4): p. e96074.
159. Jobling, M.A. and C. Tyler-Smith, *Human Y-chromosome variation in the genome-sequencing era*. Nat Rev Genet, 2017. **18**(8): p. 485-497.
160. Balanovsky, O., *Toward a consensus on SNP and STR mutation rates on the human Y-chromosome*. Human Genetics, 2017. **136**(5): p. 575-590.

161. Ballantyne, K.N., et al., *Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications*. The American Journal of Human Genetics, 2010. **87**(3): p. 341-353.
162. Ballantyne, K.N., et al., *A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages*. Forensic Science International: Genetics, 2012. **6**(2): p. 208-218.
163. Kayser, M., et al., *Evaluation of Y-chromosomal STRs: a multicenter study*. International Journal of Legal Medicine, 1997. **110**(3): p. 125-133.
164. SWGDAM, *Report on the Current Activities of the Scientific Working Group on DNA Analysis Methods Y-STR Subcommittee*. Forensic Science Communications. **6**(3): p. Available at <https://www.questia.com/library/journal/1G1-137863478/report-on-the-current-activities-of-the-scientific>.
165. Gusmão, L. and A. Carracedo, *Y chromosome-specific STRs*. Profiles DNA, 2003. **6**(1): p. 3-6.
166. Willems, T., et al., *Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates*. American journal of human genetics, 2016. **98**(5): p. 919-933.
167. Hammer, M.F. and S.L. Zegura, *The Human Y Chromosome Haplogroup Tree: Nomenclature and Phylogeography of Its Major Divisions*. Annual Review of Anthropology, 2002. **31**(1): p. 303-321.
168. de Knijff, P., *Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome*. American journal of human genetics, 2000. **67**(5): p. 1055-1061.
169. Jobling, M.A. and C. Tyler-Smith, *The human Y chromosome: an evolutionary marker comes of age*. Nat Rev Genet, 2003. **4**(8): p. 598-612.
170. Rosser, Z.H., et al., *Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language*. The American Journal of Human Genetics, 2000. **67**(6): p. 1526-1543.
171. van Oven, M., et al., *Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome*. Human Mutation, 2014. **35**(2): p. 187-191.
172. Geppert, M. and L. Roewer, *SNaPshot® Minisequencing Analysis of Multiple Ancestry-Informative Y-SNPs Using Capillary Electrophoresis*, in *DNA Electrophoresis Protocols for Forensic Genetics*, A. Alonso, Editor. 2012, Humana Press: Totowa, NJ. p. 127-140.
173. van Oven, M., et al., *Multiplex genotyping assays for fine-resolution subtyping of the major human Y-chromosome haplogroups E, G, I, J, and R in anthropological, genealogical, and forensic investigations*. ELECTROPHORESIS, 2013. **34**(20-21): p. 3029-3038.
174. van Oven, M., N. van den Tempel, and M. Kayser, *A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia*. Journal of Human Genetics, 2012. **57**(1): p. 65-69.
175. van Oven, M., A. Ralf, and M. Kayser, *An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups*. International Journal of Legal Medicine, 2011. **125**(6): p. 879.
176. Ralf, A., et al., *Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing*. Forensic Science International: Genetics, 2019. **41**: p. 93-106.
177. Wang, M., et al., *Developmental validation of a custom panel including 165 Y-SNPs for Chinese Y-chromosomal haplogroups dissection using the ion S5 XL system*. Forensic Science International: Genetics, 2019. **38**: p. 70-76.
178. Roewer, L., *Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA*. WIREs Forensic Science, 2019. **1**(4): p. e1336.
179. Kayser, M. and K.N. Ballantyne, *Forensic DNA Applications: An Interdisciplinary Perspective*, ed. D. Primorac and M. Shanfield. 2014: CRC Press Taylor & Francis Group.

180. kayser, M. and K.N. Ballantyne, *Y Chromosome in Forensic Science*, in *Forensic DNA Applications: An Interdisciplinary Perspective*, ed. D. Primorac and M. Shanfield. 2014: CRC Press Taylor & Francis Group.
181. Karafet, T.M., et al., *New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree*. *Genome research*, 2008. **18**(5): p. 830-838.
182. Consortium, Y.C., *A nomenclature system for the tree of human Y-chromosomal binary haplogroups*. *Genome research*, 2002. **12**(2): p. 339-348.
183. Hallast, P., et al., *The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades*. *Molecular biology and evolution*, 2015. **32**(3): p. 661-673.
184. Larmuseau, M.H.D., et al., *Towards a consensus Y-chromosomal phylogeny and Y-SNP set in forensics in the next-generation sequencing era*. *Forensic Science International: Genetics*, 2015. **15**: p. 39-42.
185. Willuweit, S. and L. Roewer, *The new Y Chromosome Haplotype Reference Database*. *Forensic Science International: Genetics*, 2015. **15**: p. 43-48.
186. Wang, C.-C., et al., *Convergence of Y Chromosome STR Haplotypes from Different SNP Haplogroups Compromises Accuracy of Haplogroup Prediction*. *Journal of Genetics and Genomics*, 2015. **42**(7): p. 403-407.
187. Emmerova, B., et al., *Comparison of Y-chromosomal haplogroup predictors*. *Forensic Science International: Genetics Supplement Series*, 2017. **6**: p. e145-e147.
188. Khubrani, Y.M., J.H. Wetton, and M.A. Jobling, *Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs*. *Forensic Science International: Genetics*, 2018. **33**: p. 98-105.
189. Jannuzzi, J., et al., *Male lineages in Brazilian populations and performance of haplogroup prediction tools*. *Forensic Science International: Genetics*, 2020. **44**: p. 102163.
190. Athey, T.W., *Haplogroup prediction from Y-STR values using a bayesian-allele- frequency approach*. *Journal of Genetic Genealogy*, 2006. **2**: p. 34-39.
191. Athey, T.W., *Haplogroup prediction from Y-STR values using a bayesian-allele- frequency approach*. *Journal of Genetic Genealogy*, 2005. **1**: p. 1-7.
192. Muzzio, M., et al., *Software for Y-haplogroup predictions: a word of caution*. *International Journal of Legal Medicine*, 2011. **125**(1): p. 143-147.
193. Athey, W., *Comments on the article, "Software for Y Haplogroup Predictions, a Word of Caution"*. *International Journal of Legal Medicine*, 2011. **125**(6): p. 901-903.
194. Schaffner, S.F., *The X chromosome in population genetics*. *Nature Reviews Genetics*, 2004. **5**(1): p. 43-51.
195. Khitrinskaya, I.Y., V.N. Khar'kov, and V.A. Stepanov, *Genetic diversity of the chromosome X in aboriginal Siberian populations: The structure of linkage disequilibrium and haplotype phylogeography of the ZFX locus*. *Molecular Biology*, 2010. **44**(5): p. 709-719.
196. Dobyns, W.B., et al., *Inheritance of most X-linked traits is not dominant or recessive, just X-linked*. *American Journal of Medical Genetics Part A*, 2004. **129A**(2): p. 136-143.
197. Krawczak, M., *Kinship testing with X-chromosomal markers: Mathematical and statistical issues*. *Forensic Science International: Genetics*, 2007. **1**(2): p. 111-114.
198. Szibor, R., *X-chromosomal markers: Past, present and future*. *Forensic Science International: Genetics*, 2007. **1**(2): p. 93-99.
199. Diegoli, T.M., *Forensic Application of X Chromosome STRs*, in *Forensic DNA Applications: An Interdisciplinary Perspective*, ed. D. Primorac and M. Shanfield. 2014: CRC Press Taylor & Francis Group.
200. Schmidtke, J., et al., *Prenatal exclusion without involving the putative fathers of an incestuous father-daughter parenthood*. *Prenatal Diagnosis*, 2004. **24**(8): p. 662-664.
201. Lancia, M., et al., *Using X-chromosomal markers in rape investigation*. *Forensic Science International: Genetics Supplement Series*, 2011. **3**(1): p. e55-e56.

202. Silva, D.A., et al., *Paternity testing involving human remains identification and putative half sister: Usefulness of an X-hexaplex STR markers*. Forensic Science International: Genetics Supplement Series, 2009. **2**(1): p. 230-231.
203. Li, L., et al., *Analysis of 14 highly informative SNP markers on X chromosome by TaqMan® SNP genotyping assay*. Forensic Science International: Genetics, 2010. **4**(5): p. e145-e148.
204. Tomàs, C., et al., *Utility of X-chromosome SNPs in relationship testing*. Forensic Science International: Genetics Supplement Series, 2008. **1**(1): p. 528-530.
205. Stepanov, V.A., et al., *Panel of X-linked single-nucleotide polymorphic markers for DNA identification (XSNPid) based on multiplex genotyping by multilocus PCR and MALDI-TOF mass spectrometry*. Molecular Biology, 2016. **50**(3): p. 387-397.
206. Szibor, R., et al., *Use of X-linked markers for forensic purposes*. International Journal of Legal Medicine, 2003. **117**(2): p. 67-74.
207. Nothnagel, M., et al., *Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome*. Forensic Science International: Genetics, 2012. **6**(6): p. 778-784.
208. Anderson, S., et al., *Sequence and organization of the human mitochondrial genome*. Nature, 1981. **290**(5806): p. 457-465.
209. Andrews, R.M., et al., *Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA*. Nature Genetics, 1999. **23**(2): p. 147-147.
210. Amorim, A., T. Fernandes, and N. Taveira, *Mitochondrial DNA in human identification: a review*. PeerJ, 2019. **7**: p. e7314-e7314.
211. Greenberg, B.D., J.E. Newbold, and A. Sugino, *Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA*. Gene, 1983. **21**(1): p. 33-49.
212. Yakes, F.M. and B. Van Houten, *Mitochondrial DNA damage is more extensive and persists longer than nuclear DNA damage in human cells following oxidative stress*. Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(2): p. 514-519.
213. Holland, M.M. and G. LAUC, *Forensic Aspects of mtDNA Analysis, in Forensic DNA Applications: An Interdisciplinary Perspective*, ed. D. Primorac and M. Shanfield. 2014: CRC Press Taylor & Francis Group. 85-104.
214. Wei, W., et al., *Mitochondrial DNA point mutations and relative copy number in 1363 disease and control human brains*. Acta neuropathologica communications, 2017. **5**(1): p. 13-13.
215. Lutz, S., et al., *Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HVI and HVII regions?* Forensic Science International, 2000. **113**(1): p. 97-101.
216. Satoh, M. and T. Kuroiwa, *Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell*. Experimental Cell Research, 1991. **196**(1): p. 137-140.
217. Elson, J.L., et al., *Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age*. American journal of human genetics, 2001. **68**(3): p. 802-806.
218. Loreille, O.M., et al., *High efficiency DNA extraction from bone by total demineralization*. Forensic Science International: Genetics, 2007. **1**(2): p. 191-195.
219. Melton, T., et al., *Forensic Mitochondrial DNA Analysis of 691 Casework Hairs*. Journal of Forensic Sciences, 2005. **50**(1): p. JFS2004230-8.
220. Case, J.T. and D.C. Wallace, *Maternal inheritance of mitochondrial DNA polymorphisms in cultured human fibroblasts*. Somatic Cell Genet, 1981. **7**(1): p. 103-8.
221. Giles, R.E., et al., *Maternal inheritance of human mitochondrial DNA*. Proceedings of the National Academy of Sciences of the United States of America, 1980. **77**(11): p. 6715-6719.
222. Hutchison, C.A., et al., *Maternal inheritance of mammalian mitochondrial DNA*. Nature, 1974. **251**(5475): p. 536-538.
223. Chen, X., et al., *Rearranged mitochondrial genomes are present in human oocytes*. American journal of human genetics, 1995. **57**(2): p. 239-247.

224. Wai, T., et al., *The role of mitochondrial DNA copy number in mammalian fertility*. Biology of reproduction, 2010. **83**(1): p. 52-62.
225. King, T.E., et al., *Identification of the remains of King Richard III*. Nature communications, 2014. **5**: p. 5631-5631.
226. Ballard, D., *Analysis of Mitochondrial Control Region Using Sanger Sequencing*, in *Forensic DNA Typing Protocols*, W. Goodwin, Editor. 2016, Springer New York: New York, NY. p. 143-155.
227. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(12): p. 5463-5467.
228. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA*. Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(2): p. 560-564.
229. Smith, L.M., et al., *The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis*. Nucleic acids research, 1985. **13**(7): p. 2399-2412.
230. Ansorge, W., et al., *A non-radioactive automated method for DNA sequence determination*. Journal of Biochemical and Biophysical Methods, 1986. **13**(6): p. 315-323.
231. Smith, L.M., et al., *Fluorescence detection in automated DNA sequence analysis*. Nature, 1986. **321**(6071): p. 674-679.
232. Hood, L.E., M.W. Hunkapiller, and L.M. Smith, *Automated DNA sequencing and analysis of the human genome*. Genomics, 1987. **1**(3): p. 201-212.
233. Shendure, J., et al., *DNA sequencing at 40: past, present and future*. Nature, 2017. **550**(7676): p. 345-353.
234. Gužvić, M., *History of DNA sequencing*. Journal of Medical Biochemistry, 2013. **32**(4): p. 301-312.
235. Fullwood, M.J., et al., *Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses*. Genome Res, 2009. **19**(4): p. 521-32.
236. Metzker, M.L., *Sequencing technologies - the next generation*. Nat Rev Genet, 2010. **11**(1): p. 31-46.
237. Krier, J.B., S.S. Kalia, and R.C. Green, *Genomic sequencing in clinical practice: applications, challenges, and opportunities*. Dialogues in clinical neuroscience, 2016. **18**(3): p. 299-312.
238. McCarthy, J.J., H.L. McLeod, and G.S. Ginsburg, *Genomic medicine: a decade of successes, challenges, and opportunities*. Sci Transl Med, 2013. **5**(189): p. 189sr4.
239. Marciniak, S. and G.H. Perry, *Harnessing ancient genomes to study the history of human adaptation*. Nat Rev Genet, 2017. **18**(11): p. 659-674.
240. Poinar, H.N., et al., *Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA*. Science, 2006. **311**(5759): p. 392-4.
241. Motro, Y. and J. Moran-Gilad, *Next-generation sequencing applications in clinical bacteriology*. Biomol Detect Quantif, 2017. **14**: p. 1-6.
242. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes*. Nature, 2012. **491**(7422): p. 56-65.
243. Weber-Lehmann, J., et al., *Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing*. Forensic Sci Int Genet, 2014. **9**: p. 42-6.
244. Palencia-Madrid, L. and M.M. de Pancorbo, *Next generation sequencing as a suitable analysis for authentication of degraded and ancient DNA*. Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e338-e340.
245. Lyons, E.A., et al., *A high-throughput Sanger strategy for human mitochondrial genome sequencing*. BMC genomics, 2013. **14**: p. 881-881.
246. Zhang, J., et al., *The impact of next-generation sequencing on genomics*. Journal of genetics and genomics = Yi chuan xue bao, 2011. **38**(3): p. 95-109.

247. Jager, A.C., et al., *Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories*. Forensic Sci Int Genet, 2017. **28**: p. 52-70.
248. Montano, E.A., et al., *Optimization of the Promega PowerSeq Auto/Y system for efficient integration within a forensic DNA laboratory*. Forensic Sci Int Genet, 2018. **32**: p. 26-32.
249. FTA™ Mini cards. Available from: <https://www.gelifesciences.com/en/us/shop/forensics/cards-kits-and-collectors/non-indicating-fta-mini-cards-two-sample-areas-per-card-p-00637>.
250. ForenSeq DNA Signature Prep Kit.
251. Iozzi, S., et al., *Forensic genetics in NGS era: New frontiers for massively parallel typing*. Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e418-e419.
252. Almalki, N., et al., *Systematic assessment of the performance of illumina's MiSeq FGx forensic genomics system*. Electrophoresis, 2017. **38**(6): p. 846-854.
253. Just, R.S., et al., *Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens*. Forensic Science International: Genetics, 2017. **28**: p. 1-9.
254. Silvia, A.L., N. Shugarts, and J. Smith, *A preliminary assessment of the ForenSeq FGx System: next generation sequencing of an STR and SNP multiplex*. Int J Legal Med, 2017. **131**(1): p. 73-86.
255. Xavier, C. and W. Parson, *Evaluation of the Illumina ForenSeq DNA Signature Prep Kit - MPS forensic application for the MiSeq FGx benchtop sequencer*. Forensic Sci Int Genet, 2017. **28**: p. 188-194.
256. Kocher, S., et al., *Inter-laboratory validation study of the ForenSeq DNA Signature Prep Kit*. Forensic Sci Int Genet, 2018. **36**: p. 77-85.
257. Moreno, L.I., M.B. Galusha, and R. Just, *A closer look at Verogen's Forenseq DNA Signature Prep kit autosomal and Y-STR data for streamlined analysis of routine reference samples*. Electrophoresis, 2018. **39**(21): p. 2685-2693.
258. Churchill, J.D., et al., *Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling*. Forensic Science International: Genetics, 2016. **20**: p. 20-29.
259. Caratti, S., et al., *MiSeq FGx sequencing system: A new platform for forensic genetics*. Forensic Science International: Genetics Supplement Series, 2015. **5**: p. e98-e100.
260. Calafell, F., et al., *An assessment of a massively parallel sequencing approach for the identification of individuals from mass graves of the Spanish Civil War (1936-1939)*. Electrophoresis, 2016. **37**(21): p. 2841-2847.
261. Zhang, Q., et al., *Evaluation of the performance of Illumina's ForenSeq system on serially degraded samples*. Electrophoresis, 2018. **39**(21): p. 2674-2684.
262. Kostrzewa, G., M. Konarzewska, and W. Pepiński, *Application of massively parallel sequencing (MPS) in paternity testing – case report*. Archiwum Medycyny Sądowej i Kryminologii/Archives of Forensic Medicine and Criminology, 2017. **67**(1): p. 61-67.
263. Ma, Y., et al., *Next generation sequencing: Improved resolution for paternal/maternal duos analysis*. Forensic Sci Int Genet, 2016. **24**: p. 83-85.
264. Ramani, A., et al., *Ancestry prediction in Singapore population samples using the Illumina ForenSeq kit*. Forensic Sci Int Genet, 2017. **31**: p. 171-179.
265. King, J.L., et al., *Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit*. Forensic Sci Int Genet, 2018. **36**: p. 60-76.
266. Parson, W., et al., *Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements*. Forensic Sci Int Genet, 2016. **22**: p. 54-63.
267. Novroski, N.M.M., et al., *Characterization of genetic sequence variation of 58 STR loci in four major population groups*. Forensic Sci Int Genet, 2016. **25**: p. 214-226.

268. Wendt, F.R., et al., *Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system*. Forensic Sci Int Genet, 2016. **24**: p. 18-23.
269. Churchill, J.D., et al., *Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System*. Forensic Science International: Genetics, 2017. **30**: p. 81-92.
270. Devesse, L., et al., *Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups*. Forensic Science International: Genetics, 2018. **34**: p. 57-61.
271. Kim, S.Y., et al., *Massive parallel sequencing of short tandem repeats in the Korean population*. Electrophoresis, 2018. **39**(21): p. 2702-2707.
272. Phillips, C., et al., *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, 2018. **39**(21): p. 2708-2724.
273. Salvador, J.M., et al., *Filipino DNA variation at 12 X-chromosome short tandem repeat markers*. Forensic Sci Int Genet, 2018. **36**: p. e8-e12.
274. Casals, F., et al., *Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations*. Forensic Sci Int Genet, 2017. **30**: p. 66-70.
275. Wendt, F.R., et al., *Flanking region variation of ForenSeq™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans*. Forensic Science International: Genetics, 2017. **28**: p. 146-154.
276. SWGDAM. *Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*. 2017; Available from: <https://www.swgdam.org/publications>.
277. Update, N.R.C.U.C.o.D.F.S.A., *The Evaluation of Forensic DNA Evidence*. 1996, Washington, DC: National Academy Press.
278. SWGDAM. *Interpretation Guidelines for Y-Chromosome STR Typing*. 2014; Available from: <https://www.swgdam.org/publications>.
279. Jobling, M.A., A. Pandya, and C. Tyler-Smith, *The Y chromosome in forensic analysis and paternity testing*. International Journal of Legal Medicine, 1997. **110**(3): p. 118-124.
280. Ballantyne, K.N., et al., *Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats*. Human Mutation, 2014. **35**(8): p. 1021-1032.
281. Brenner, C.H., *Fundamental problem of forensic mathematics—The evidential value of a rare haplotype*. Forensic Science International: Genetics, 2010. **4**(5): p. 281-291.
282. Buckleton, J.S., M. Krawczak, and B.S. Weir, *The interpretation of lineage markers in forensic DNA testing*. Forensic science international. Genetics, 2011. **5**(2): p. 78-83.
283. Andersen, M.M., et al., *Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory*. Forensic Science International: Genetics, 2013. **7**(2): p. 264-271.
284. Andersen, M.M. and D.J. Balding, *How convincing is a matching Y-chromosome profile?* PLOS Genetics, 2017. **13**(11): p. e1007028.
285. *The Public Authority for Civil Information*.
286. *The Nations Online Project*. Available from: <https://www.nationsonline.org/>.
287. Casey, M.S., *The History of Kuwait*. 2007, Westport, Connecticut: Greenwood Press.
288. Potter, L.G.E., *Sectarian Politics in the Persian Gulf*. 1st Edition ed. 2014: London: Hurst; New York: Oxford University Press.
289. Alsmadi, O., et al., *Genetic substructure of Kuwaiti population reveals migration history*. PLoS One, 2013. **8**(9): p. e74913.
290. Theyab, J.B., S. Al-Bustan, and M.H. Crawford, *The genetic structure of the Kuwaiti population: mtDNA Inter- and intra-population variation*. Hum Biol, 2012. **84**(4): p. 379-403.
291. Triki-Fendri, S., et al., *Genetic structure of Kuwaiti population revealed by Y-STR diversity*. Annals of Human Biology, 2010. **37**(6): p. 827-835.

292. Triki-Fendri, S., et al., *Genetic structure of the Kuwaiti population revealed by paternal lineages*. Am J Hum Biol, 2016. **28**(2): p. 203-212.
293. Mohammad, T., et al., *Genetic structure of nomadic Bedouin from Kuwait*. Heredity, 2009. **103**(5): p. 425-433.
294. Al-Kandari, Y. and Y. Al-Kandari, *Consanguineous Marriage and its Relationship with Sociocultural Variables in Urban and Bedouin Geographical Regions in Kuwait*. Arabian Humanities, 2018(10).
295. Radovanovic, Z., N. Shah, and J. Behbehani, *Prevalence and Social Correlates of Consanguinity in Kuwait*. Annals of Saudi Medicine, 1999. **19**(3): p. 206-210.
296. Al-Kandari, Y.Y. and D.E. Crews, *The effect of consanguinity on congenital disabilities in the Kuwaiti population*. J Biosoc Sci, 2011. **43**(1): p. 65-73.
297. Al-Awadi, S.A., et al., *Consanguinity among the Kuwaiti population*. Clinical Genetics, 1985. **27**(5): p. 483-486.
298. Alenizi, M., et al., *STR data for the AmpF ℓ STR $^{\circ}$ Identifiler $^{\circ}$ loci in Kuwaiti population*. Legal Medicine, 2008. **10**(6): p. 321-325.
299. Al-Enizi, M., et al., *Population genetic analyses of 15 STR loci from seven forensically-relevant populations residing in the state of Kuwait*. Forensic Sci Int Genet, 2013. **7**(4): p. e106-7.
300. Al-enizi, M., et al., *Population data on 25 autosomal STRs for 500 unrelated Kuwaitis*. Forensic Science International: Genetics, 2014. **12**: p. 126-127.
301. *PrepFiler Express $^{\text{TM}}$ and PrepFiler Express BTA $^{\text{TM}}$ Forensic DNA Extraction Kits User Guide*. Available from: http://tools.thermofisher.com/content/sfs/brochures/cms_086049.pdf.
302. *PowerQuant $^{\circ}$ System instructions for Use of Products*. Available from: <https://www.promega.co.uk/resources/protocols/technical-manuals/101/powerquant-system-protocol/>.
303. Promega. *PowerPlex $^{\circ}$ Fusion System for Use on the Applied Biosystems $^{\circ}$ Genetic Analyzers*. Available from: <https://www.promega.com/~media/Files/Resources/Protocols/Technical%20Manuals/101/PowerPlex%20Fusion%20System%20Protocol.pdf>.
304. *PowerPlex $^{\circ}$ Y23 System for Use on the Applied Biosystems $^{\circ}$ Genetic Analyzers*. Available from: <https://www.promega.co.uk/~media/files/resources/protocols/technical-manuals/101/powerplex-y23-system-protocol.pdf?la=en>.
305. *ForenSeq $^{\text{TM}}$ DNA Signature Prep Reference Guide System Instructions for Use of Products*. Available from: <https://verogen.com/support/>.
306. Woerner, A.E., J.L. King, and B. Budowle, *Fast STR allele identification with STRait Razor 3.0*. Forensic Sci Int Genet, 2017. **30**: p. 18-23.
307. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics (Oxford, England), 2009. **25**(14): p. 1754-1760.
308. Thorvaldsdóttir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*. Briefings in Bioinformatics, 2012. **14**(2): p. 178-192.
309. Excoffier, L. and H.E. Lischer, *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows*. Mol Ecol Resour, 2010. **10**(3): p. 564-7.
310. Gouy, A. and M. Zieger, *STRAF-A convenient online tool for STR data evaluation in forensic genetics*. Forensic Sci Int Genet, 2017. **30**: p. 148-151.
311. Fisher, R.A., *Standard calculations for evaluating a blood-group system*. Heredity (Edinb), 1951. **5**(1): p. 95-102.
312. Brenner CH, M.J., *Paternity index calculations in single locus hypervariable DNA probes: validation and other studies*. In: Proceedings for the international symposium on human identification 1989, 1990. **Promega, Madison**: p. pp 21–52.

313. Ohno, Y., I.M. Sebetan, and S. Akaishi, *A simple method for calculating the probability of excluding paternity with any number of codominant alleles*. Forensic Sci Int, 1982. **19**(1): p. 93-8.
314. Botstein, D., et al., *Construction of a genetic linkage map in man using restriction fragment length polymorphisms*. American Journal of Human Genetics, 1980. **32**(3): p. 314-331.
315. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*. Genetics, 2000. **155**(2): p. 945-959.
316. Evanno, G., S. Regnaut, and J. Goudet, *Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study*. Mol Ecol, 2005. **14**(8): p. 2611-20.
317. Paschou, P., et al., *Tracing sub-structure in the European American population with PCA-informative markers*. PLoS genetics, 2008. **4**(7): p. e1000114-e1000114.
318. Paschou, P., et al., *PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations*. PLOS Genetics, 2007. **3**(9): p. e160.
319. Chakraborty, R., *Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing*. Human Biology, 1992. **64**(2): p. 141-159.
320. Rousset, F., *genepop'007: a complete re-implementation of the genepop software for Windows and Linux*. Molecular Ecology Resources, 2008. **8**(1): p. 103-106.
321. *Snipper App suite version 2.5*. Available from: <http://mathgene.usc.es/snipper/>.
322. Santos, C., et al., *Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data*, in *Forensic DNA Typing Protocols*, W. Goodwin, Editor. 2016, Springer New York: New York, NY. p. 255-285.
323. Wang, Z., et al., *Population genetics of 30 insertion-deletion polymorphisms in two Chinese populations using Qiagen Investigator® DIPplex kit*. Forensic Science International: Genetics, 2014. **11**: p. e12-e14.
324. Almheiri, M., et al., *Study of insertion-deletion polymorphisms (indels) in the UAE population*. Forensic Science International: Genetics Supplement Series, 2019. **7**(1): p. 561-563.
325. Al-Snan, N.R., et al., *Population genetics of 30 insertion/deletion polymorphisms in the Bahraini population*. in press.
326. *CLUMPAK - Clustering Markov Packager Across K*. Available from: <http://clumpak.tau.ac.il/>.
327. Kopelman, N.M., et al., *Clumpak: a program for identifying clustering modes and packaging population structure inferences across K*. Molecular ecology resources, 2015. **15**(5): p. 1179-1191.
328. Earl, D.A. and B.M. vonHoldt, *STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method*. Conservation Genetics Resources, 2012. **4**(2): p. 359-361.
329. *Structure Harvester*. Available from: <http://taylor0.biology.ucla.edu/structureHarvester/>.
330. Poulsen, L., et al., *Population and forensic data for three sets of forensic genetic markers in four ethnic groups from Iran: Persians, Lurs, Kurds and Azeris*. Forensic Sci Int Genet, 2015. **17**: p. 43-46.
331. Tomas, C., et al., *Results for five sets of forensic genetic markers studied in a Greek population sample*. Forensic Science International: Genetics, 2015. **16**(Supplement C): p. 132-137.
332. Saiz, M., et al., *Allelic frequencies and statistical data from 30 INDEL loci in Uruguayan population*. Forensic Science International: Genetics, 2014. **9**: p. e27-e29.
333. *Qiagen supplementary material: Population data for analysis of results from the Investigator DIPplex kit*. Available at <http://www.qiagen.com/literature/render.aspx?id=200253>.
334. Guo, S.W. and E.A. Thompson, *Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles*. Biometrics, 1992. **48**(2): p. 361-372.
335. Bonferroni, C.E., *Teoria statistica delle classi e calcolo delle probabilit *. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936. **8**: p. 3-62.

336. Excoffier, L., P.E. Smouse, and J.M. Quattro, *Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data*. Genetics, 1992. **131**(2): p. 479-491.
337. Wright, S., *The genetical structure of populations*. Ann Eugen, 1951. **15**(4): p. 323-54.
338. Nelson, M.S., et al., *Detection of a primer-binding site polymorphism for the STR locus D16S539 using the Powerplex 1.1 system and validation of a degenerate primer to correct for the polymorphism*. J Forensic Sci, 2002. **47**(2): p. 345-9.
339. Wang, L., et al., *A comparative study of insertion/deletion polymorphisms applied among Southwest, South and Northwest Chinese populations using Investigator® DIPplex*. Forensic Science International: Genetics, 2016. **21**: p. 10-14.
340. Akhteruzzaman, S., et al., *Genetic polymorphism of 30 InDel markers for forensic use in Bangladeshi population*. Forensic Science International: Genetics Supplement Series, 2013. **4**(1): p. e348-e349.
341. A., A.A.S., et al., *Consanguinity among the Kuwaiti population*. Clinical Genetics, 1985. **27**(5): p. 483-486.
342. Slatkin, M., *Linkage disequilibrium — understanding the evolutionary past and mapping the medical future*. Nature reviews. Genetics, 2008. **9**(6): p. 477-485.
343. Pereira, R., et al., *Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing*. PLoS One, 2012. **7**(1): p. e29684.
344. Hares, D.R., *Selection and implementation of expanded CODIS core loci in the United States*. Forensic Science International: Genetics, 2015. **17**: p. 33-34.
345. Ensenberger, M.G., et al., *Developmental validation of the PowerPlex® Fusion 6C System*. Forensic Science International: Genetics, 2016. **21**: p. 134-144.
346. Teebi, A.S., *Autosomal recessive disorders among Arabs: an overview from Kuwait*. Journal of medical genetics, 1994. **31**(3): p. 224-233.
347. Balding, D.J., *Weight-of-Evidence for Forensic DNA Profiles*. 2005: John Wiley & Sons.
348. Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies*. Genetics, 2003. **164**(4): p. 1567-87.
349. Hubisz, M.J., et al., *Inferring weak population structure with the assistance of sample group information*. Mol Ecol Resour, 2009. **9**(5): p. 1322-32.
350. John, S.E., et al., *Kuwaiti population subgroup of nomadic Bedouin ancestry-Whole genome sequence and analysis*. Genom Data, 2015. **3**: p. 116-27.
351. Amigo, J., et al., *pop.STR—An online population frequency browser for established and new forensic STRs*. Forensic Science International: Genetics Supplement Series, 2009. **2**(1): p. 361-362.
352. Phillips, C., et al., *Global population variability in Qiagen Investigator HDplex STRs*. Forensic Sci Int Genet, 2014. **8**(1): p. 36-43.
353. Phillips, C., et al., *Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel*. Forensic Sci Int Genet, 2011. **5**(3): p. 155-69.
354. *pop.STR*. Available from: <http://spsmart.cesga.es/popstr.php>.
355. Dixit, S., et al., *Forensic genetic analysis of population of Madhya Pradesh with PowerPlex Fusion 6C™ Multiplex System*. International Journal of Legal Medicine, 2019. **133**(3): p. 803-805.
356. El Andari, A., et al., *Population genetic data for 23 STR markers from Lebanon*. Forensic Sci Int Genet, 2013. **7**(4): p. e108-13.
357. Hedjazi, A., et al., *Allele frequencies for 15 autosomal STR loci in Fars province population, southwest of Iran*. Leg Med (Tokyo), 2013. **15**(4): p. 226-8.

358. Khubrani, Y.M., J.H. Wetton, and M.A. Jobling, *Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity*. Forensic Sci Int Genet, 2019. **39**: p. 97-102.
359. Alsafiah, H.M., et al., *Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler((R)) PCR amplification kit*. Forensic Sci Int Genet, 2017. **31**: p. e59-e61.
360. Perez-Miranda, A.M., et al., *Qatari DNA variation at a crossroad of human migrations*. Hum Hered, 2006. **61**(2): p. 67-79.
361. Alshamali, F., et al., *STR population diversity in nine ethnic populations living in Dubai*. Forensic Sci Int, 2005. **152**(2-3): p. 267-79.
362. Jones, R.J., et al., *Population data for 21 autosomal short tandem repeat markers in the Arabic population of the United Arab Emirates*. Forensic Sci Int Genet, 2017. **28**: p. e41-e42.
363. Al-Snan, N.R., et al., *Population genetic data of the 21 autosomal STRs included in GlobalFiler kit of a population sample from the Kingdom of Bahrain*. PLoS One, 2019. **14**(8): p. e0220620.
364. Farhan, M.M., et al., *Population genetic data for 20 autosomal STR loci in an Iraqi Arab population: Application to the identification of human remains*. Forensic Sci Int Genet, 2016. **25**: p. e10-e11.
365. Nei, M., F. Tajima, and Y. Tateno, *Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data*. J Mol Evol, 1983. **19**(2): p. 153-70.
366. Takezaki, N., M. Nei, and K. Tamura, *POPTREEW: Web Version of POPTREE for Constructing Population Trees from Allele Frequency Data and Computing Some Other Quantities*. Molecular Biology and Evolution, 2014. **31**(6): p. 1622-1624.
367. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms*. Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
368. Butler, J.M., *New resources for the forensic genetics community available on the NIST STRBase website*. Forensic Science International: Genetics Supplement Series, 2008. **1**(1): p. 97-99.
369. Borsuk, L.A., et al., *Sequence-based US population data for the SE33 locus*. Electrophoresis, 2018. **39**(21): p. 2694-2701.
370. *STRidER; Forensic STR Sequence Structure Guide*.
371. *NCBI; Variation Viewer*
372. Gettings, K.B., et al., *STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci*. Forensic Sci Int Genet, 2017. **31**: p. 111-117.
373. Khubrani, Y.M., et al., *Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia*. Forensic Sci Int Genet, 2019. **43**: p. 102164.
374. Alsafiah, H.M., et al., *Sequence-based Saudi population data for the SE33 locus*. Forensic Science International: Genetics Supplement Series, 2019.
375. Durrett, R. and V. Limic, *On the quantity and quality of single nucleotide polymorphisms in the human genome*. Stochastic Processes and their Applications, 2001. **93**(1): p. 1-24.
376. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat*. American journal of human genetics, 1998. **62**(6): p. 1408-1415.
377. Kidd, K.K., et al., *Selecting microhaplotypes optimized for different purposes*. ELECTROPHORESIS, 2018. **39**(21): p. 2815-2823.
378. Oldoni, F., K.K. Kidd, and D. Podini, *Microhaplotypes in forensic genetics*. Forensic Science International: Genetics, 2019. **38**: p. 54-69.
379. Kidd, K.K., et al., *Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics*. Forensic Science International: Genetics, 2014. **12**: p. 215-224.
380. King, J.L., et al., *STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems*. Forensic Science International: Genetics, 2017. **29**: p. 21-28.
381. Crow, J.F. and M. Kimura, *An Introduction to Population Genetics Theory*. 1970: New York: Harper & Row.

382. Cann, H.M., et al., *A human genome diversity cell line panel*. Science, 2002. **296**(5566): p. 261-2.
383. Guo, F., et al., *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, 2017. **31**: p. 135-148.
384. Browning, S.R. and E.A. Thompson, *Detecting rare variant associations by identity-by-descent mapping in case-control studies*. Genetics, 2012. **190**(4): p. 1521-1531.
385. Nelson, M.R., et al., *An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people*. Science (New York, N.Y.), 2012. **337**(6090): p. 100-104.
386. Kidd, K.K. and W.C. Speed, *Criteria for selecting microhaplotypes: mixture detection and deconvolution*. Investigative Genetics, 2015. **6**(1): p. 1.
387. Leal, S.M., *Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium*. Genetic epidemiology, 2005. **29**(3): p. 204-214.
388. Bazyar, R., et al., *Introduction and Assessment of 25 Informative Autosomal SNPs for Human Genetically Identification*. Journal of Military Medicine, 2018. **20**(2): p. 201-211.
389. Purps, J., et al., *Validation of a combined autosomal/Y-chromosomal STR approach for analyzing typical biological stains in sexual-assault cases*. Forensic Science International: Genetics, 2015. **19**: p. 238-242.
390. *Y-DNA Haplogroup Predictor - NEVGEN*. Available from: <https://www.nevgen.org/>.
391. Desmarais, D., et al., *Development of a highly polymorphic STR marker for identity testing purposes at the human androgen receptor gene (HUMARA)*. J Forensic Sci, 1998. **43**(5): p. 1046-9.
392. Kishida, T., et al., *Duplex PCR of the Y-27H39 and HPRT loci with reference to Japanese population data on the HPRT locus*. Nihon Hoigaku Zasshi, 1997. **51**(2): p. 67-9.
393. Krüger, J., et al., *Zur Verwendung des Polymorphismus der sauren Erythrocytenphosphatase bei der Vaterschaftsbegutachtung*. Deutsche Zeitschrift für die gesamte gerichtliche Medizin, 1968. **64**(2): p. 127-146.
394. Al-Snan, N.R., et al., *Geographical structuring and low diversity of paternal lineages in Bahrain shown by analysis of 27 Y-STRs*. Molecular Genetics and Genomics, 2020. **295**(6): p. 1315-1324.
395. Ferri, G., et al., *Molecular characterisation and population genetics of the DYS458 .2 allelic variant*. Forensic Science International: Genetics Supplement Series, 2008. **1**(1): p. 203-205.
396. Capelli, C., et al., *Phylogenetic evidence for multiple independent duplication events at the DYS19 locus*. Forensic Science International: Genetics, 2007. **1**(3): p. 287-290.
397. Glusman, G., et al., *Kaviar: an accessible system for testing SNV novelty*. Bioinformatics, 2011. **27**(22): p. 3216-3217.
398. Fakhro, K.A., et al., *The Qatar genome: a population-specific tool for precision medicine in the Middle East*. Human Genome Variation, 2016. **3**: p. 16016.
399. Hunter-Zinck, H., et al., *Population Genetic Structure of the People of Qatar*. American Journal of Human Genetics, 2010. **87**(1): p. 17-25.
400. Rodriguez-Flores, J.L., et al., *Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations*. Genome Research, 2016. **26**(2): p. 151-162.

Appendix 1 – Participant information sheet and consent

A1.1 DIPplex Project



Participant Information Sheet

Name of department: Department of Pure and Applied Chemistry

Title of the study: Validation study of insertion and deletion polymorphisms as forensic markers in Kuwaiti population.

Introduction

Chief Investigator:

Name: Dr Penny Haddrill
Status: PhD supervisor
Tel.: 0141 548 4337
E-mail: penny.haddrill@strath.ac.uk

Co-Investigator:

Name: Mahdi Haidar
Status: PhD Student
Tel.: UK: +44 (0)141 548 5992 ext.: 4519
Kuwait: +965 66150080
E-mail: mahdi.haidar@strath.ac.uk

What is the purpose of this investigation?

The aim of the study is to validate a set of insertion/deletion (indel) markers for forensic use in the Kuwaiti population. These are alternative markers to standard DNA profiling methods, which would be useful in allowing investigators to analyse difficult forensic samples, such as degraded samples, as well as helping to solve paternity cases. It is envisaged that the project will develop an informative set of insertion/deletion markers that can be used as a complementary technique alongside existing systems for forensic analysis of biological samples.

Do you have to take part?

This study involves participants donating blood samples. Participation in this research is entirely voluntary and you have the right to refuse to participate without giving a reason; refusing to participate will not negatively affect you in any way. You also have the right to withdraw from this research at any time up to the completion of the project without detriment and without giving a reason, and ask for your data to be destroyed.

What will you do in the project?

You will be required to provide small blood sample, collected by one of the trained investigators using a sterile lancet to make a 'finger prick', and blood deposited onto a collection card. DNA will be extracted from the samples and used to validate insertion/deletion polymorphisms as forensic markers in the Kuwaiti populations. This work will take place in the Kuwait Identification Centre, General Department of Criminal Evidence. No payments will be provided for taking part in this research.

Why have you been invited to take part?

To study insertion and deletion markers to be applied as a routine analysis in the Kuwait Forensic Centre, it is essential that these samples are collected from volunteers who originate from Kuwaiti populations; these samples cannot be simulated. All Kuwaiti participants over 18 in age are welcomed to provide blood samples.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263



What are the potential risks to you in taking part?

The collection of blood using a lancet carries a small risk of infection to the donor, which will be minimised with the use of alcohol wipes and participants will be given a plaster following donation. In addition, participants will only be asked to handle their own samples. In this way, the chances of infection will be substantially reduced and will be negligible.

What happens to the information in the project?

All participants' information will be kept confidential. In order to preserve anonymity, samples will be labelled with a code that does not contain any information allowing the participant to be identified, except for by the investigators. This will allow a participant's data to be identified, removed and destroyed should they subsequently wish to withdraw from the study. Once the project is completed, the codes linking identity with the samples will be deleted and after this point participants will no longer be able to withdraw from the study, as the destruction of the codes will mean that their data will be fully anonymised.

Samples will be stored in a locked cabinet that will only be accessible by the named investigators, in a storage room in a restricted area in the General Department of Forensic Evidence. This is a restricted government centre accessed only by authorised personnel. All data outputs will be stored electronically on password-protected computers only accessible by the investigators. Electronic data may be retained indefinitely in this form, and no information will be put onto any databases. All blood samples will be securely disposed prior to the end date of the research, which is estimated to be around the 1st June 2018.

The outcomes of this study will be written into a PhD thesis by co-investigator Mahdi Haidar. In addition, it is envisaged that the outcomes of this study will be written into journal/conference publication(s). In neither form of publication will any information be included that could allow the participants to be identified.

The University of Strathclyde is registered with the Information Commissioner's Office who implements the Data Protection Act 1998. All personal data on participants will be processed in accordance with the provisions of the Data Protection Act 1998.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.

What happens next?

If you are happy to be involved in the project, please sign the consent form provided to confirm this. Please note that participants will not be informed of the specific results of the tests but will be supplied with electronic copies of any resulting publications upon request.

If you do not want to be involved in the project, we would like to thank you for your attention.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263



Researcher contact details:

Mahdi Haidar

UK:

Centre for Forensic Science
Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: +44 (0)141 548 5992 ext.: 4519

Kuwait:

Kuwait Identification Centre
General Department of Criminal Evidence
Ministry of Interior
Al-Dajeej
Kuwait

Mobile: +965 66150080

E-mail:

mahdi.haidar@strath.ac.uk

Chief Investigator details:

Dr Penny Haddrill,
Centre for Forensic Science
Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: 0141 548 4377
E-mail: penny.haddrill@strath.ac.uk

This investigation was granted ethical approval by the Department of Pure and Applied Chemistry Ethics Committee.

If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Secretary to the University Ethics Committee
Research & Knowledge Exchange Services
University of Strathclyde
Graham Hills Building
50 George Street
Glasgow
G1 1QE

Telephone: 0141 548 3707

Email: ethics@strath.ac.uk

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263



Consent Form

Name of department: Department of Pure and Applied Chemistry

Title of the study: Validation study of insertion and deletion polymorphisms as forensic markers in Kuwaiti population.:

- I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences. If I exercise my right to withdraw and I don't want my data to be used, any data which have been collected from me will be destroyed. However, I understand that upon the completion of the study I will no longer be able to withdraw.
- I understand that any information recorded in the investigation will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project
- I understand that I will be asked to donate a blood sample, collected by the investigators using a sterile lancet.
- I consent to the DNA in my sample being analysed
- I understand that all of my biological samples will be securely destroyed within one month of the end of the project, which is estimated to be on the 1st June 2018.
- I consent to the taking of biological samples from me, and understand that they will be the property of Kuwait Forensic DNA department.

(PRINT NAME)	
Signature of Participant:	Date:

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

A1.2 ForenSeq and Fusion 6C projects



Participant Information Sheet

Name of department: Department of Pure and Applied Chemistry

Title of the study: Genetic analysis of the Kuwaiti population using massively parallel sequencing of forensic markers.

Introduction

Chief Investigator:

Name: Dr Penny Haddrill

Status: PhD supervisor

Tel.: 0141 548 4337

E-mail: penny.haddrill@strath.ac.uk

Co-Investigator:

Name: Mahdi Haidar

Status: PhD Student

Tel.: UK: +44 (0)141 548 5992

Kuwait: +965 66150080

E-mail: mahdi.haidar@strath.ac.uk

What is the purpose of this investigation?

The aim of the study is to use massively parallel sequencing (MPS) technology to validate a set of short tandem repeats (STRs) and single nucleotide polymorphism (SNPs) markers in the Kuwaiti population. This will demonstrate how well the markers work, provide data on levels of polymorphism and genetic variation, investigate population substructure and increase the amount of population genetic data for the Kuwait population.

Why have you been invited to take part?

The aim of the project is to study a set of forensic markers using high-throughput sequencing technology in a sample of individuals from Kuwait. It is essential that these samples are collected from volunteers who originate from Kuwaiti populations; these samples cannot be simulated. All Kuwaiti participants over 18 in age are welcomed to provide blood samples.

Do you have to take part?

This study involves participants donating blood samples. Participation in this research is entirely voluntary and you have the right to refuse to participate without giving a reason; refusing to participate will not negatively affect you in any way. You also have the right to withdraw from this research at any time up to the completion of the project (June 2019) without detriment and without giving a reason, and ask for your samples and data to be destroyed.

What will you do in the project?

You will be required to provide small blood sample, collected by a trained investigator using a sterile lancet to make a 'finger prick', and blood deposited onto a collection card. DNA will be extracted from the samples and analysed at the University of Leicester in the UK. No payments will be provided for taking part in this research. When you sign the consent form, you will also be asked to provide information about your ethnicity, which will allow the researchers to look for genetic associations between different population groupings'.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263



What are the potential risks to you in taking part?

The collection of blood using a lancet carries a small risk of infection to the donor, which will be minimised with the use of alcohol wipes and participants will be given a plaster following donation. In addition, participants will only be asked to handle their own samples. In this way, the chances of infection will be substantially reduced and will be negligible.

What happens to the information in the project?

All participants' information will be kept confidential. In order to preserve anonymity, samples will be labelled with a code that does not contain any information allowing the participant to be identified, except for by the investigators. This will allow a participant's data to be identified, removed and destroyed should they subsequently wish to withdraw from the study. Once the project is completed, the codes linking identity with the samples will be deleted and after this point participants will no longer be able to withdraw from the study, as the destruction of the codes will mean that their data will be fully anonymised.

Samples will be stored in a locked cabinet that will only be accessible by the named investigators, in a storage room in a restricted area in the General Department of Forensic Evidence. This is a restricted government centre accessed only by authorised personnel. All the samples will be transported to the University of Strathclyde and stored in a restricted laboratory space with keypad access only for authorised laboratory personnel.

All data outputs will be stored electronically on password-protected computers only accessible by the investigators. Electronic data may be retained indefinitely in this form, and no information will be put onto any databases. All blood samples will be securely disposed within one month of the end date of the research, which is estimated to be around the 1st June 2019.

The outcomes of this study will be written into a PhD thesis by co-investigator Mahdi Haidar. In addition, it is envisaged that the outcomes of this study will be written into journal/conference publication(s). In neither form of publication will any information be included that could allow the participants to be identified.

The University of Strathclyde is registered with the Information Commissioner's Office who implements the Data Protection Act 1998. All personal data on participants will be processed in accordance with the provisions of the Data Protection Act 1998.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.

What happens next?

If you are happy to be involved in the project, please sign the consent form provided to confirm this. Please note that participants will not be informed of the specific results of the tests but will be supplied with electronic copies of any resulting publications upon request.

If you do not want to be involved in the project, we would like to thank you for your attention.

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263



Researcher contact details:

Mahdi Haidar

UK:

Centre for Forensic Science
Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: +44 (0)141 548 5992

Kuwait:

Kuwait Identification Centre
General Department of Criminal Evidence
Ministry of Interior
Al-Dajeej
Kuwait
Mobile: +965 66150080

E-mail:

mahdi.haidar@strath.ac.uk

Chief Investigator details:

Dr Penny Haddrill,
Centre for Forensic Science
Department of Pure and Applied Chemistry
University of Strathclyde
Royal College, 204 George Street, Glasgow, G1 1XW
Telephone: 0141 548 4377
E-mail: penny.haddrill@strath.ac.uk

This investigation was granted ethical approval by the Department of Pure and Applied Chemistry Ethics Committee.

If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Secretary to the University Ethics Committee
Research & Knowledge Exchange Services
University of Strathclyde
Graham Hills Building
50 George Street
Glasgow
G1 1QE

Telephone: 0141 548 3707

Email: ethics@strath.ac.uk

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

Consent Form

Name of department: Department of Pure and Applied Chemistry

Title of the study: Genetic analysis of the Kuwaiti population using massively parallel sequencing of forensic markers:

- I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences. If I exercise my right to withdraw and I don't want my data to be used, any data which have been collected from me will be destroyed. However, I understand that upon the completion of the study I will no longer be able to withdraw.
- I understand that any information recorded in the investigation will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project
- I understand that I will be asked to donate a blood sample, collected by the investigators using a sterile lancet.
- I consent to the DNA in my sample being analysed
- I understand that all of my biological samples will be securely destroyed within one month of the end of the project, which is estimated to be on the 1st June 2019.
- I consent to the taking of biological samples from me, and understand that they will be the property of the University of Strathclyde.

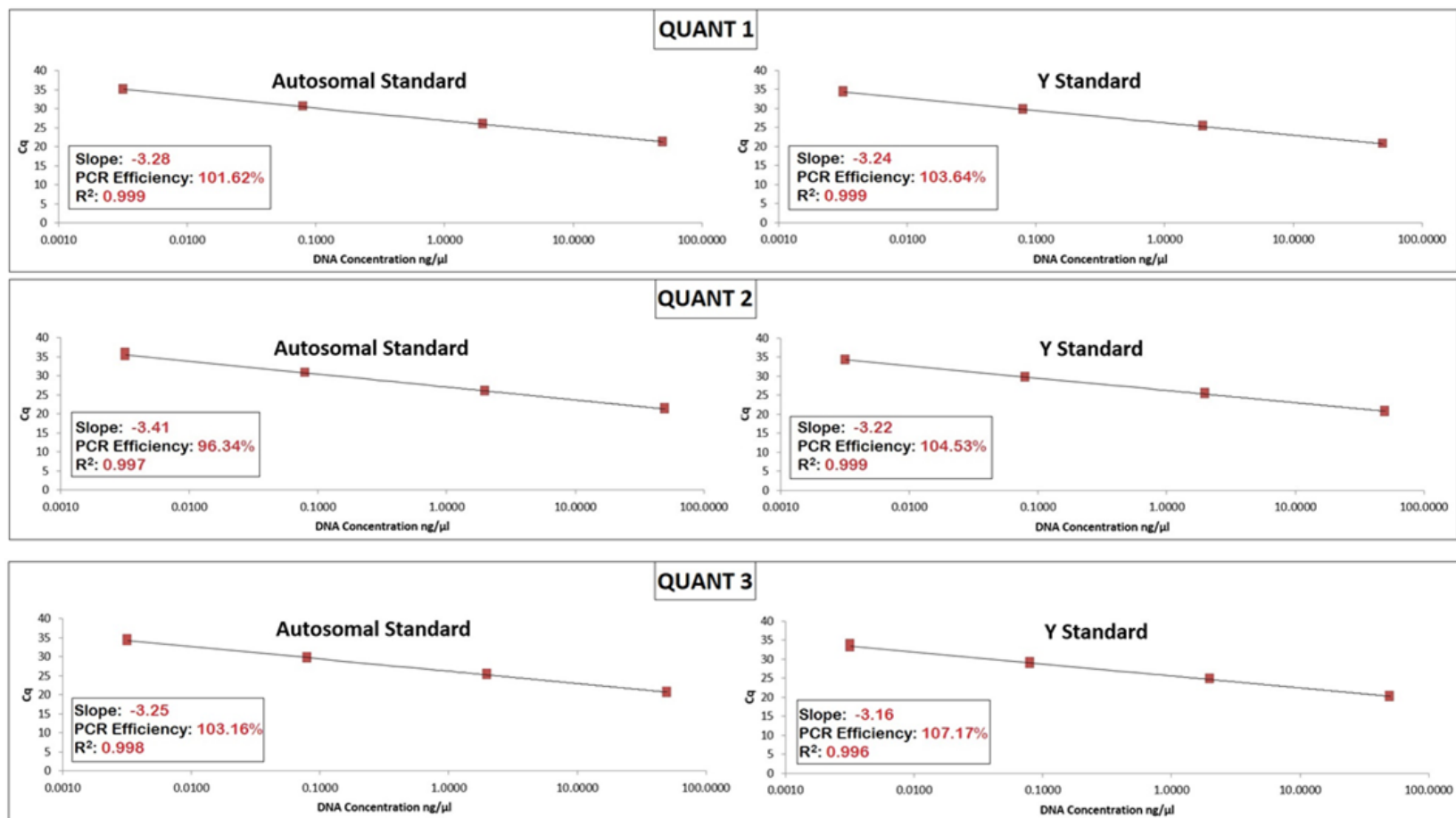
(PRINT NAME)	
Ethnicity:	
Signature of Participant:	Date:

The place of useful learning

The University of Strathclyde is a charitable body, registered in Scotland, number SC015263

Appendix 2 – DNA quantification results

A2.1 PowerQuant™ System Standard curves



A2.2 Autosomal DNA Concentrations

Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)
1	0.57	26	0.16	51	0.30	76	0.31	101	0.37	126	0.35
2	0.67	27	0.25	52	0.62	77	0.84	102	0.21	127	0.20
3	0.40	28	0.51	53	0.36	78	0.64	103	0.49	128	0.82
4	0.15	29	0.18	54	0.35	79	0.68	104	0.61	129	0.23
5	0.49	30	0.10	55	0.25	80	0.74	105	0.26	130	0.43
6	0.64	31	0.66	56	0.23	81	0.85	106	0.37	131	0.20
7	1.03	32	0.53	57	0.22	82	0.97	107	0.42	132	0.32
8	1.53	33	0.18	58	0.32	83	0.31	108	0.27	133	0.32
9	0.24	34	0.46	59	0.86	84	1.58	109	0.70	134	0.50
10	0.64	35	0.38	60	0.32	85	0.62	110	0.27	135	0.59
11	0.84	36	0.34	61	0.62	86	0.67	111	0.58	136	0.75
12	0.50	37	0.33	62	0.51	87	0.83	112	0.56	137	1.03
13	0.22	38	0.38	63	0.27	88	0.21	113	0.91	138	0.47
14	0.24	39	0.11	64	0.55	89	0.57	114	0.36	139	0.29
15	0.59	40	0.25	65	0.40	90	0.27	115	0.83	140	0.41
16	0.25	41	0.37	66	0.88	91	0.16	116	0.25	141	0.94
17	0.61	42	0.50	67	0.34	92	0.44	117	0.67	142	0.38
18	0.66	43	0.43	68	0.40	93	0.44	118	0.49	143	0.49
19	0.13	44	0.48	69	0.61	94	0.48	119	0.29	144	0.57
20	0.40	45	0.34	70	0.62	95	0.39	120	0.44	145	0.57
21	0.76	46	0.52	71	0.21	96	0.32	121	0.51	146	0.36
22	0.46	47	0.33	72	0.74	97	0.76	122	0.29	147	1.14
23	0.43	48	1.21	73	0.61	98	0.47	123	0.64	148	1.28
24	0.45	49	0.20	74	0.24	99	0.44	124	0.22	149	0.79
25	0.23	50	0.27	75	0.20	100	0.25	125	0.25	150	0.68

A2.3 Y-Chromosome Concentrations

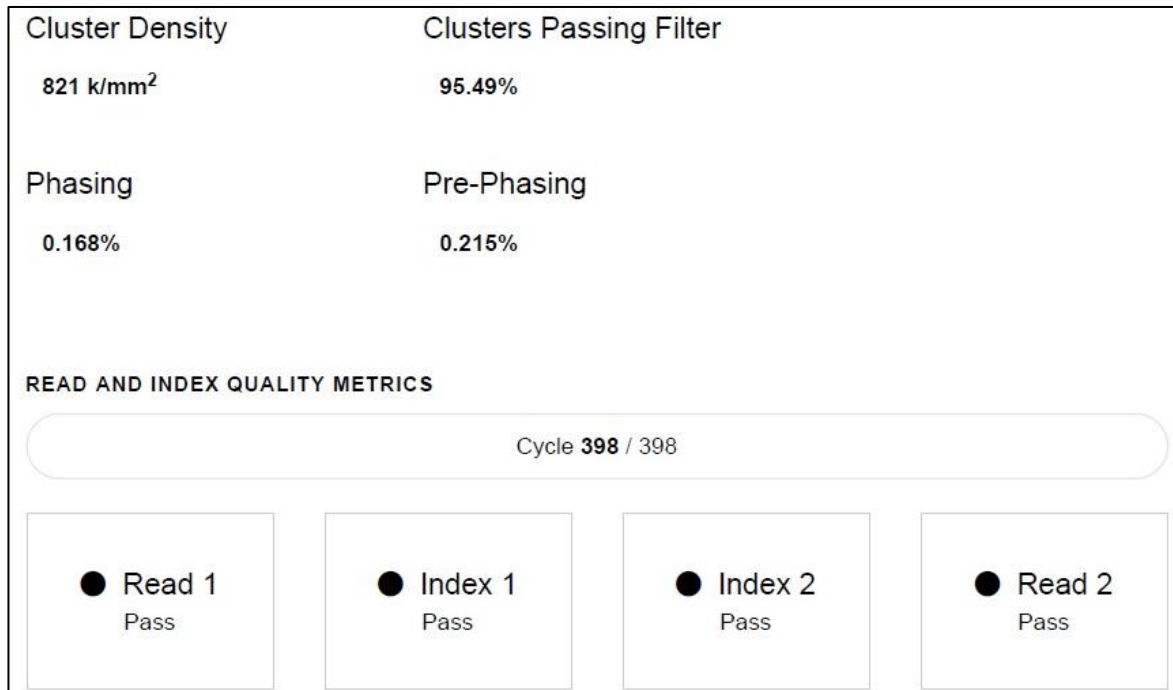
Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)	Sample Name	CONC (ng/μL)
1	-	26	0.14	51	0.32	76	0.20	101	-	126	0.29
2	0.73	27	0.29	52	-	77	-	102	-	127	0.13
3	0.33	28	0.36	53	0.34	78	-	103	-	128	-
4	0.20	29	-	54	0.36	79	-	104	-	129	0.21
5	0.42	30	-	55	0.20	80	-	105	-	130	0.43
6	0.57	31	-	56	0.22	81	-	106	-	131	0.17
7	0.99	32	0.56	57	0.17	82	-	107	-	132	0.31
8	-	33	0.18	58	0.27	83	-	108	-	133	0.31
9	0.22	34	0.35	59	0.61	84	-	109	0.76	134	0.34
10	0.56	35	0.35	60	0.29	85	-	110	0.24	135	0.46
11	0.71	36	0.36	61	0.45	86	-	111	0.47	136	0.39
12	0.59	37	0.22	62	0.57	87	-	112	-	137	0.98
13	0.22	38	0.31	63	-	88	-	113	-	138	0.56
14	0.29	39	0.08	64	0.54	89	-	114	0.36	139	-
15	0.48	40	0.19	65	-	90	-	115	0.77	140	-
16	0.24	41	0.28	66	-	91	-	116	0.24	141	-
17	0.51	42	0.50	67	0.33	92	-	117	-	142	-
18	0.51	43	0.35	68	0.39	93	-	118	-	143	-
19	0.19	44	0.39	69	-	94	-	119	0.25	144	-
20	0.28	45	0.30	70	0.56	95	-	120	0.39	145	-
21	-	46	0.41	71	0.17	96	-	121	0.49	146	-
22	0.41	47	0.27	72	-	97	-	122	-	147	-
23	0.42	48	-	73	-	98	-	123	-	148	-
24	0.38	49	0.21	74	0.19	99	-	124	-	149	0.66
25	0.19	50	0.25	75	0.20	100	-	125	0.24	150	0.75

Appendix 3 – The default stutter filter percentages used by AUS

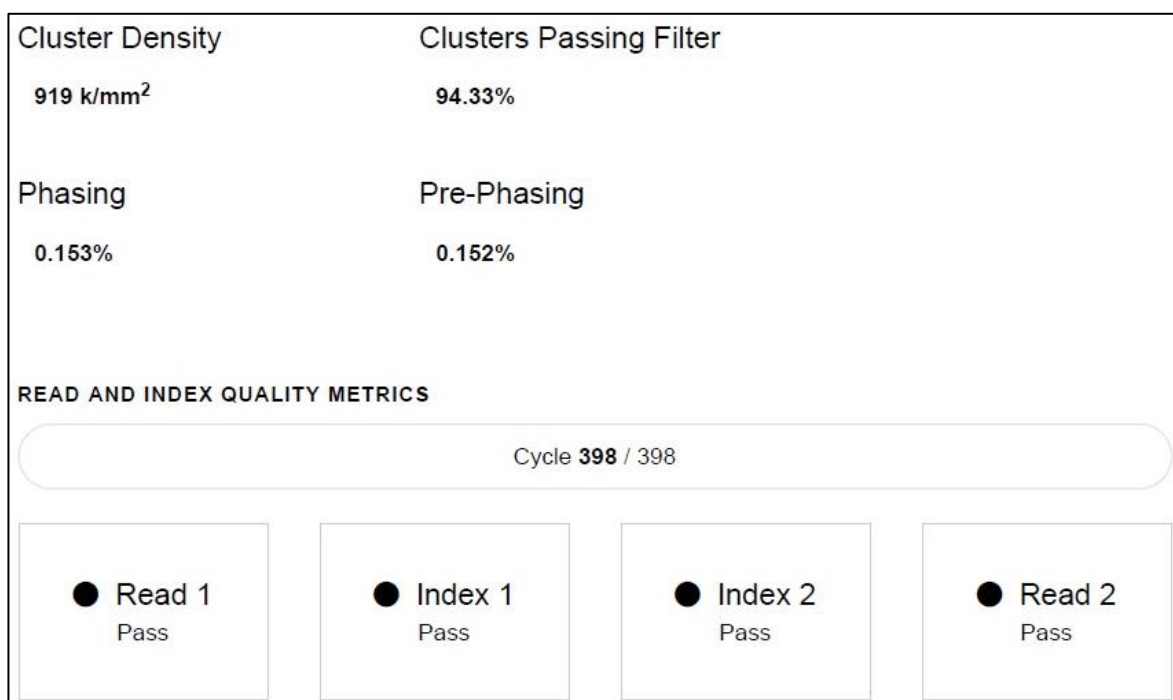
Locus	Stutter Filter (%)
Amelogenin	0
D1S1656	25
TPOX	10
D2S441	7.5
D2S1338	20
D3S1358	15
D4S2408	7.5
FGA	25
D5S818	12.5
CSF1PO	10
D6S1043	12.5
D7S820	10
D8S1179	25
D9S1122	12.5
D10S1248	20
TH01	10
vWA	22
D12S391	33
D13S317	12.5
PentaE	10
D16S539	20
D17S1301	20
D18S51	22
D19S433	12.5
D20S482	15
D21S11	10
PentaD	7.5
D22S1045	20

Appendix 4 – Quality metrics from four MiSeq FGx® runs performed to sequence 376 Kuwaiti samples.

Run 1



Run 2



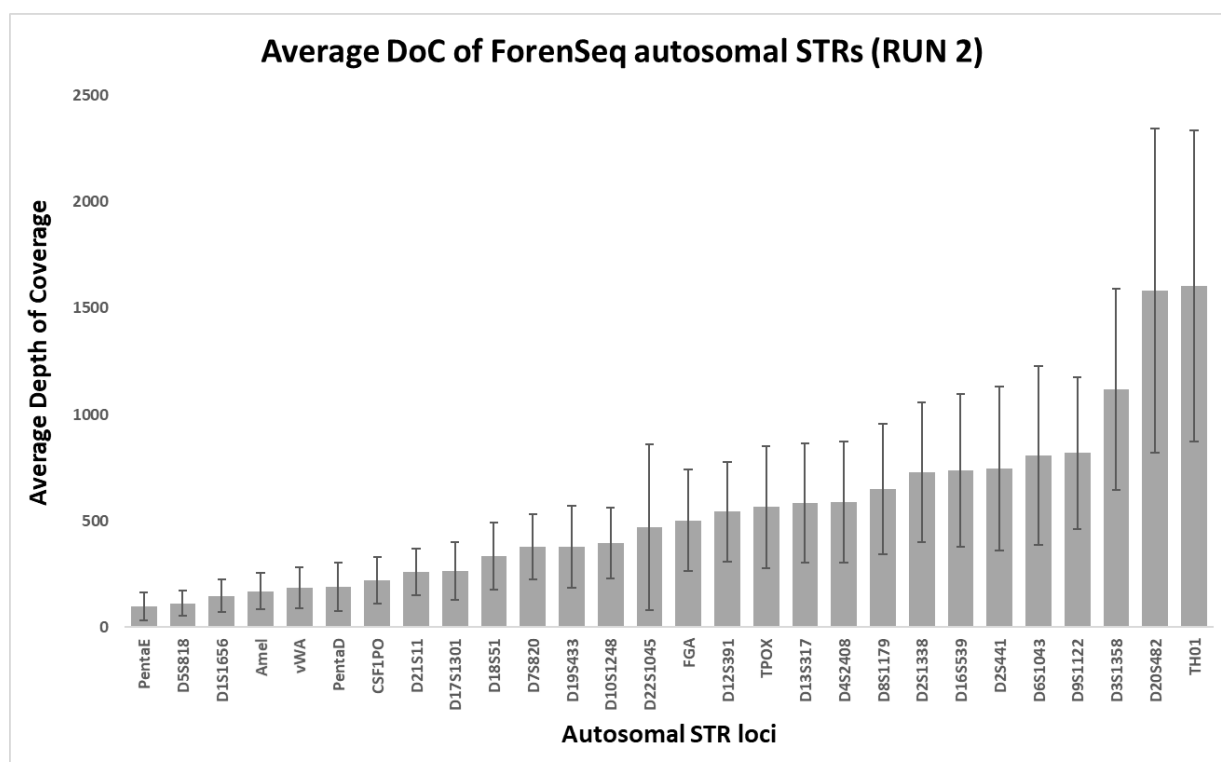
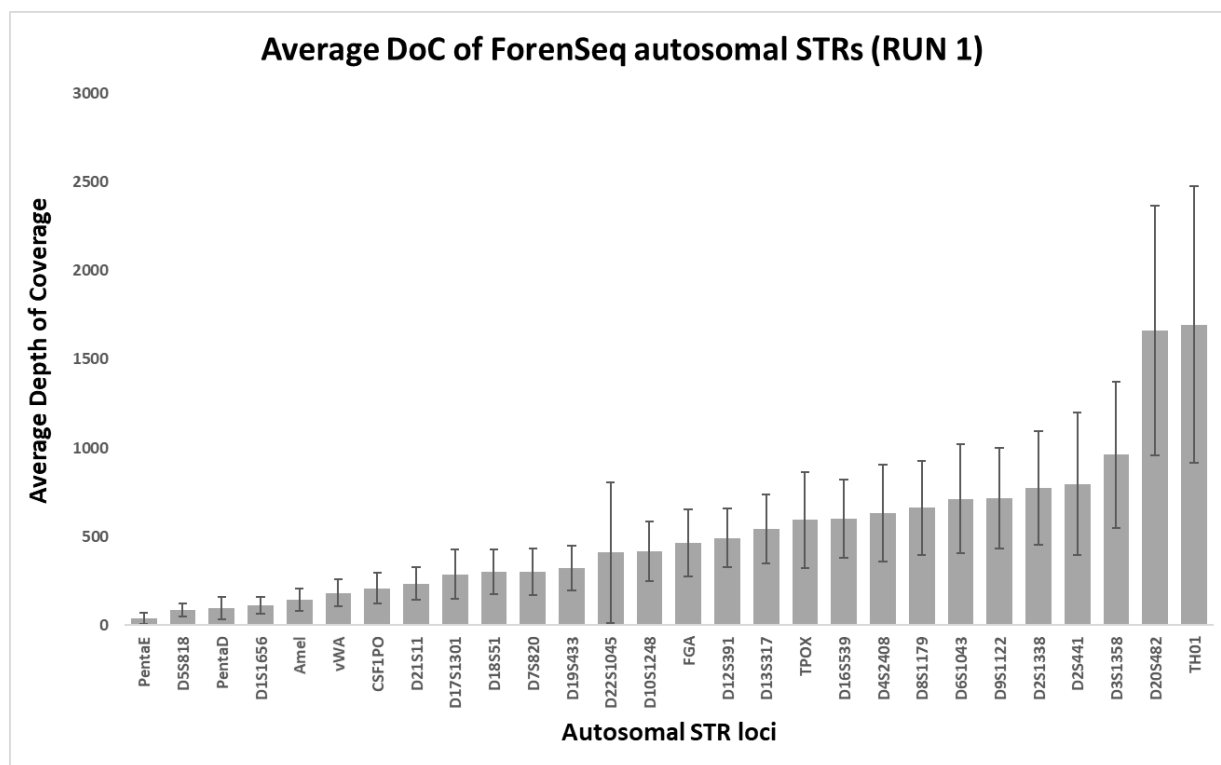
Run 3

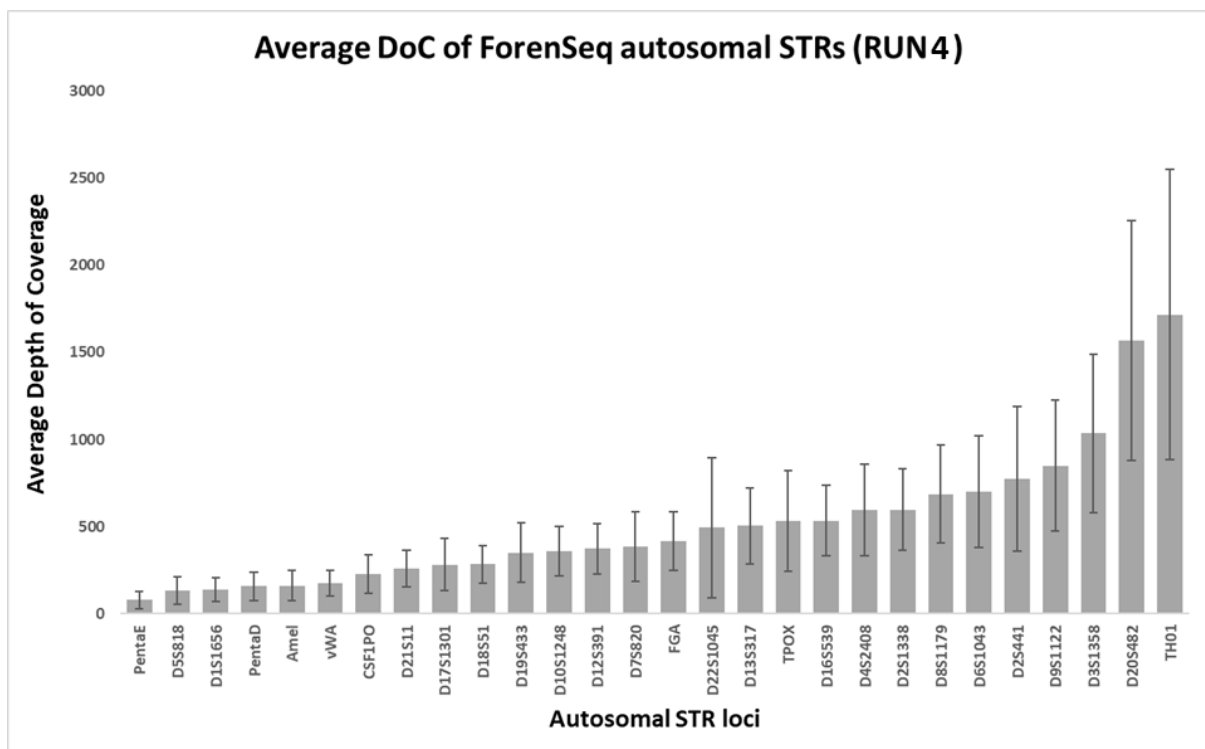
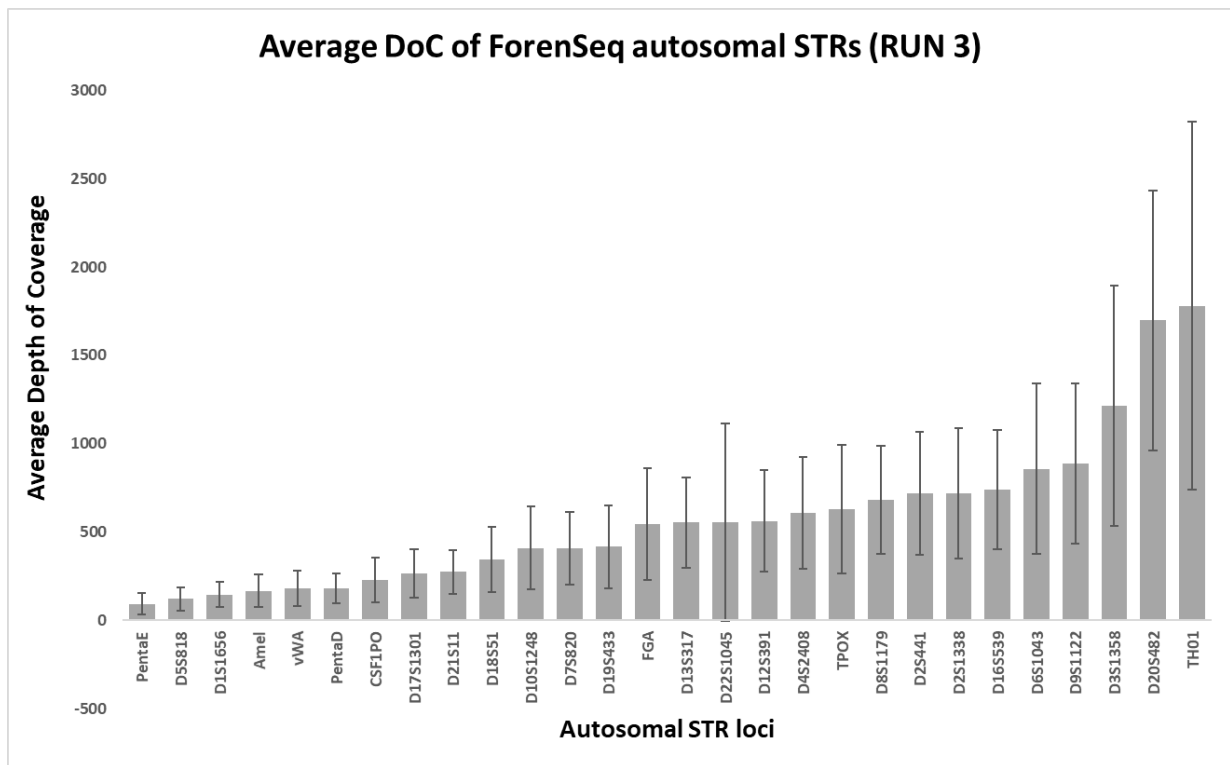
Cluster Density	Clusters Passing Filter
970 k/mm ²	94.29%
Phasing	Pre-Phasing
0.120%	0.164%
READ AND INDEX QUALITY METRICS	
Cycle 398 / 398	
● Read 1 Pass	● Index 1 Pass
● Index 2 Pass	● Read 2 Pass

Run 4

Cluster Density	Clusters Passing Filter
873 k/mm ²	94.78%
Phasing	Pre-Phasing
0.133%	0.235%
READ AND INDEX QUALITY METRICS	
Cycle 398 / 398	
● Read 1 Pass	● Index 1 Pass
● Index 2 Pass	● Read 2 Pass

Appendix 5 – Average depth of coverage (DoC) of 27 autosomal STRs collected using four MiSeq FGx® runs to sequence 376 Kuwaiti samples.





Frequencies of length-based, repeat sequence-based (sequence-based), and flanking sequence-based (flank-based) alleles among 365 Kuwaiti individuals. All the sequences shown are as reported by the UAS flanking sequence report. To avoid confusion with STR repeat region sequences, flanking region sequences are coloured red. Flanking variants are highlighted in bold. Allele nomenclature is in line with ISFG recommendations. Where a single length-based allele can be subdivided into multiple sequence-based or flank-based alleles, the different alleles are assigned alphabetically.

220

Appendix 7 – Allele frequencies of ForenSeq™ DNA Signature Prep Kit novel autosomal STR loci observed in literature.

Frequencies of novel alleles (as identified using STRait Razor v3) observed in the literature (Middle East: ME; NA: North America; EA: East Asia; OC: Oceania; EUR: Europe; SE: South Asia; AF: Africa; SA: Saudi Arabia. The alleles were identified as being novel due to length-based (LB), sequence-based (SB) or flank-based (rs) variants.

Locus	Nomenclature according to Parson et al. 2016	Type	ME	NA	EA	QC	EUR	SE	AF	Khurbrani et al. 2019	Gettings et al. 2017
										SA	Genebank
CSF	CSF1PO [CE 12]-GRCh38-Chr5-150076318-150076389 ATCT ACCT [ATCT]10	SB	0.006	0	0	0	0	0	0	0.022	0
D10S1248	D10S1248 [CE9]-GRCh38-Chr10 129294226-129294318 [GGAA]9	LB	0.009	0	0	0	0	0	0.005	0.011	MH167056.1
D10S1248	D10S1248 [CE 19]-GRCh38-Chr10 129294226-129294318 [GGAA]19	LB	0	0	0	0	0	0	0	0	MH167067.1
D125391	D125391 [CE 18]-GRCh38-Chr12-12296981-12297168 [AGAT]9 [AGAC]8 AGAT	SB	0	0	0	0	0	0	0	0	MH167120.1
D125391	D125391 [CE 18]-GRCh38-Chr12-12296981-12297168 [AGAT]10 [AGAC]7 AGAT 12297179-G (rs138635218)	FB	0	0	0	0	0	0	0	0	MH167123.1
D125391	D125391 [CE 19]-GRCh38-Chr12-12296981-12297168 AGAC [AGAT]11 [AGAC]6 AGAT	SB	0	0	0	0	0	0	0	0	MKS569924.1
D125391	D125391 [CE 19]-GRCh38-Chr12-12296981-12297168 [AGAT]9 [AGAC]10	SB	0	0	0	0	0	0.002	0	0	0
D125391	D125391 [CE 19]-GRCh38-Chr12-12296981-12297168 [AGAT]11 [AGAC]7 AGAT 12297179-G (rs138635218)	FB	0	0	0	0	0	0	0	0	MH167132.1
D125391	D125391 [CE 19]-GRCh38-Chr12-12296981-12297168 [AGAT]12 [AGAC]6 AGAT 12297179-G (rs138635218)	FB	0	0	0	0	0	0	0	0	MH167135.1
D125391	D125391 [CE 20]-GRCh38-Chr12-12296981-12297168 [AGAT]10 [AGAC]10	SB	0.003	0	0	0	0	0	0	0	0
D125391	D125391 [CE 20]-GRCh38-Chr12-12296981-12297168 [AGAT]13 [AGAC]6 AGAT 12297179-G (rs138635218)	FB	0	0	0	0	0	0	0	0	MH167151.1
D125391	D125391 [CE 21]-GRCh38-Chr12-12296981-12297168 [AGAT]15 [AGAC]5 AGAT	SB	0	0	0	0.018	0	0.002	0	0	MKS569929.1
D125391	D125391 [CE 23]-GRCh38-Chr12-12296981-12297168 [AGAT]12 [AGAC]11	SB	0	0	0	0.054	0.013	0.002	0	0.006	MH167177.1
D125391	D125391 [CE 25]-GRCh38-Chr12-12296981-12297168 [AGAT]15 [AGAC]10	SB	0	0.008	0	0	0	0	0	0	MH167190.1
D125391	D125391 [CE 26]-GRCh38-Chr12-12296981-12297168 [AGAT]17 [AGAC]8 AGAT	SB	0	0	0	0	0	0	0	0.006	MH167197.1
D125391	D125391 [CE 27]-GRCh38-Chr12-12296981-12297168 [AGAT]18 [AGAC]8 AGAT	SB	0	0	0.004	0	0	0	0	0	MH167200.1
D135317	D135317 [CE 10]-GRCh38-Chr13-82147986-82148107 [TATC]10 82148001-A; 82148069-T (rs146621667) (rs9546005)	FB	0	0	0	0	0	0	0	0	MH167214.1
D135317	D135317 [CE 10]-GRCh38-Chr13-82147986-82148107 [TATC]10 82148069-T 82148077-82148080 DEL (rs9546005) (rs1442523705)	FB	0.003	0	0	0	0	0	0	0	0
D165539	D165539 [CE 8]-GRCh38-Chr16-86352664-86352781 [GATA]8 86352692-G (rs563997442)	FB	0	0	0	0	0	0	0	0.006	MKS70018.1
D18551	D18551 [CE 15.2]-GRCh38-Chr18-63281662-63281796 [AGAA]15 AG 63281740-G (rs535823682)	FB	0	0	0	0	0	0	0	0	MH167295.1
D18551	D18551 [CE 16.2]-GRCh38-Chr18-63281662-63281796 [AGAA]16 AG 63281740-G (rs535823682)	FB	0	0	0	0	0	0	0.005	0	MH167298.1
D18551	D18551 [CE 17.2]-GRCh38-Chr18-63281662-63281796 [AGAA]17 AG 63281740-G (rs535823682)	FB	0	0	0	0	0	0	0	0	MKS569959.1
D195433	D195433 [CE 9]-GRCh38-Chr19-29926205-29926352 [CCTT]7 CCTA CCTT CTTT CCTT	SB	0	0	0	0	0	0	0.005	0	MH174810.1
D1S1656	D1S1656 [CE 9]-GRCh38-Chr1-230769555-230769682 [TCTA]9	SB	0	0	0	0	0	0.010	0	0	MKS70031.1
D1S1656	D1S1656 [CE 10]-GRCh38-Chr1-230769555-230769682 CCTA [TCTA]9	SB	0	0	0	0	0	0.012	0.005	0	MH174834.1
D1S1656	D1S1656 [CE 14]-GRCh38-Chr1-230769555-230769682 CTTA [TCTA]13	SB	0	0	0	0	0	0	0	0	MKS70032.1
D1S1656	D1S1656 [CE 15.3]-GRCh38-Chr1-230769555-230769682 CCTA [TCTA]12 TCA [TCTA]2	SB	0.009	0	0	0	0	0.002	0	0	MKS70033.1
D1S1656	D1S1656 [CE 17]-GRCh38-Chr1-230769555-230769682 CTTA [TCTA]16	SB	0.003	0	0	0	0	0.002	0	0	MH174859.1
D21S11	D21S11 [CE 27]-GRCh38-Chr21-19181939-19182111 [TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10	SB	0	0	0	0	0	0	0	0	MH174718.1
D21S11	D21S11 [CE 28]-GRCh38-Chr21-19181939-19182111 [TCTA]6 [TCTG]5 [TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10	SB	0	0	0.002	0	0	0	0	0	0
D21S11	D21S11 [CE 28]-GRCh38-Chr21-19181939-19182111 [TCTA]4 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11	SB	0	0.008	0	0	0	0	0	0	0
D21S11	D21S11 [CE 29.2]-GRCh38-Chr21-19181939-19182111 [TCTA]5 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10 TA TACT	SB	0	0	0	0	0	0	0	0	MH174740.1
D21S11	D21S11 [CE 30]-GRCh38-Chr21-19181939-19182111 [TCTA]2 TATA [TCTA]3 [TCTG]5 [TCTA]3 TCA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11	SB	0	0	0	0	0	0	0	0	MH174749.2
D21S11	D21S11 [CE 32]-GRCh38-Chr21-19181939-19182111 [TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]14	SB	0	0.008	0.002	0	0	0	0	0	MKS569898.1
D21S11	D21S11 [CE 35.2]-GRCh38-Chr21-19181939-19182111 [TCTA]6 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]14 TA TCTA	SB	0	0	0	0	0.003	0	0	0	0
D21S11	D21S11 [CE 36]-GRCh38-Chr21-19181939-19182111 [TCTA]11 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11	SB	0	0	0	0	0	0	0	0.006	0
D22S1045	D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 [ATT]12 ACT [ATT]2 37140182-A (rs554502154)	FB	0	0	0	0	0	0	0	0.006	0
D2S1338	D2S1338 [CE 14]-GRCh38-Chr2-218014856-218014964 [GGAA]8 [GGCA]6	SB	0	0	0	0	0	0	0	0	MKS569967.1
D2S1338	D2S1338 [CE 15]-GRCh38-Chr2-218014856-218014964 [GGAA]9 [GGCA]6	SB	0	0	0	0	0.003	0	0	0.006	0
D2S1338	D2S1338 [CE 19]-GRCh38-Chr2-218014856-218014964 [GGAA]11 GGGA [GGCA]7	SB	0	0	0	0	0	0	0	0	MH105137.1
D2S1338	D2S1338 [CE 20]-GRCh38-Chr2-218014856-218014964 [GGAA]2 GGAC [GGAA]11 [GGCA]6	SB	0	0	0	0	0	0	0.005	0	MH105150.1
D2S1338	D2S1338 [CE 24]-GRCh38-Chr2-218014856-218014964 [GGAA]2 GGAC [GGAA]16 [GGCA]5	SB	0	0	0.007	0	0	0	0	0	MKS569981.1
D2S441	D2S441 [CE 13.3]-GRCh38-Chr2-68011918-68012017 [TCTA]4 TCA [TCTA]9	SB	0	0	0	0	0	0	0	0	MKS70009.1
D3S1358	D3S1358 [CE 13]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]9	SB	0	0	0	0	0	0	0	0.006	0
D3S1358	D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]4 [TCTA]11	SB	0	0	0	0	0	0	0	0	MH166978.1
D3S1358	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 TCTC [TCTA]13	SB	0.006	0	0	0	0	0.002	0	0.011	MK990348.1
D3S1358	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]4 [TCTA]12	SB	0.003	0	0	0	0	0.005	0	0	MH166983.1
D3S1358	D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 TCTC [TCTA]14	SB	0	0	0	0	0	0	0	0.006	MK990350.1
D3S1358	D3S1358 [CE 18]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]4 [TCTA]13	SB	0.003	0	0	0	0	0.002	0	0.006	MH166987.1
D3S1358	D3S1358 [CE 19]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]16	SB	0	0	0	0	0	0	0	0	MK990352.1
D3S1358	D3S1358 [CE 20]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]16	SB	0.003	0	0	0	0	0	0	0	MH166990.1
D6S1043	D6S1043 [CE 7]-GRCh38-Chr6-91740160-91740292 [ATCT]7	SB	0	0	0	0	0.003	0	0	0	0
D6S1043	D6S1043 [CE 16]-GRCh38-Chr6-91740160-91740292 [ATCT]4 ATGT [ATCT]11	SB	0	0	0	0	0	0	0.010	0	MKS70038.1
D6S1043	D6S1043 [CE 22]-GRCh38-Chr6-91740160-91740292 [ATCT]5 ATGT [ATCT]16	SB	0	0	0.002	0	0	0	0	0	MH166952.1
D7S820	D7S820 [CE 12]-GRCh38-Chr7-84160191-84160297 [TATC]6 TGTC [TATC]5 84160204-A (rs7789995)	FB	0	0	0	0	0	0	0	0	MKS70004.1
D7S820	D7S820 [CE 13]-GRCh38-Chr7-84160191-84160297 [TATC]13 84160204-A; 84160286-A (rs7789995) (rs16887642)	FB	0	0	0	0	0	0	0	0	MKS70005.1
FGA	FGA [CE 17.2]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG [AAAG]10 AA AAAA [GAAA]3	SB	0	0	0	0	0	0	0	0	MH232604.1
FGA	FGA [CE 21]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG AAAG AAAC [AAAG]11 AGAA AAAA [GAAA]3	SB	0.182	0.097	0.111	0.054	0.177	0.122	0.067	0	0
FGA	FGA [CE 22]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG AAAG AAAC [AAAG]12 AGAA AAAA [GAAA]3	SB	0.160	0.081	0.144	0.054	0.161	0.164	0.192	0	0
FGA	FGA [CE 24]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG AAAG AAAC [AAAG]14 AGAA AAAA [GAAA]3	SB	0.178	0.177	0.197	0.286	0.146	0.224	0.173	0	0
FGA	FGA [CE 25]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG [AAAG]5 AAGG [AAAG]11 AGAA AAAA [GAAA]3	SB	0	0	0	0	0.006	0.005	0	0	0
FGA	FGA [CE 29]-GRCh38-Chr4-154587713-154587823 [GGAA]2 GGAG [AAAG]5 AAGG [AAAG]9 GAAG AGAA AAAA [GAAA]3	SB	0	0	0	0	0	0	0.005	0	0
PentaD	PentaD [CE 8]-GRCh38-Chr21-43636100-43636278 [AAAGA]8 43636273-G (rs186259515)	FB	0	0	0	0	0	0	0	0	MK990337.1
PentaD	PentaD [CE 12]-GRCh38-Chr21-43636100-43636278 [AAAGA]11 AAAGG	SB	0	0	0	0	0	0	0	0	MK990341.1
PentaE	PentaE [CE 15.4]-GRCh38-Chr15-96830996-96831114 [TCTTT]15 TCCT	SB	0.003	0	0	0	0	0	0	0	MH232654.1
PentaE	P4entaE [CE 16.4]-GRCh38-Chr15-96830996-96831114 [TCTTT]16 TCCT	SB	0.009	0	0	0	0	0.002	0	0.011	MN983128.1
PentaE	PentaE [CE 18.4]-GRCh38-Chr15-96830996-96831114 [TCTTT]18 TCCT	SB	0	0	0	0	0	0.002	0	0	0
TH01	TH01 [CE 10.3]-GRCh38-Chr11-2171056-2171127 [AATG]7 ATG [AATG]3	SB	0	0	0	0	0	0	0	0	MK990333.1
TPOX	TPOX [CE 14]-GRCh38-Chr2-1489651-1489689 [AATG]14	SB	0	0	0	0	0	0	0	0	MG988078.1
vWA	vWA [CE 13]-GRCh38-Chr12-5983950-5984049 [TAGA]9 [CAGA]3 TAGA	SB	0	0	0	0	0	0	0	0	MH167073.1
vWA	vWA [CE 13]-GRCh38-Chr12-5983950-5984049 [TAGA]3 TGGA [TAGA]3 [CAGA]4 [TAGA]2 5983970-G (rs75219269)	FB	0	0	0	0	0	0	0	0	MKS569942.1
vWA	vWA [CE 17]-GRCh38-Chr12-5983950-5984049 [TAGA]13 [CAGA]4	SB	0	0	0	0	0	0	0	0	MH167089.1

IDs were given following the work of Borsuk et al. three additional IDs was suggested for these motifs: D6*, D7* and D8*.

238

(continued)

[illegible]

Appendix 9. Allele frequencies for iiSNPs haplotypes.

Frequencies of the iiSNPs and iiSNP-flanking variants among 369 Kuwaiti individuals for 88 of the ForenSeq™ DNA Signature Prep Kit iiSNPs. All the sequences shown are as reported by the UAS flanking region report. iiSNP bases are highlighted in bold, flanking region variants are in red, and novel alleles identified by STRait Razor are in green. Allele nomenclature is according to ISFG recommendations.

Locus	Sequences with flanking region	Nomenclature	Total	Targeted iiSNP			iiSNP-flank		
				Allele	Ob	Freq	Allele	Ob	Freq
rs1005533	GCAAAAAGCAAGAGCCGTGGAATT A AGTCGCCGCTGTT C AGGGGAGGCATAAGGAGCTGGA	rs1005533 [CE A]-GRCh38-Chr20:40858446-40858561	738	A	384	0.5203			
rs1005533	GGACTGGGTGGGCTCGGCAGCTTCCCTGGTCTTGCCCTGCACTCCTCACCCAGC	rs1005533-A							
rs1005533	GCAAAAAGCAAGAGCCGTGGAATT G AGTCGCCGCTGTT C AGGGGAGGCATAAGGAGCTGGA	rs1005533 [CE G]-GRCh38-Chr20:40858446-40858561	738	G	354	0.4797			
rs1005533	GGACTGGGTGGGCTCGGCAGCTTCCCTGGTCTTGCCCTGCACTCCTCACCCAGC	rs1005533-G							
rs10092491	AAACTGAAG C TTTCCTTATAGAGATTTATCCTAGTTAGTTTGC	rs10092491 [CE C]-GRCh38-Chr8:28553546-28553610	738	C	480	0.6504			
rs10092491	AAACTGAAG T TTTCCTTATAGAGATTTATCCTAGTTAGTTTGC	rs10092491-C							
rs10092491	AAACTGAAGTTTTCCTTATAGAGATTTATCCTAGTTAGTTTGC	rs10092491 [CE T]-GRCh38-Chr8:28553546-28553610	738	T	258	0.3496			
rs10092491	AAACTGAAGTTTTCCTTATAGAGATTTATCCTAGTTAGTTTGC	rs10092491-T							
rs1015250	AAAGGTTACTAAGTGATGGAGTT A GGAAAAGAACCCAGGTGTTTATT G CTGTCCCAGTGATTT	rs1015250 [CE G]-GRCh38-Chr9:1823726-1823793	738	G	533	0.7222	G	533	0.7222
rs1015250	TTCA	rs1015250-G							
rs1015250	AAAGGTTACTAAGTGATGGAGTT A GGAAAAGAACCCAGGTGTTTATT C CTGTCCCAGTGATTT	rs1015250 [CE C]-GRCh38-Chr9:1823726-1823793	738	C	205	0.2778	C	48	0.0650
rs1015250	TTCA	rs1015250-C							
rs1015250	AAAGGTTACTAAGTGATGGAGTT G GGAAAAGAACCCAGGTGTTTATT C CTGTCCCAGTGATTT	rs1015250 [CE C]-GRCh38-Chr9:1823726-1823793					C1	157	0.2127
rs1015250	TTCA	rs6475200-G; rs1015250-C							
rs1024116	CAT G CACTTAATAAAGTATGCCTTTGTATTTACTTTTGCTCACTTCCCA	rs1024116 [CE C]-GRCh38-Chr18:77720385-77720433	738	G	371	0.5027			
rs1024116	CATA C ACTTAATAAAGTATGCCTTTGTATTTACTTTTGCTCACTTCCCA	rs1024116-C							
rs1024116		rs1024116 [CE T]-GRCh38-Chr18:77720385-77720433	738	A	367	0.4973			
rs1024116		rs1024116-T							
rs1028528	CTTACTCGACATC A CTGTGTGCAGATCCGCGGAGGT	rs1028528 [CE A]-GRCh38-Chr22:47966528-47966563	738	A	464	0.6287			
rs1028528	CTTACTCGACATC G CTGTGTGCAGATCCGCGGAGGT	rs1028528-A							
rs1028528		rs1028528 [CE G]-GRCh38-Chr22:47966528-47966563	738	G	274	0.3713			
rs1028528		rs1028528-G							
rs10488710	TTTACTGTATTAGGAGTTCCCACTTGTTCTTTTTCTGCAAATGTGGCA G TCGGTTTTATTTT	rs10488710 [CE C]-GRCh38-Chr11:115336442-115336505	734	G	316	0.4305			
rs10488710	TTTACTGTATTAGGAGTTCCCACTTGTTCTTTTTCTGCAAATGTGGCA A CTCGGTTTTATTTT	rs10488710-C							
rs10488710		rs10488710 [CE G]-GRCh38-Chr11:115336442-115336505	734	C	418	0.5695			
rs10488710		rs10488710-G							
rs10495407	TCTGCTTCTGGAGATCTCCACTTCCTCTTGTTGCATTGGATTCTCATTGAA A TCTATTCCAT	rs10495407 [CE A]-GRCh38-Chr1:238275955-238276021	738	A	239	0.3238			
rs10495407	TC	rs10495407-A							
rs10495407	TCTGCTTCTGGAGATCTCCACTTCCTCTTGTTGCATTGGATTCTCATTGAA G TCTATTCCA	rs10495407 [CE G]-GRCh38-Chr1:238275955-238276021	738	G	499	0.6762			
rs10495407	TTC	rs10495407-G							
rs1058083	TTTGCTCAGAGTATCC A GAGTTAGCCACTAGG	rs1058083 [CE A]-GRCh38-Chr13:99385963-99385994	738	A	244	0.3306			
rs1058083	TTTGCTCAGAGTATCC G GAGTTAGCCACTAGG	rs1058083-A							
rs1058083		rs1058083 [CE G]-GRCh38-Chr13:99385963-99385994	738	G	494	0.6694			
rs1058083		rs1058083-G							
rs10773760	TGTTAGCCGTCGGGACCAGCTTCTGTCTGGAAGTTCGTC A AATTGCAGTT A GTCC	rs10773760 [CE A]-GRCh38-Chr12:130277100-130277155	738	A	463	0.6274			
rs10773760	TGTTAGCCGTCGGGACCAGCTTCTGTCTGGAAGTTCGTC A AATTGCAGTT G GTCC	rs10773760-A							
rs10773760		rs10773760 [CE G]-GRCh38-Chr12:130277100-130277155	738	G	275	0.3726			
rs10773760		rs10773760-G							
rs10776839	CAGCTGAGGAGCC C G A TTTGCCGTCAGATCAGAGCC C AGTTGCCCGGTCTGCCG C AGCT	rs10776839 [CE T]-GRCh38-Chr9:134525445-134525507	738	T	358	0.4851	T	358	0.4851
rs10776839	CT	rs7037930-G; rs10776839-T							

rs10776839	CAGCTGAGGAGCCCCAAGGTTGCCGTCAGATCAGAGCCCCAGTTGCCCGGTCTGCCGCAGCTCT	rs10776839 [CE G]-GRCh38-Chr9:134525445-134525507 rs10776839-G	738	G	380	0.5149	G	121	0.1640
rs10776839	CAGCTGAGGAGCCCCGAGGTTGCCGTCAGATCAGAGCCCCAGTTGCCCGGTCTGCCGCAGCTCT	rs10776839 [CE G]-GRCh38-Chr9:134525445-134525507 rs7037930-G; rs10776839-G					G1	258	0.3496
rs10776839	CAGCTGAGGAGCCCCGAGGTGGCCGTCAGATCAGAGCCCCAGTTGCCCGGTCTGCCGCAGCTCT	rs10776839 [CE G]-GRCh38-Chr9:134525445-134525507 rs7037930-G; rs10776839-G; rs1317939293-G					G2	1	0.0014
rs1109037	CCAGTTTCTCCAGAGTGGAAAGACTTTTCATCTCGCACTGGCACCTTGAGACCCCGGGTCTGATGAACTGGGGGG	rs1109037 [CE A]-GRCh38-Chr2:9945582-9945659 rs1109037-A	738	A	343	0.4648	A	201	0.2724
rs1109037	CCAGTTTCTCCAGAGTGGAAAGACTTTTCATCTCGCACTGGCACCTTGAGACCCCGGGTCTGATGAACTGGGAGG	rs1109037 [CE A]-GRCh38-Chr2:9945582-9945659 rs1109037-A; rs1109038-A					A1	142	0.1924
rs1109037	CCAGTTTCTCCGAGTGGAAAGACTTTTCATCTCGCACTGGCACCTTGAGACCCCGGGTCTGATGAACTGGGGGG	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs1109037-G	738	G	395	0.5352	G	287	0.3889
rs1109037	CCAGTTTCTCCGAGTGGAAAGACTTTTCATCTCGCACTGGCACCTTGAGACCCCGGGTCTGATGAACTGGGAGG	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs1109037-G; rs1109038-A					G1	105	0.1423
rs1109037	CCAGTTTCTCGAGTGGAAAGACTTTTCATCTCGCACTGGCACCTTGAGACCCCGGGTCTGATGAACTGGGAGG	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs1109037-G; rs1109038-A; rs999755320-T					G2	2	0.0027
rs1109037	CCAGTTTCTCCGAGTGGAAAGACTTTTCATCTCGCACTGGCGGACCTTGAGACCCCGGGTCTGATGAACTGGGGGG	rs1109037 [CE G]-GRCh38-Chr2:9945582-9945659 rs1109037-G; rs1572313857-G					G3	1	0.0014
rs1294331	AGTGTAGTTATGGATTTTTATTGAATTTTTG	rs1294331 [CE C]-GRCh38-Chr1:233312640-233312670 rs1294331-C	738	G (C)	447	0.6057	G(C)	447	0.6057
rs1294331	AGTATAGTTATGGATTTTTATTGAATTTTTG	rs1294331 [CE T]-GRCh38-Chr1:233312640-233312670 rs1294331-T	738	A (T)	291	0.3943	A(T)	290	0.3930
rs1294331	AGTATAGCTATGGATTTTTATTGAATTTTTG	rs1294331 [CE T]-GRCh38-Chr1:233312640-233312670 rs1294331-T; rs922804818-G					A1(T)	1	0.0014
rs12997453	TTTTATGCTTTAAAGATACAGGTTATCTGTATTACATTGGGTTTTACCTACCTT	rs12997453 [CE G]-GRCh38-Chr2:181548493-181548547 rs12997453-G	738	G	531	0.7195	G	531	0.7195
rs12997453	TTTTATGCTTTAAAGATACAGGTTATCTGTATTACATTAGTTTTTACCTACCTT	rs12997453 [CE A]-GRCh38-Chr2:181548493-181548547 rs12997453-A	738	A	207	0.2805	A	148	0.2005
rs12997453	TTTTATGCTTTAAAGATATAGGTTATCTGTATTACATTAGTTTTTACCTACCTT	rs12997453 [CE A]-GRCh38-Chr2:181548493-181548547 rs72883670-T; rs12997453-A					A1	59	0.0799
rs13182883	TGAGGGGAGGGGTCCCTTCTGGCCTAGTAGAGGGCCTGGCCTGCAGTGAGCATTCAAATCCTCAAGGAACAGGGTGGGGAGGTGGGACAAAGGCAGGAAGAAAGTAACGGAGAGCCTGGGGAGACA	rs13182883 [CE G]-GRCh38-Chr5:137297586-137297711 rs13182883-G	738	G	440	0.5962	G	440	0.5962
rs13182883	TGAGGGGAGGGGTCCCTTCTGGCCTAGTAGAGGGCCTGGCCTGCAGTGAGCATTCAAATCCTCAAGGAACAGGGTGGGGAGGTGGGACAAAGGCAGGAAGAAAGTAACGGAGAGCCTGGGGAGACA	rs13182883 [CE A]-GRCh38-Chr5:137297586-137297711 rs13182883-A	738	A	298	0.4038	A	297	0.4024
rs13182883	TGAGAGGAGGGGTCCCTTCTGGCCTAGTAGAGGGCCTGGCCTGCAGTGAGCATTCAAATCCTCAAGGAACAGGGTGGGGAGGTGGGACAAAGGCAGGAAGAAAGTAACGGAGAGCCTGGGGAGACA	rs13182883 [CE A]-GRCh38-Chr5:137297586-137297711 rs13182883-A; rs1336721677-A					A1	1	0.0014
rs13218440	CCTCTGAGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTTGAGGTACAGCTCCACTGCCCTCTGAGTGGCCCTCCATGAAAATGCCTCATGTCTGTGTCCCTAAACTGTAGGCCTCTGGGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTTGAGGTACAGCTCCACTGCCCTCTGAGTGGCCCTCCATGAAAATGCCTCATGTCTGTGTCCCTAAACTGTAGG	rs13218440 [CE A]-GRCh38-Chr6:12059715-12059838 rs13218440-A	738	A	307	0.4160			
rs13218440	CCTCTGAGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTTGAGGTACAGCTCCACTGCCCTCTGAGTGGCCCTCCATGAAAATGCCTCATGTCTGTGTCCCTAAACTGTAGGCCTCTGGGCAGCCTCCTGGAATACTCAGCTGGGATGGGTTGGGGCTGCTTGAGGTACAGCTCCACTGCCCTCTGAGTGGCCCTCCATGAAAATGCCTCATGTCTGTGTCCCTAAACTGTAGG	rs13218440 [CE G]-GRCh38-Chr6:12059715-12059838 rs13218440-G	738	G	431	0.5840			
rs1335873	AGGGTGCAGGTATGTATTGTTGGCCGTGGTATACTGAGTACATAGCTAGGTACCTGGTACGGGA	rs1335873 [CE A]-GRCh38-Chr13:20327551-20327614 rs1335873-A	738	T(A)	442	0.5989	T(A)	442	0.5989
rs1335873	AGGGTGCAGGTATGTATTGTTGGCCGTGGGAATACTGAGTACATAGCTAGGTACCTGGTACGGGA	rs1335873 [CE T]-GRCh38-Chr13:20327551-20327614 rs1335873-T	738	A(T)	296	0.4011	A(T)	295	0.3997
rs1335873	AGGGTGCAGGTATGTATTGTTGGCCGTGGGAATACTGAGTACATAGCTAGGTACCTGGTACAGGA	rs1335873 [CE T]-GRCh38-Chr13:20327551-20327614 20327554-T; rs1335873-T					A1(T)	1	0.0014
rs1336071	AAAAAGCATCAGATTAACAAAGATAAGAAAAATAGAAATTCAGATGGACAAAACAG	rs1336071 [CE T]-GRCh38-Chr6:93827491-93827549 rs1336071-T	738	A(T)	379	0.5136	A(T)	379	0.5136
rs1336071	AAAAAGCATCAGGTTAAACAAAGATAAGAAAAATAGAAATTCAGATGGACAAAACAG	rs1336071 [CE C]-GRCh38-Chr6:93827491-93827549 rs1336071-C	738	G (C)	359	0.4864	G(C)	332	0.4499
rs1336071	AAAAAGCATCAGGTTAAACAAAGATAAGAAAAATAGAAATTCAGATGGACAAAACAG	rs1336071 [CE C]-GRCh38-Chr6:93827491-93827549 93827520.1T; rs1336071-C					G1(C)	7	0.0095

rs1336071	AAAAAGCATCAGGTTAAACAAAGATAAGAAAAA A TAGAATTTTCAGATGGACAAAACAG	rs1336071 [CE C]-GRCh38-Chr6:93827491-93827549 rs73756355-T; rs1336071-C						G2(C)	18	0.0244
rs1336071	AAAAAGCATCAGGTTAAACAAAGATAAGAAAAA T TAGAATTTTCAGATGGACAAAACA	rs1336071 [CE C]-GRCh38-Chr6:93827491-93827549 rs1336071-C; rs540177118-delCT						G3(C)	2	0.0027
rs1355366	TTCAGAGCCACTGGAGGCCTCGAGGATG GGG ACTTGCCAAAGCCAGTTGTGGCCTAAGCACA TGTGCCAGCTGGAT	rs1355366 [CE C]-GRCh38-Chr3:191088272-191088347 rs1355366-C	738	G(C)	338	0.4580				
rs1355366	TTCAGAGCCACTGGAGGCCTCGAGGATG AGG ACTTGCCAAAGCCAGTTGTGGCCTAAGCACA TGTGCCAGCTGGAT	rs1355366 [CE T]-GRCh38-Chr3:191088272-191088347 rs1355366-T	738	A(T)	400	0.5420				
rs1360288	GGG G AGAGGCTCCCTGCAGCCTCTGGGGAGTAGAGGAGACCTGGGAAGATACTGGCTGCAT CCCCTCAACAGATGCCCC	rs1360288 [CE T]-GRCh38-Chr9:126205735-126205813 rs1360288-T	738	T	240	0.3252	T	240	0.3252	
rs1360288	GGG G AGAGGCTCCCTGCAGCCTCTGGGGAGTAGAGGAGACCTGGGAAGAC ACT GGCTGCAT CCCCTCAACAGATGCCCC	rs1360288 [CE C]-GRCh38-Chr9:126205735-126205813 rs1360288-C	738	C	498	0.6748	C	496	0.6721	
rs1360288	GGG T AGAGGCTCCCTGCAGCCTCTGGGGAGTAGAGGAGACCTGGGAAGAC ACT GGCTGCAT CCCCTCAACAGATGCCCC	rs1360288 [CE C]-GRCh38-Chr9:126205735-126205813 rs116412791-T; rs1360288-C					C1	2	0.0027	
rs1382387	GAAGGAGAAACACCTGAAC TTT CAAT TC CCTGCAGTGGGCAGATGC	rs1382387 [CE A]-GRCh38-Chr16:80072444-80072489 rs1382387-A	738	T(A)	567	0.7683				
rs1382387	GAAGGAGAAACACCTGAAC TTT CAAG T CCCTGCAGTGGGCAGATGC	rs1382387 [CE C]-GRCh38-Chr16:80072444-80072489 rs1382387-C	738	G(C)	171	0.2317				
rs1413212	GGTGGAGCATGGGG G CATTTCA	rs1413212 [CE C]-GRCh38-Chr1:242643488-242643508 rs1413212-C	738	G(C)	507	0.6870				
rs1413212	GGTGGAGCATGGGG A CATTTCA	rs1413212 [CE T]-GRCh38-Chr1:242643488-242643508 rs1413212-T	738	A(T)	231	0.3130				
rs1454361	GGGAGGAGGGAAATACACCCTGAGCTGCTTGT T TTCTAAATGGATACTGAAAAGTGTCTTA CTGATGATGGAC	rs1454361 [CE A]-GRCh38-Chr14:25381580-25381654 rs1454361-A	738	T(A)	324	0.4390				
rs1454361	GGGAGGAGGGAAATACACCCTGAGCTGCATGTTGT TT CTAAATGGATACTGAAAAGTGTCTTA CTGATGATGGAC	rs1454361 [CE T]-GRCh38-Chr14:25381580-25381654 rs1454361-T	738	A(T)	414	0.5610				
rs1463729	CTTTGGCAGC A TACACTCATAGCCAC G TGCAGCCATTCCCACCCAAAAGCA T GCACAT	rs1463729 [CE C]-GRCh38-Chr9:124119138-124119195 rs1463729-C	738	G(C)	346	0.4688	G(C)	345	0.4675	
rs1463729	CTTTGGCAGC G TACACTCATAGCCAC G TGCAGCCATTCCCACCCAAAAGCA T GCACAT	rs1463729 [CE C]-GRCh38-Chr9:124119138-124119195 rs1463729-C; rs553955089-C					G1(C)	1	0.0014	
rs1463729	CTTTGGCAGC A TACACTCATAGCCAC A TGCAGCCATTCCCACCCAAAAGCA T GCACAT	rs1463729 [CE T]-GRCh38-Chr9:124119138-124119195 rs1463729-T	738	A(T)	392	0.5312	A(T)	388	0.5257	
rs1463729	CTTTGGCAGC A TACACTCATAGCCAC A TGCAGCCATTCCCACCCAAAAGCA C GCACAT	rs1463729 [CE T]-GRCh38-Chr9:124119138-124119195 rs114709250-G; rs1463729-T					A1(T)	4	0.0054	
rs1490413	TCAGAACTGCCTGGTGTGG A CTGGGCTGATGTGGGTTCTTTGCAGAACTGGCTGG	rs1490413 [CE G]-GRCh38-Chr1:4307219-4307273 rs1490413-G	738	G	426	0.5772	G	426	0.5772	
rs1490413	TCAGAACTGCCTGGTGTGG A CTGGGCTGATGTGGGTTCTTTGCAGAACTGGCTGG	rs1490413 [CE A]-GRCh38-Chr1:4307219-4307273 rs1490413-A	738	A	312	0.4228	A	311	0.4214	
rs1490413	TCAGAACTGCCTGGTGTGG G CTGGGCTGATGTGGGTTCTTTGCAGAACTGGCTGG	rs1490413 [CE A]-GRCh38-Chr1:4307219-4307273 rs113167966-G; rs1490413-A					A1	1	0.0014	
rs1498553	ACTTCAGATGTTCAAAGCCAGAC G GAGAATAAAAGGATGGCTATGAGATCTATGGAAGTGCTGA G	rs1498553 [CE C]-GRCh38-Chr11:5687776-5687839 rs1498553-C	738	C	314	0.4255				
rs1498553	ACTTCAGATGTTCAAAGCCAGATGAGAATAAAAGGATGGCTATGAGATCTATGGAAGTGCTGA G	rs1498553 [CE T]-GRCh38-Chr11:5687776-5687839 rs1498553-T	738	T	424	0.5745				
rs1523537	TCTTAATACATTCATTTCTGC A TGGGTGGGGTTTCAGTCTGCAACAAGATCTTGTAGGGACGCT ATCGCT	rs1523537 [CE C]-GRCh38-Chr20:52679563-52679632 rs1523537-C	738	C	342	0.4634	C	342	0.4634	
rs1523537	TCTTAATACATTCATTTCTGC A TGGGTGGGGTTTCAGTCTGCAACAAGATCTTGTAGGGATGCT ATCGCT	rs1523537 [CE T]-GRCh38-Chr20:52679563-52679632 rs1523537-T	738	T	396	0.5366	T	385	0.5217	
rs1523537	TCTTAATACATTCATTTCTGC G TGGGTGGGGTTTCAGTCTGCAACAAGATCTTGTAGGGATGCT ATCGCT	rs1523537 [CE T]-GRCh38-Chr20:52679563-52679632 rs77195753-G; rs1523537-T					T1	11	0.0149	
rs1528460	CTTAACATATTTAAG A CTGAAAAATGTTACAGTAAAAGTTTTGTTTAACTCCTGCATTTGCCAAAC	rs1528460 [CE C]-GRCh38-Chr15:54918491-54918555 rs1528460-C	738	C	294	0.3984				

rs1528460	CTTAACATATTTAAGATTGAAAAATGTTACAGTAAAAGTTTTGTTTAACTCTGCATTTGCCAAAC	rs1528460 [CE T]-GRCh38-Chr15:54918491-54918555 rs1528460-T	738	T	444	0.6016				
rs159606	GTTTCTCATCTGTTATTATTTGTTTACATCTGCTCCTATATTTATTCTCTC	rs159606 [CE A]-GRCh38-Chr5:17374761-17374814 rs159606-A	738	A	250	0.3388				
rs159606	GTTTCTCATCTGTTATTATTTGTTTACGTCTGCTCCTATATTTATTCTCTC	rs159606 [CE G]-GRCh38-Chr5:17374761-17374814 rs159606-G	738	G	488	0.6612				
rs1821380	ATGGAGCCACTGAACTGCAGTGCAAAAATGCAGTAAGGCATACAGATAGAAGAAGGAGAAATGT CAGGAAAAGACAG	rs1821380 [CE G]-GRCh38-Chr15:39021164-39021239 rs1821380-G	738	C(G)	245	0.3320	C(G)	245	0.3320	
rs1821380	ATGGAGCCACTGAACTGCAGTGCAAAAATGCAGTAAGGGATACAGATAGAAGAAGGAGAAATGT CAGGAAAAGACAG	rs1821380 [CE C]-GRCh38-Chr15:39021164-39021239 rs1821380-C	738	G(C)	493	0.6680	G(C)	491	0.6653	
rs1821380	ATGGAGCCACTGAACTGCAGTGCAAAAATGCAGTAAGGGATACAGATAGAAGAAGGAGAAACG TCAGGAAAAGACAG	rs1821380 [CE C]-GRCh38-Chr15:39021164-39021239 rs76591310-G; rs1821380-C					G1(C)	2	0.0027	
rs1886510	GGGTAAATTTTAGTAATTCTTAAAAGAATAAGCAATATTGCAAGTGTTGTTGTGAAAATCCAGG CGTCA	rs1886510 [CE A]-GRCh38-Chr13:21800531-21800600 rs1886510-A	738	T(A)	251	0.3401				
rs1886510	GGGTAAATTTTAGTAATTCTTAAAAGAATAAGCAATATTGCGAAGTGTTGTTGTGAAAATCCAG GCGTCA	rs1886510 [CE G]-GRCh38-Chr13:21800531-21800600 rs1886510-G	738	C(G)	487	0.6599				
rs1979255	GATGAGCAAGAGTTCCAACGTTCCATGCCCTGACCAACACAAGCTA	rs1979255 [CE C]-GRCh38-Chr4:189396884-189396929 rs1979255-C	738	G(C)	237	0.3211	G(C)	237	0.3211	
rs1979255	GATCAGCAAGAGTTCCAACGTTCCATGCCCTGACCAACACAAGCTA	rs1979255 [CE G]-GRCh38-Chr4:189396884-189396929 rs1979255-G	738	C(G)	501	0.6789	C(G)	500	0.6775	
rs1979255	GATCAGCAAGAGTTCCAACGTTCCATGCCCTGATCAACACAAGCTA	rs1979255 [CE G]-GRCh38-Chr4:189396884-189396929 rs1979255-G; 189396896-G>A					C1(G)	1	0.0014	
rs2040411	AAGTGCAATTTTCATGA	rs2040411 [CE A]-GRCh38-Chr22:47440656-47440672 rs2040411-A	738	A	437	0.5921				
rs2040411	AAGTGCGTATTTTCATGA	rs2040411 [CE G]-GRCh38-Chr22:47440656-47440672 rs2040411-G	738	G	301	0.4079				
rs2046361	TGAAACGATCATTTTCAAATTTAAATATTAGAAAGGTGAAGTGTAACAATGACCAAAAATAGAC AAATC	rs2046361 [CE A]-GRCh38-Chr4:10967394-10967463 rs2046361-A	738	T(A)	271	0.3672	T(A)	271	0.3672	
rs2046361	TGAAACGATCATTTTCAAATTTAAATTAATAGAAAGGTGAAGTGTAACAATGACCAAAAATAGAC CAAATC	rs2046361 [CE T]-GRCh38-Chr4:10967394-10967463 rs2046361-T	738	A(T)	467	0.6328	A(T)	465	0.6301	
rs2046361	TGAAACGATCATTTTCAAATTTAAATTAATAGAAAGGTGAAGTGTAACAATGACCAAAAATAGAC CAAATC	rs2046361 [CE T]-GRCh38-Chr4:10967394-10967463 rs2046361-T; rs116699455-T					A1(T)	2	0.0027	
rs2056277	AAGCATGTGTCAACCGCCAAACTGGGTGTTAGGGAGACAGGCATGAATGAGATGGGA	rs2056277 [CE T]-GRCh38-Chr8:138386821-138386877 rs2056277-T	738	T	127	0.1721	T	127	0.1721	
rs2056277	AAGCATGTGTCAACCGCCAAACTGGGTGTTAGGGAGACAGGCATGAATGAGACGGGA	rs2056277 [CE C]-GRCh38-Chr8:138386821-138386877 rs2056277-C	738	C	611	0.8279	C	589	0.7981	
rs2056277	AAGCATGTGTCAACTGCCAAACTGGGTGTTAGGGAGACAGGCATGAATGAGACGGGA	rs2056277 [CE C]-GRCh38-Chr8:138386821-138386877 rs149318691-T; rs2056277-C					C1	22	0.0298	
rs2076848	GAAATTATTGATAATACACAGGTATCCTGGCCTCACCACCAGAAATCAGGGCTTGATGGACCT GAAGCGGTCCCAGG	rs2076848 [CE A]-GRCh38-Chr11:134797628-134797704 rs2076848-A	736	T(A)	405	0.5503	T(A)	383	0.5204	
rs2076848	GAAATTATTGATAATACACAGGTATCCTGGCCTCACCACCAGAAATCAGGGCTTGATGGACCT GAAGCGGTCCCAGG	rs2076848 [CE A]-GRCh38-Chr11:134797628-134797704 rs7947725-T; rs2076848-A					T1(A)	22	0.0299	
rs2076848	GAAATTATTGATAATACACAGGTATCCTGGCCTCACCACCAGAAATCAGGGCATGATGGACCT GAAGCGGTCCCAGG	rs2076848 [CE T]-GRCh38-Chr11:134797628-134797704 rs2076848-T	736	A(T)	331	0.4497	A(T)	329	0.4470	
rs2076848	GAAATTATTGATAATACACAGGTATCCTGGCCTCACCACCAGAAATCAGGGCATGATGGACCT GAAGCGGTCCCAGG	rs2076848 [CE T]-GRCh38-Chr11:134797628-134797704 rs7947725-T; rs2076848-T					A1(T)	2	0.0027	
rs2107612	ACTAATTATGTGTTTTTCTAAATCATATTGTCTACTTTTCTCCAAAACA	rs2107612 [CE A]-GRCh38-Chr12:779120-779169 rs2107612-A	738	A	486	0.6585				
rs2107612	ACTAATTATGTGTTTTTCTAAATCATATTGTCTGCTTTTCTCCAAAACA	rs2107612 [CE G]-GRCh38-Chr12:779120-779169 rs2107612-G	738	G	252	0.3415				
rs2111980	CCTTCGAGCTCCAGCCTGGTGCCTCCGCTCCGTGACTACTGGCAAAGATCT	rs2111980 [CE C]-GRCh38-Chr12:105934430-105934481 rs2111980-C	738	G(C)	360	0.4878	G(C)	360	0.4878	
rs2111980	CCTTCAAGCTCCAGCCTGGTGCCTCCGCTCCGTGACTACTGGCAAAGATCT	rs2111980 [CE T]-GRCh38-Chr12:105934430-105934481 rs2111980-T	738	A(T)	378	0.5122	A(T)	372	0.5041	

rs2111980	CCTTCAAGCTCCAGCCTGGTGCCTCCGCTCCATGACTACTGGCAAAGATCT	rs2111980 [CE T]-GRCh38-Chr12:105934430-105934481 rs79578959-T; rs2111980-T						A1(T)	6	0.0081
rs221956	TTCCCTCCAGCTCTCCTCTCCCTTTCTGAGCCCTCAGCAAACCTGACTTTAG	rs221956 [CE C]-GRCh38-Chr21:42186845-42186896 rs221956-C	738	C	462	0.6260				
rs221956	TTCCCTCCAGCTCTCCTCTCCCTTTCTGAGCCCTCAGCAAATTGACTTTAG	rs221956 [CE T]-GRCh38-Chr21:42186845-42186896 rs221956-T	738	T	276	0.3740				
rs214955	CCTGCCATTAAATTTTTGCGTCACCTTTTCAGTCTTTGTTTGCAAGCATCCAAGGGCTGCCAGA ATAAAGAA	rs214955 [CE C]-GRCh38-Chr6:152376519-152376590 rs214955-C	738	G(C)	327	0.4431	G(C)	327	0.4431	
rs214955	CCTGCCATTAAATTTTTGCATCACCTTTTCAGTCTTTGTTTGCAAGCATCCAAGGGCTGCCAGA ATAAAGAA	rs214955 [CE T]-GRCh38-Chr6:152376519-152376590 rs214955-T	738	A(T)	411	0.5569	A(T)	397	0.5379	
rs214955	CCTGCCATTAAATTTTTGCATCACCTTTTCAGTCTTTGTTTGCAAGCATCAAAGGGCTGCCAGA ATAAAGAA	rs214955 [CE T]-GRCh38-Chr6:152376519-152376590 rs117020413-T; rs214955-T					A1(T)	14	0.0190	
rs2269355	TCCCAGATTCTCTCCACAGTCCC	rs2269355 [CE C]-GRCh38-Chr12:6836737-6836759 rs2269355-C	738	C	302	0.4092				
rs2269355	TCCCAGATTCTCTGCACAGTCCC	rs2269355 [CE G]-GRCh38-Chr12:6836737-6836759 rs2269355-G	738	G	436	0.5908				
rs2342747	GGAGGAAGAAAACAGAGAGTCTTGACCGTAGAGGGGACAACAAAGAATGAGCTT	rs2342747 [CE A]-GRCh38-Chr16:5818670-5818723 rs2342747-A	738	A	226	0.3062				
rs2342747	GGAGGAAGAAAACAGAGAGTCTTGACCGTGAGGGGACAACAAAGAATGAGCTT	rs2342747 [CE G]-GRCh38-Chr16:5818670-5818723 rs2342747-G	738	G	512	0.6938				
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE G]-GRCh38-Chr3:110582177-110582286 rs2399332-G	738	C(G)	481	0.6518	C(G)	12	0.0163	
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE G]-GRCh38-Chr3:110582177-110582286 rs2399332-T; rs2399333-T; rs2399332-G					C1(G)	468	0.6341	
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE G]-GRCh38-Chr3:110582177-110582286 rs2399333-T; rs2399332-G					C2(G)	1	0.0014	
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE T]-GRCh38-Chr3:110582177-110582286 rs2399332-T	738	A(T)	257	0.3482	A(T)	251	0.3401	
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE T]-GRCh38-Chr3:110582177-110582286 rs2399333-T; rs2399332-T					A1(T)	3	0.0041	
rs2399332	ATTAAAAATCAGCAAAATACATGAACAGATACTTCTCAAAGAAGCCAACAAACTTGAAAAAT GTTCAACGCTACTAATCATGAGAGAAATGTAATCAAAACCCAAT	rs2399332 [CE T]-GRCh38-Chr3:110582177-110582286 rs2399334-T; rs2399332-T					A2(T)	3	0.0041	
rs251934	GAGGCTTTTAAGTAGATGGGACAGCCAGATATCTACTACTTCATCTGCCCCAGAC	rs251934 [CE A]-GRCh38-Chr5:175351637-175351691 rs251934-A	738	T(A)	461	0.6247				
rs251934	GAGGCTTTTAAGTAGACGGGACAGCCAGATATCTACTACTTCATCTGCCCCAGAC	rs251934 [CE G]-GRCh38-Chr5:175351637-175351691 rs251934-G	738	C(G)	277	0.3753				
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE A]-GRCh38-Chr4:46327592-46327708 rs279844-A	738	A	476	0.6450	A	473	0.6409	
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE A]-GRCh38-Chr4:46327592-46327708 rs279844-A; 46327708-T>C					A1	3	0.0041	
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE T]-GRCh38-Chr4:46327592-46327708 rs279844-T	738	T	262	0.3550	T	1	0.0014	
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE T]-GRCh38-Chr4:46327592-46327708 rs279844-T; rs279845-A					T1	248	0.3360	
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE T]-GRCh38-Chr4:46327592-46327708 rs279844-T; rs61621790-A					T2	12	0.0163	
rs279844	AAGGATAACAGATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT TCCTTGATATTAAGTTCAGTGTCATTTTGACCAGATATTAATACTCACAACCTCTCTAACT	rs279844 [CE T]-GRCh38-Chr4:46327592-46327708 rs279844-T; rs279845-A; 46327691-C>T					T3	1	0.0014	
rs2830795	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTTATTGTCTAAAGAGCAAAGAAGTCCTA TTAT	rs2830795 [CE A]-GRCh38-Chr21:27235795-27235862 rs2830795-A	738	A	553	0.7493	A	474	0.6423	
rs2830795	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTTATTGTCTAAAGAGCAAAGAAGTCCTA TTAT	rs2830795 [CE A]-GRCh38-Chr21:27235795-27235862 rs12626695-C; rs2830795-A					A1	75	0.1016	
rs2830795	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTTATTGTCTAATGAGCAAAGAAGTCCTA TTAT	rs2830795 [CE A]-GRCh38-Chr21:27235795-27235862 rs79319609-T; rs2830795-A					A2	4	0.0054	

rs2830795	ACTGGGTTCACTTCTATAGACATAGGACACACCATTTTATTGTCTAAAGGGCAAAGAAGTCCTA TTAT	rs2830795 [CE G]-GRCh38-Chr21:27235795-27235862 rs2830795-G	738	G	185	0.2507	G	169	0.2290
rs2830795	ACTGGGTTCACTCTCTATAGACATAGGACACACCATTTTATTGTCTAAAGGGCAAAGAAGTCCTA TTAT	rs2830795 [CE G]-GRCh38-Chr21:27235795-27235862 rs12626695-C; rs2830795-G					G1	16	0.0217
rs2831700	ATTTGGCTAAACTATTGCCGGAGATAAGTTAGAA	rs2831700 [CE A]-GRCh38-Chr21:28307343-28307376 rs2831700-A	738	A	432	0.5854			
rs2831700	ATTTGGCTAAACTATTGCCGGAGATGAGTTAGAA	rs2831700 [CE G]-GRCh38-Chr21:28307343-28307376 rs2831700-G	738	G	306	0.4146			
rs321198	ATACAATTCTCAAATGAAATAACTAAATAAGGAAGCTGTGTTCCCTTTCTCCTACACACAGGCTTC AGGTTACCTGTTTTCCCTTTTGATTCCACTTCCGTGTGAAGCAAGCAGT	rs321198 [CE C]-GRCh38-Chr7:137344994-137345108 rs321198-C	738	C	427	0.5786	C	427	0.5786
rs321198	ATACAATTCTCAAATGAAATAACTAAATAAGGAAGCTGTGTTCCCTTTCTCCTACACACAGGCTTC AGGTTACCTGTTTTCCCTTTTGATTCCACTTCTGTGTGAAGCAAGCAGT	rs321198 [CE T]-GRCh38-Chr7:137344994-137345108 rs321198-T	738	T	311	0.4214	T	308	0.4173
rs321198	ATACAATTCTCAAATGAAATAACTAAATAAGGAAGCTGTGTTCCCTTTCTCCTACACACAGGCTTC AGGTTACCTGTTTTCCCTTTTGATTCCACTTCTGCGTGAAGCAAGCAGT	rs321198 [CE T]-GRCh38-Chr7:137344994-137345108 rs321198-T; rs186446125-C					T1	3	0.0041
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE A]-GRCh38-Chr13:106285996-106286115 rs354439-A	738	T(A)	295	0.3997	T(A)	293	0.3970
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE A]-GRCh38-Chr13:106285996-106286115 rs354439-A; 106286115-106286116 INS [CC]					T1(A)	1	0.0014
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE A]-GRCh38-Chr13:106285996-106286115 rs354439-A; 106285999-T>C					T2(A)	1	0.0014
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE T]-GRCh38-Chr13:106285996-106286115 rs354439-T	738	A(T)	443	0.6003	A(T)	441	0.5976
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE T]-GRCh38-Chr13:106285996-106286115 rs354439-T; rs894755414-A					A1(T)	1	0.0014
rs354439	TGTTCTGGTGGCTTCTCTTTCCCTTATGTATCTCTCATGTATCACATTCCCTTTTAAGCACAAAT ATTCTGAATATCATTCACTGTTTTCTATCGCAACCTGCAATTTGAGAGTTAAGAA	rs354439 [CE T]-GRCh38-Chr13:106285996-106286115 rs354439-T; rs574448517-C					A2(T)	1	0.0014
rs3780962	CTGTCCTTACGGGTGAAAGCTGATATCTTGACCTTGTTTCATC	rs3780962 [CE A]-GRCh38-Chr10:17151313-17151354 rs3780962-A	738	T(A)	330	0.4472			
rs3780962	CTGTCCTCACGGGTGAAAGCTGATATCTTGACCTTGTTTCATC	rs3780962 [CE G]-GRCh38-Chr10:17151313-17151354 rs3780962-G	738	C(G)	408	0.5528			
rs430046	AAGGTCATACAATGAATGGTGTGATGTAAACGCTTGGGAGGCGATTCTGAGGGTAGGTGCTG GGTT	rs430046 [CE C]-GRCh38-Chr16:77983107-77983174 rs430046-C	738	C	460	0.6233	C	460	0.6233
rs430046	AAGGTCATACAATGAATGGTGTGATGTAAACGCTTGGGAGGTGATTCTGAGGGTAGGTGCTG GGTT	rs430046 [CE T]-GRCh38-Chr16:77983107-77983174 rs409820-A; rs430044-T; rs430046-T	738	T	278	0.3767	T	277	0.3753
rs430046	AAGGTCATACAATGAATGGTGTGATGTAAACGCTTGGGAGGTGATTCTGAGGGTAGGTGCTG GGTGT	rs430046 [CE T]-GRCh38-Chr16:77983107-77983174 rs409820-A; rs430044-T; rs430046-T; rs190617353-G					T1	1	0.0014
rs4364205	TTCCAGTCTAGATATCCACCCATGAGAAATATATCCACAAAAAGATTGAGCAA	rs4364205 [CE G]-GRCh38-Chr3:32376108-32376161 rs4364205-G	738	G	441	0.5976			
rs4364205	TTCCAGTCTAGATATCCACCCATGAGAAATATATCCACAAAAATATTGAGCAA	rs4364205 [CE T]-GRCh38-Chr3:32376108-32376161 rs4364205-T	738	T	297	0.4024			
rs445251	CATGTGCATTGGAGTTTTGATCACGAACCACTTGCAAGTTTACATTAAATTTGAATTTGAGGCC GGGT	rs445251 [CE C]-GRCh38-Chr20:15144244-15144311 rs369438-C; rs445251-C	738	G(C)	445	0.6030	G(C)	445	0.6030
rs445251	CATGTGCATTGGAGTTTTGATCACGAACCACTTGCAAGTTTACATTAAATTTGAATTTGAGGCC AGGT	rs445251 [CE G]-GRCh38-Chr20:15144244-15144311 rs445251-G	738	C(G)	293	0.3970	C(G)	288	0.3902
rs445251	TATGTGCATTGGAGTTTTGATCACGAACCACTTGCAAGTTTACATTAAATTTGAATTTGAGGCCA GGT	rs445251 [CE G]-GRCh38-Chr20:15144244-15144311 rs445251-G; rs117702247-A					C1(G)	5	0.0068
rs4606077	TGGGATCTGACTCCCCACAGCCACCCAAAGCCAGGGAACCTCTCACTGCCCTTCGGGCTTC AGCACAGGGCTGTCTCCACGCGGCAGGGCCTGTGCTTTCACCTGG	rs4606077 [CE C]-GRCh38-Chr8:143574562-143574669 rs4606077-C; rs1869434-A	738	C	484	0.6558	C	484	0.6558
rs4606077	TGGGATCTGACTCCCCACAGCCTACCCAAAGCCGGGGAACCTCTCACTGCCCTTCGGGCTTC AGCACAGGGCTGTCTCCACGCGGCAGGGCCTGTGCTTTCACCTGG	rs4606077 [CE T]-GRCh38-Chr8:143574562-143574669 rs4606077-T	738	T	254	0.3442	T	180	0.2439
rs4606077	TGGGATCTGACTCCCCACAGCCTACCCAAAGCTGGGGAACCTCTCACTGCCCTTCGGGCTTC AGCACAGGGCTGTCTCCACGCGGCAGGGCCTGTGCTTTCACCTGG	rs4606077 [CE T]-GRCh38-Chr8:143574562-143574669 rs4606077-T; rs58774517-T					T1	73	0.0989
rs4606077	TGGGATCTGACTCCCCACAGCCTACCCAAAGCCGGGGAACCTCTCACTGCCCTTCGGGCTTC AGCACAGGGCTGTCTCCACGCGGCAGGGCCTGTGCTTTCACCTGG	rs4606077 [CE T]-GRCh38-Chr8:143574562-143574669 rs4606077-T; rs975955864-C					T2	1	0.0014

rs560681	TCCATCTCTGTTTACTCAGGTCACAGG A CCCTGGGGCCTCCAAGAGTT	rs560681 [CE G]-GRCh38-Chr1:160816871-160816918 rs560681-G	738	G	303	0.4106	G	303	0.4106
rs560681	TCCATCTCT A TTTACTCAGGTCACAGG A CCCTGGGGCCTCCAAGAGTT	rs560681 [CE A]-GRCh38-Chr1:160816871-160816918 rs560681-A	738	A	435	0.5894	A	423	0.5732
rs560681	TCCATCTCT A TTTACTCAGGTCACAGG G CCCTGGGGCCTCCAAGAGTT	rs560681 [CE A]-GRCh38-Chr1:160816871-160816918 rs560681-A; rs60615385-G					A1	12	0.0163
rs576261	GTCACCAACCCTGGCCTCACAACCTCTCTC	rs576261 [CE A]-GRCh38-Chr19:39069160-39069188 rs576261-A	738	A	429	0.5813			
rs576261	GTCACCACCCCTGGCCTCACAACCTCTCTC	rs576261 [CE C]-GRCh38-Chr19:39069160-39069188 rs576261-C	738	C	309	0.4187			
rs6444724	TTGGAATGGGAGGAAAGGGAAGGACTAAATTGTTGAACACTGGTTACCGTCTAGGTATTTA CAAACCTTG	rs6444724 [CE C]-GRCh38-Chr3:193489543-193489613 rs6444724-C	738	C	299	0.4051			
rs6444724	TTGGAATGGGAGGAAAGGGAAGGACTAAATTGTTGAACACTGGTTACTGTGCTAGGTATTTA CAAACCTTG	rs6444724 [CE T]-GRCh38-Chr3:193489543-193489613 rs6444724-T	738	T	439	0.5949			
rs6811238	CACTGAGAGGAGAAGACTGTGTGTTTTAAAGCCAGGTTTGTTTAAAGGGTTATGATAGTATTAA	rs6811238 [CE G]-GRCh38-Chr4:168742417-168742480 rs6811238-G	738	G	375	0.5081			
rs6811238	CACTGAGAGGAGAAGACTGTGTGTTTTAAAGCCAGGTTTGTTTAAAGTGTATGATAGTATTAA	rs6811238 [CE T]-GRCh38-Chr4:168742417-168742480 rs6811238-T	738	T	363	0.4919			
rs717302	AATAAGCTTTAGAAAGGCATATCGTATTA A CTGTGT A GTGAACGCTGTGCATTAGGTTTAGCTC	rs717302 [CE A]-GRCh38-Chr5:2879245-2879308 rs717302-A	738	A	396	0.5366			
rs717302	AATAAGCTTTAGAAAGGCATATCGTATTA A CTGTGT G GTGAACGCTGTGCATTAGGTTTAGCTC	rs717302 [CE G]-GRCh38-Chr5:2879245-2879308 rs717302-G	738	G	342	0.4634			
rs719366	GTCACTTCTCGGCAGCATT C AGCACTGTGACCACAGCATCTTTAACTCTTTTATTATCCTTTCC TGCTTTTCCTCCTCCCACTTAG C AGCTACTCCTCTGGGGGCCTGTCCTTTACTC	rs719366 [CE A]-GRCh38-Chr19:27972398-27972517 rs719366-A	738	T(A)	540	0.7317			
rs719366	GTCACTTCTCGGCAGCATT C AGCACTGTGACCACAGCATCTTTAACTCTTTTATTATCCTTTCC TGCTTTTCCTCCTCCCACTTAG C AGCTACTCCTCTGGGGGCCTGTCCTTTACTC	rs719366 [CE G]-GRCh38-Chr19:27972398-27972517 rs719366-G	738	C(G)	198	0.2683			
rs722098	GAAATATCCTTG A TAAAGGATTTAAATTTTGATGTGCTGAATATTTCTT	rs722098 [CE A]-GRCh38-Chr21:15313267-15313315 rs722098-A	738	A	513	0.6951			
rs722098	GAAATATCCTTG G TAAAGGATTTAAATTTTGATGTGCTGAATATTTCTT	rs722098 [CE G]-GRCh38-Chr21:15313267-15313315 rs722098-G	738	G	225	0.3049			
rs722290	GTGTTTCAGATTT C AGATTTTGAAATATTTGCATATACATACT G AGATGTCTTG	rs722290 [CE C]-GRCh38-Chr14:52749993-52750048 rs722290-C	738	G(C)	340	0.4607			
rs722290	GTGTTTCAGATTT C AGATTTTGAAATATTTGCATATACATACT C AGATGTCTTG	rs722290 [CE G]-GRCh38-Chr14:52749993-52750048 rs722290-G	738	C(G)	398	0.5393			
rs727811	CAACGACTT A AAATCATCTGCATCTCCAGCAATCTCATGATTTCACTT A TATTTTGCT	rs727811 [CE T]-GRCh38-Chr6:164624247-164624311 rs1390470-T; rs727811-T	738	A(T)	346	0.4688	A(T)	346	0.4688
rs727811	CAACGACTT A CAATCATCTGCATCTCCAGCAATCTCATGATTTCACTT G TATTTTGCT	rs727811 [CE G]-GRCh38-Chr6:164624247-164624311 rs727811-G	738	C(G)	392	0.5312	C(G)	3	0.0041
rs727811	CAACGACTT A CAATCATCTGCATCTCCAGCAATCTCATGATTTCACTT A TATTTTGCT	rs727811 [CE G]-GRCh38-Chr6:164624247-164624311 rs1390470-T; rs727811-G					C1(G)	389	0.5271
rs729172	AGCCTCATTAAATATGACCAAGGCTCCTCTGCAGAC A GAATGTATGTAACCG	rs729172 [CE T]-GRCh38-Chr16:5556181-5556221 rs729172-T	738	A(T)	256	0.3469	A(T)	256	0.3469
rs729172	AGCCTCATTAAATATGACCAAGGCTCCTCTGCAGAC C GAATGTATGTAACCG	rs729172 [CE G]-GRCh38-Chr16:5556181-5556221 rs729172-G	738	C(G)	482	0.6531	C(G)	481	0.6518
rs729172	AGCCTCATTAAATATGACCAAGGCTCCTCTGCAGAC C AATGTATGTAACCG	rs729172 [CE G]-GRCh38-Chr16:5556181-5556221 rs729172-G; rs562809565-T					C1(G)	1	0.0014
rs733164	CCAGGCTCAGCTTT C AGCCCCAGGTCCCACCAACAGGCCATCCCACTTGAAAA A TTTGCTG ACATTCTGAGCCGGGGCC	rs733164 [CE A]-GRCh38-Chr22:27420770-27420849 rs733164-A	738	A	248	0.3360			
rs733164	CCAGGCTCAGCTTT C AGCCCCAGGTCCCACCAACAGGCCATCCCACTTGAAAA G TTTGCTG ACATTCTGAGCCGGGGCC	rs733164 [CE G]-GRCh38-Chr22:27420770-27420849 rs733164-G	738	G	490	0.6640			
rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGCGCCG G CTGG A GAGGAGACCTGCT G GCTTCCCGTTGAATTCGGTG AC A CTC	rs735155 [CE C]-GRCh38-Chr10:3331961-3332088 rs735155-C	738	G(C)	320	0.4336	G(C)	313	0.4241

rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGGCCGCG G CTGG A GAGGGAGACCTGCTGGCTTCCCGTTGAATTCGGTG ACACAC	rs735155 [CE C]-GRCh38-Chr10:3331961-3332088 rs7905965-T; rs735155-C					G1(C)	3	0.0041
rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGGCCGCG G CTGG A GAGGGAGACCTGCTGGCTTCCCGTTGAATTCGGTG ACGCTC	rs735155 [CE C]-GRCh38-Chr10:3331961-3332088 rs373487413-C; rs735155-C					G2(C)	3	0.0041
rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGGCCGCG A CTGG A GAGGGAGACCTGCTGGCTTCCCGTTGAATTCGGTG ACACTC	rs735155 [CE C]-GRCh38-Chr10:3331961-3332088 rs735155-C; rs79799511-T					G3(C)	1	0.0014
rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGGCCGCG G CTGG A GAGGGAGACCTGCTGACTTCCCGTTGAATTCGGTG ACACTC	rs735155 [CE T]-GRCh38-Chr10:3331961-3332088 rs735155-T	738	A(T)	418	0.5664	A(T)	416	0.5637
rs735155	GACGCCGGCTCCAGAAGGGACCTAACCTGGAGAAAACCGGAGAGCTGGCGCTGAAGGGTCG GGAGAGGCCTGCTGGGCCGCG G CTGGAGGGAGACCTGCTGACTTCCCGTTGAATTCGGTGAC ACTC	rs735155 [CE T]-GRCh38-Chr10:3331961-3332088 rs735155-T; 3331999-3332000 DEL [CT]					A1(T)	2	0.0027
rs737681	CCG C GGGAGCTCTGCACAAGCCGCAGGGTCACATGTGAGGCCATCTCCACC C TATCCCTGG GCCCAGGTGCAGTCTCTC	rs737681 [CE C]-GRCh38-Chr7:156198068-156198147 rs737681-C	738	C	429	0.5813	C	429	0.5813
rs737681	CCG C GGGAGCTCTGCACAAGCCGCAGGGTCACATGTGAGGCCATCTCCACC T TATCCCTGG GCCCAGGTGCAGTCTCTC	rs737681 [CE T]-GRCh38-Chr7:156198068-156198147 rs737681-T	738	T	309	0.4187	T	308	0.4173
rs737681	CCGGGGGAGCTCTGCACAAGCCGCAGGGTCACATGTGAGGCCATCTCCACC T TATCCCTGG GCCCAGGTGCAGTCTCTC	rs737681 [CE T]-GRCh38-Chr7:156198068-156198147 rs737681-T; 156198071-C>G					T1	1	0.0014
rs740598	TCAAATAGCAATGGCTCGTCTATGGTTAGTCTC G CAGCCACATT C TCAGAACTGCTCAAACCT GGCCTGC	rs740598 [CE G]-GRCh38-Chr10:116747355-116747426 rs740598-G	738	G	338	0.4580	G	338	0.4580
rs740598	TCAAATAGCAATGGCTCGTCTATGGTTAGTCTC A CAGCCACATT C TCAGAACTGCTCAAACCT GGCCTGC	rs740598 [CE A]-GRCh38-Chr10:116747355-116747426 rs740598-A	738	A	400	0.5420	A	398	0.5393
rs740598	TCAAATAGCAATGGCTCGTCTATGGTTAGTCTC A CAGCCACATT T TCAGAACTGCTCAAACCT GGCCTGC	rs740598 [CE A]-GRCh38-Chr10:116747355-116747426 rs740598-A; rs189367495-T					A1	2	0.0027
rs740910	AAGT A TAAACAGTT T GCTAAGTAAGGTGAGTGGTATAATCATAT G TTTGTA ¹ AAAAAGCAAACAAA	rs740910 [CE G]-GRCh38-Chr17:5803260-5803323 rs740910-G	738	G	189	0.2561	G	189	0.2561
rs740910	AAGT A TAAACAGTT T GCTAAGTAAGGTGAGTGGTATAATCATAT A TTTGTA ¹ AAAAAGCAAACAAA	rs740910 [CE A]-GRCh38-Chr17:5803260-5803323 rs740910-A	738	A	549	0.7439	A	461	0.6247
rs740910	AAGT G TAAACAGTT T GCTAAGTAAGGTGAGTGGTATAATCATAT A TTTGTA ¹ AAAAAGCAAACAAA	rs740910 [CE A]-GRCh38-Chr17:5803260-5803323 rs60810599-G; rs740910-A					A1	87	0.1179
rs740910	AAGT A TAAACAGTT C GGTAAGTAAGGTGAGTGGTATAATCATAT A TTTGTA ¹ AAAAAGCAAACAAA	rs740910 [CE A]-GRCh38-Chr17:5803260-5803323 rs740910-A; 5803273-G>A					A2	1	0.0014
rs763869	TGTTTATATTATTCTAACTCAATTGCATTACATT	rs763869 [CE A]-GRCh38-Chr8:1427433-1427467 rs763869-A	738	T(A)	428	0.5799			
rs763869	TGTTTATATTATTCTAACTCAACTGCATTACATT	rs763869 [CE G]-GRCh38-Chr8:1427433-1427467 rs763869-G	738	C(G)	310	0.4201			
rs8037429	GAGTTACGTAG	rs8037429 [CE C]-GRCh38-Chr15:53324706-53324716 rs8037429-C	738	C	379	0.5136			
rs8037429	GAGTTATGTAG	rs8037429 [CE T]-GRCh38-Chr15:53324706-53324716 rs8037429-T	738	T	359	0.4864			
rs8078417	GGG G ACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTC G TGCCCCCGTGAAGCTTCAGCTC CCCTCCCGGCTGTCTTGAGGCTCTTCTCACACTCAGATGC	rs8078417 [CE C]-GRCh38-Chr17:82503992-82504093 rs8078417-C	738	C	528	0.7154	C	525	0.7114
rs8078417	GGG G ACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTC A TGCCCCCGTGAAGCTTCAGCTCC CCTCCCGGCTGTCTTGAGGCTCTTCTCACACTCAGATGC	rs8078417 [CE C]-GRCh38-Chr17:82503992-82504093 rs138630479-A; rs8078417-C	738				C1	2	0.0027
rs8078417	GGG G ACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTC G TGCCCCCGTGAAGCTTCAGCTCC CCTCCCGGCTGTCTTGAGGCTCTTCTCACACTCAGATGC	rs8078417 [CE C]-GRCh38-Chr17:82503992-82504093 rs78650971-C; rs8078417-C					C2	1	0.0014
rs8078417	GGG G ACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTC G TGCCCCCGTGAAGCTTCAGCTC CCCTCCCTGGCTGTCTTGAGGCTCTTCTCACACTCAGATGC	rs8078417 [CE T]-GRCh38-Chr17:82503992-82504093 rs8078417-T	T	210	0.2846	T	208	0.2818	
rs8078417	GGG C ACGCCTGGCGCTGCGAGGGAGGCCCGAGCCTC G TGCCCCCGTGAAGCTTCAGCTCC CCTCCCTGGCTGTCTTGAGGCTCTTCTCACACTCAGATGC	rs8078417 [CE T]-GRCh38-Chr17:82503992-82504093 rs78650971-C; rs8078417-T					T1	2	0.0027
rs826472	TAGCATTTTATGCATCAATAAGATGTTAATT A TGATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAG TCACTAT CTCTCATCAGGAT	rs826472 [CE T]-GRCh38-Chr10:2364342-2364443 rs826472-T	738	T	277	0.3753	T	273	0.3699

rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATCATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAGTCACTATCTCTCATCAGGAT	rs826472 [CE T]-GRCh38-Chr10:2364342-2364443 rs826472-T; 2364376-G>C						T1	2	0.0027
rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAGTCACTCTCTCATCAGGAT	rs826472 [CE T]-GRCh38-Chr10:2364342-2364443 rs826472-T, 2364427-2364430 DEL [CTAT]						T2	2	0.0027
rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAGTCACTATCTCTCACCAGGAT	rs826472 [CE C]-GRCh38-Chr10:2364342-2364443 rs826472-C	738	C	461	0.6247		C	436	0.5908
rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAGTCACTATCTCTCACCAGGAT	rs826472 [CE C]-GRCh38-Chr10:2364342-2364443 rs138540224-C; rs826472-C						C1	23	0.0312
rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTT GTCTCTGTTATATTAGTCACTATCTCTCACCAGGAT	rs826472 [CE C]-GRCh38-Chr10:2364342-2364443 rs826472-C; rs950907051-G						C2	1	0.0014
rs826472	TAGCATTTTTATGCATCAATAAGATGTTAATTATGATCAGATATTTACAATGATGCTGAATTTTG TCTCTGTTATATTAGTCACTATCTCTCACCAGGAT	rs826472 [CE C]-GRCh38-Chr10:2364342-2364443 rs826472-C; rs1588389080-A						C3	1	0.0014
rs873196	TGCCCTTGTAAATGTGAACATGCCTGATTGACTCCAACCTCCTGCCAGCCTTGGCATGCTCATTT GTGAACC	rs873196 [CE C]-GRCh38-Chr14:98379189-98379259 rs873196-C	738	C	229	0.3103				
rs873196	TGCCCTTGTAAATGTGAACATGCCTGATTGACTCCAACCTCCTGCCAGCCTTGGCATGCTCATTT GTGAACC	rs873196 [CE T]-GRCh38-Chr14:98379189-98379259 rs873196-T	738	T	509	0.6897				
rs876724	AGTATATTTTTGTACACTCTGCTAACTGCCTGCTCATAGATATTCAAATTTAGTAGATGTAGA TA	rs876724 [CE T]-GRCh38-Chr2:114970-115036 rs876724-T; rs300773-T	738	T	177	0.2398		T	177	0.2398
rs876724	AGTACATTTTTGTACACTCTGCTAACTGCCTGCTCATAGATATTCAAATTTAGTAGATGTAGA CA	rs876724 [CE C]-GRCh38-Chr2:114970-115036 rs876724-C	738	C	561	0.7602		C	178	0.2412
rs876724	AGTACATTTTTGTACACTCTGCTAACTGCCTGCTCATAGATATTCAAATTTAGTAGATGTAGA TA	rs876724 [CE C]-GRCh38-Chr2:114970-115036 rs876724-C; rs300773-T						C1	372	0.5041
rs876724	AGTACATTTTTGTACACTCTGCTAACTGCCTGCTCATAGATATTCAAATTTAGTAGATGTAGA TA	rs876724 [CE C]-GRCh38-Chr2:114970-115036 rs876724-C; rs114448669-C; rs300773-T						C2	8	0.0108
rs876724	AGTACATTTTTCTCACACTCTGCTAACTGCCTGCTCATAGATATTCAAATTTAGTAGATGTAGA TA	rs876724 [CE C]-GRCh38-Chr2:114970-115036 rs876724-C; rs77642176-C; rs300773-T						C3	3	0.0041
rs891700	GTGTTAGCAGTAAACATTTTCATCAAATTTCCATTCTTTTTTTTGAAGCCTACTTGCATAGTT CTAAGG	rs891700 [CE A]-GRCh38-Chr1:239718572-239718642 rs891700-A	738	A	322	0.4363		A	322	0.4363
rs891700	GTGTTAGCAGTAAACATTTTCATCAAATTTCCATTCTTTTTTTTGAAGCCTGCTTGCATAGTT CTAAGG	rs891700 [CE G]-GRCh38-Chr1:239718572-239718642 rs891700-G	738	G	416	0.5637		G	385	0.5217
rs891700	GTGTTAGCAGTAAACATTTTCATCAAATTTCCATTCTTTTTTTTGAAGCCTGCTTGCATAGTT CTAAGG	rs891700 [CE G]-GRCh38-Chr1:239718572-239718642 rs891700-G; rs12047255-A						G1	24	0.0325
rs891700	GTGTTACCAGTAAACATTTTCATCAAATTTCCATTCTTTTTTTTGAAGCCTGCTTGCATAGT TCTAAGG	rs891700 [CE G]-GRCh38-Chr1:239718572-239718642 rs891700-G; rs12047255-C; rs543563536-INS [T]						G2	6	0.0081
rs891700	GTGTTAGCAGTAAACATTTTCATCAAATTTCCATTCTTTTTTTTGAAGCCTGCTTGCATAGTT CTAAG	rs891700 [CE G]-GRCh38-Chr1:239718572-239718642 rs891700-G; rs543563536-DEL [T]						G3	1	0.0014
rs901398	GTGCAAAGTAGCTGAATATCAGCCCCGTTGATAGCTAACATTAGT	rs901398 [CE C]-GRCh38-Chr11:11074649-11074693 rs901398-C	738	C	225	0.3049				
rs901398	GTGCAAAGTAGCTGAATATCAGCCCCGTTGATAGCTAACATTAGT	rs901398 [CE T]-GRCh38-Chr11:11074649-11074693 rs901398-T	738	T	513	0.6951				
rs914165	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAAGAGGGCGACACCCTCAGGCAGCTCTG CTGAGCCCCGCCCCACCCAGTGCAAAAAGGTGACTGGTCTGCACTC	rs914165 [CE G]-GRCh38-Chr21:41043962-41044069 rs914165-G	738	G	379	0.5136		G	379	0.5136
rs914165	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAAGAGGGCAACACCCTCAGGCAGCTCTG CTGAGCCCCGCCCCACCCAGTGCAAAAAGGTGACTGGTCTGCACTC	rs914165 [CE A]-GRCh38-Chr21:41043962-41044069 rs914165-A	738	A	359	0.4864		A	351	0.4756
rs914165	CAAGCAGCAGAGCCTGGATGCTGATGGGCACCAAAGAGGGCAACACCCTCAGGCAGCTCTG CTGAGCCCCGCCCCACCCAGTGCAAAAAGGTGACTGGTCTGCACTC	rs914165 [CE A]-GRCh38-Chr21:41043962-41044069 rs914165-A; rs755095-G						A1	8	0.0108
rs917118	CCATGAAGATGGAGTCAACATTTTACAAGACGCTCGTTGACCTCAGTCATCTCTTACCACTTGT GCTC	rs917118 [CE C]-GRCh38-Chr7:4417342-4417409 rs917118-C	738	C	465	0.6301				
rs917118	CCATGAAGATGGAGTCAACATTTTACAAGATGCTCGTTGACCTCAGTCATCTCTTACCACTTGT GCTC	rs917118 [CE T]-GRCh38-Chr7:4417342-4417409 rs917118-T	738	T	273	0.3699				
rs938283	CATACATTGAAGTCCTAACCCCTAGTACGTTAGATGTGACCGCATTTGGAGAT	rs938283 [CE C]-GRCh38-Chr17:79472374-79472426 rs938283-C	738	C	154	0.2087				
rs938283	CATACATTGAAGTCCTAACCCCTAGTACGTTAGATGTGACCGTATTTGGAGAT	rs938283 [CE T]-GRCh38-Chr17:79472374-79472426 rs938283-T	738	T	584	0.7913				

rs964681	GACACGGGCATTTGGGGCCACAGTGCTCAGACAGCAACCTCTGGTTCTTACCAATCC	rs964681 [CE C]-GRCh38-Chr10:130900152-130900208 rs964681-C	738	C	280	0.3794			
rs964681	GACATGGGCATTTGGGGCCACAGTGCTCAGACAGCAACCTCTGGTTCTTACCAATCC	rs964681 [CE T]-GRCh38-Chr10:130900152-130900208 rs964681-T	738	T	458	0.6206			
rs987640	ACAGGTACATTCACTTAACAGGCTCTCTTTCCACCCTGTAGAAATACAAAAATAAGACTTAATA CAGACGATGG	rs987640 [CE T]-GRCh38-Chr22:33163486-33163560 rs987640-T	738	T	373	0.5054	T	373	0.5054
rs987640	ACAGGTACATTCACTTAACAGGCTCTCTTTCCACCCTGTAGAAATACAAAAATAAGACTTAAT ACAGACGATGG	rs987640 [CE A]-GRCh38-Chr22:33163486-33163560 rs987640-A	738	A	365	0.4946	A	307	0.4160
rs987640	ACCGGTACATTCACTTAACAGGCTCTCTTTCCACCCTGTAGAAATACAAAAATAAGACTTAAT ACAGACGATGG	rs987640 [CE A]-GRCh38-Chr22:33163486-33163560 rs17793354-C; rs987640-A					A1	58	0.0786
rs9905977	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGGCTCAGTCAGCATCAAGATTCTTTCTGCTTTTC CCCTCTGCCCTCCCTGGCTTGTGACGTTTGTCCCTCAGGCTTGGCCCTCTCGTGGCC	rs9905977 [CE A]-GRCh38-Chr17:3016058-3016176 rs9905977-A	738	A	219	0.2967	A	219	0.2967
rs9905977	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGGCTCAGTCAGCGTCAAGATTCTTTCTGCTTTTC CCCTCTGCCCTCCCTGGCTTGTGACGTTTGTCCCTCAGGCTTGGCCCTCTCGTGGCC	rs9905977 [CE G]-GRCh38-Chr17:3016058-3016176 rs9905977-G	738	G	519	0.7033	G	379	0.5136
rs9905977	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGGCTCAGTCAGCGTCAAGATTCTTTCTGCTTTTC CCCTCTGCCCTCCCTAGCTTGTGACGTTTGTCCCTCAGGCTTGGCCCTCTCGTGGCC	rs9905977 [CE G]-GRCh38-Chr17:3016058-3016176 rs9905977-G; rs28582109-A					G1	11	0.0149
rs9905977	TGGTGTCCAAGGAGGGCTGGGTGACTCGTGGCTCAGTCAGCGTCAAGATTCTTTCTGCTTTTC CCCTCTGCCCTCCCTGGCTTGTGACGTTTGTCCCTCAGGCTTGGCCCTCTCGTGGCC	rs9905977 [CE G]-GRCh38-Chr17:3016058-3016176 rs9905977-G; rs73298992-T					G2	129	0.1748
rs993934	TTATGAACCTAAGCTAATATACTCTGGAGACTGTTATCACAACTTTGCTCTATTGCCTTACAAA GCAAA	rs993934 [CE A]-GRCh38-Chr2:123351571-123351640 rs993934-A	738	T(A)	427	0.5786			
rs993934	TTACGAACCTAAGCTAATATACTCTGGAGACTGTTATCACAACTTTGCTCTATTGCCTTACAAA GCAAA	rs993934 [CE G]-GRCh38-Chr2:123351571-123351640 rs993934-G	738	C(G)	311	0.4214			
rs9951171	GGGAGAAAGTCCTCGTTGTTCTCTGGGATGCAACATGAGAGAGCAGCACACTGAGGCTTT ATGGGTTGCCCTG	rs9951171 [CE G]-GRCh38-Chr18:9749816-9749890 rs9951171-G	738	G	428	0.5799	G	425	0.5759
rs9951171	GGGAGAACAGTCCTCGTTGTTCTCTGGGATGCAACATGAGAGAGCAGCACACTGAGGCTTT ATGGGTTGCCCTG	rs9951171 [CE G]-GRCh38-Chr18:9749816-9749890 rs145524126-C; rs9951171-G					G1	3	0.0041
rs9951171	GGGAGAAAGTCCTCGTTGTTCTCTGGGATGCAACATGAGAGAGCAGCACACTGAGGCTTT ATGGATTGCCCTG	rs9951171 [CE A]-GRCh38-Chr18:9749816-9749890 rs9951171-A	738	A	310	0.4201	A	309	0.4187
rs9951171	GGGAGAAAGTCCTCGTTGTTCTCTGGGATGCAACATGAGAGAGCAGCACACTGAGGTTT TGAATTGCCCTG	rs9951171 [CE A]-GRCh38-Chr18:9749816-9749890 rs9951171-A; 9749874-C>T					A1	1	0.0014
rs907100	TGATGCCCTGGCATCAAAGAAGGCTCCAACCTGGCTTCTTTCTGTGTTTTCCAAGGCTTGGA AAG	rs907100 [CE C]-GRCh38-Chr2:238654924-238654989 rs907100-C	738	C	386	0.5230	C	200	0.2710
rs907100	TGATGCCCTGGCATCAAAGAAGGCTCCAACCTGAGCTTCTTTCTGTGTTTTCCAAGGCTTGGA AAG	rs907100 [CE C]-GRCh38-Chr2:238654924-238654989 rs907100-C; rs11689319-A					C1	186	0.2520
rs907100	TGATGCCCTGGCATGAAAGAAGGCTCCAACCTGGCTTCTTTCTGTGTTTTCCAAGGCTTGGA AAG	rs907100 [CE G]-GRCh38-Chr2:238654924-238654989 rs907100-G	738	G	352	0.4770	G	352	0.4770
rs6955448	TAAGTTGGAATAGGTGATGCAAAGCCCTGCCGTTACTACTATTGTTGAAAAGTTCATGGGAGA AAGTTGTAAA	rs6955448 [CE T]-GRCh38-Chr7:4270677-4270749 rs6950322-A; rs6955448-T	738	T	207	0.2805	T	207	0.2805
rs6955448	TAAGTTGGGATAGGTGATGCAAAGCCCTGCCGTTACTACTATTGTTGAAAAGTTCACGGGAGA AAGTTGTAAA	rs6955448 [CE C]-GRCh38-Chr7:4270677-4270749 rs6955448-C	738	C	531	0.7195	C	530	0.7182
rs6955448	TAAGTTGGGATAGGTGATTCAAAGCCCTGCCGTTACTACTATTGTTGAAAAGTTCACGGGAGA AAGTTGTAAA	rs6955448 [CE C]-GRCh38-Chr7:4270677-4270749 rs140855431-T; rs6955448-C					C1	1	0.0014
rs4530059	CAGAGCTCCAGAAGCAACTCCAGCACACAGAGAGGCGCTGATGTGCCTGTCAGGTGCTGCTA CTGAGGAAGCCGTTGCTGGTCTCCGGAAGCTCTTGATATCCACAGGAGTGCCGTCTGCCTGGC CTCC	rs4530059 [CE A]-GRCh38-Chr14:104302784-104302911 rs4530059-A	738	A	271	0.3672	A	32	0.0434
rs4530059	CAGAGCTCCAGAAGCAACTCCAGCACACAGAGAGGCGCTGATGTGCCTGTCAGGTGCTGCTA CTGAGGAAGCCGTTGCTGGTCTCCGGAAGCTCTTGATATCCACAGGAGTGCCGTCTGCCTGGC CTCC	rs4530059 [CE A]-GRCh38-Chr14:104302784-104302911 rs4530059-A; rs4450333-T					A1	239	0.3238
rs4530059	CAGAGCTCCAGAAGCAACTCCAGCACACGGAGAGGCGCTGATGTGCCTGTCAGGTGCTGCTA CTGAGGAAGCCGTTGCTGGTCTCCGGAAGCTCTTGATATCCACAGGAGTGCCGTCTGCCTGGC CTCC	rs4530059 [CE G]-GRCh38-Chr14:104302784-104302911 rs4530059-G	738	G	467	0.6328	G	466	0.6314
rs4530059	CAGAGCTCCAGAAGCAACTCCAGCACACGGAGAGGCACTGATGTGCCTGTCAGGTGCTGCTA CTGAGGAAGCCGTTGCTGGTCTCCGGAAGCTCTTGATATCCACAGGAGTGCCGTCTGCCTGGC CTCC	rs4530059 [CE G]-GRCh38-Chr14:104302784-104302911 rs4530059-G; rs999246876-A					G1	1	0.0014

rs338882	GCCTGTGCATACACACGTTTGGGACAAGGGCTGGATTCTTCGGCTGGGATGTCTCTCAGAGC TCTTGACTTGGTCCCTTTGGCTGGGGCTTGCCGTGAGGTGTGGGCTGCGCCACG	rs338882 [CE A]-GRCh38-Chr5:179263618-179263733 rs338882-A	738	T(A)	404	0.5474
rs338882	GCCTGTGCACACACACGTTTGGGACAAGGGCTGGATTCTTCGGCTGGGATGTCTCTCAGAGC TCTTGACTTGGTCCCTTTGGCTGGGGCTTGCCGTGAGGTGTGGGCTGCGCCACG	rs338882 [CE G]-GRCh38-Chr5:179263618-179263733 rs338882-G	738	C(G)	334	0.4526

Appendix 10. Hardy-Weinberg Equilibrium (HWE) test.

Observed and expected heterozygosity and Fisher's Exact test *P*-values for target iiSNPs and target iiSNPs with flanking region variants. Significant departures from HWE are highlighted in grey.

Locus	Targeted iiSNPs			iiSNP-flank		
	Observed heterozygosity	Expected heterozygosity	<i>P</i> -value	Observed heterozygosity	Expected heterozygosity	<i>P</i> -value
rs1109037	0.47154	0.498	0.34418	0.68835	0.71829	0.17828
rs1005533	0.50407	0.49985	0.91657	0.50407	0.49985	0.91677
rs10092491	0.4878	0.45537	0.19992	0.4878	0.45537	0.20765
rs1015250	0.38753	0.40178	0.51733	0.41734	0.42949	0.49611
rs1024116	0.48509	0.50066	0.60335	0.48509	0.50066	0.60313
rs1028528	0.4336	0.46749	0.18416	0.4336	0.46749	0.18049
rs10488710	0.44687	0.49101	0.08664	0.44687	0.49101	0.08918
rs10495407	0.44173	0.43854	0.90683	0.44173	0.43854	0.90573
rs1058083	0.43902	0.44322	0.90684	0.43902	0.44322	0.90616
rs10773760	0.49593	0.46819	0.26889	0.49593	0.46819	0.26714
rs10776839	0.46612	0.50023	0.21117	0.59621	0.61642	0.21295
rs1294331	0.45257	0.47831	0.32241	0.45528	0.47937	0.23757
rs12997453	0.38211	0.40418	0.30844	0.40921	0.43628	0.24819
rs13182883	0.47154	0.48214	0.74472	0.47154	0.48323	0.8005
rs13218440	0.44173	0.48654	0.08611	0.44173	0.48654	0.08703
rs1335873	0.42276	0.48108	0.02303	0.42547	0.48217	0.01829
rs1336071	0.46883	0.50031	0.24981	0.50136	0.53392	0.08211
rs1355366	0.49864	0.49714	1	0.49864	0.49714	1
rs1360288	0.43902	0.43949	1	0.44444	0.44314	1
rs1382387	0.28455	0.35652	0.00025	0.28455	0.35652	0.00013
rs1413212	0.41463	0.43065	0.47076	0.41463	0.43065	0.47185
rs1454361	0.49864	0.49323	0.9152	0.49864	0.49323	0.91524
rs1463729	0.49864	0.49873	1	0.50407	0.50571	0.52477
rs1490413	0.46612	0.48873	0.39412	0.46612	0.48988	0.47529
rs1498553	0.47696	0.48956	0.66766	0.47696	0.48956	0.67006
rs1523537	0.50949	0.498	0.67518	0.52575	0.51357	0.92883
rs1528460	0.47696	0.47999	0.91504	0.47696	0.47999	0.91385
rs159606	0.43902	0.44861	0.72449	0.43902	0.44861	0.72508
rs1821380	0.43631	0.44414	0.81515	0.43631	0.44775	0.16055
rs1886510	0.44715	0.44948	1	0.44715	0.44948	1
rs1979255	0.39295	0.43661	0.05748	0.39566	0.43845	0.10166
rs2040411	0.52846	0.48368	0.08307	0.52846	0.48368	0.08367
rs2046361	0.42547	0.46536	0.11728	0.43089	0.46878	0.17068
rs2056277	0.30081	0.28533	0.36229	0.3523	0.33298	0.67324
rs2076848	0.48641	0.49562	0.75515	0.51902	0.5292	0.64107
rs2107612	0.40108	0.45034	0.03937	0.40108	0.45034	0.04044
rs2111980	0.47696	0.50038	0.40795	0.48238	0.50859	0.02738
rs214955	0.46883	0.49419	0.3418	0.4878	0.51463	0.58384
rs221956	0.44986	0.46888	0.44024	0.44986	0.46888	0.44211
rs2269355	0.47154	0.48417	0.66671	0.47154	0.48417	0.66708
rs2342747	0.44444	0.42549	0.4657	0.44444	0.42549	0.46481

rs2399332	0.43631	0.45455	0.49211	0.45799	0.48254	0.56928
rs251934	0.44173	0.46956	0.2692	0.44173	0.46956	0.26616
rs279844	0.4336	0.45858	0.30606	0.45257	0.47666	0.176
rs2830795	0.36585	0.37619	0.58112	0.50949	0.52493	0.60994
rs2831700	0.47696	0.48608	0.74916	0.47696	0.48608	0.7484
rs321198	0.50136	0.48831	0.66766	0.50407	0.49171	0.86108
rs338882	0.45799	0.49643	0.14665	0.45799	0.49643	0.1438
rs354439	0.4336	0.47999	0.0649	0.44173	0.48595	0.07582
rs3780962	0.48238	0.49509	0.67652	0.48238	0.49509	0.67592
rs430046	0.44444	0.47023	0.3187	0.44444	0.47125	0.42169
rs4364205	0.5122	0.48162	0.23308	0.5122	0.48162	0.23215
rs445251	0.49051	0.47944	0.66652	0.49864	0.48473	0.50141
rs4530059	0.43089	0.46536	0.18001	0.4607	0.4952	0.25208
rs4606077	0.43902	0.45205	0.64285	0.4878	0.5013	0.61605
rs560681	0.46883	0.48466	0.59228	0.48509	0.50333	0.65938
rs576261	0.46883	0.48744	0.52157	0.46883	0.48744	0.52276
rs6444724	0.46883	0.48266	0.59298	0.46883	0.48266	0.59318
rs6811238	0.54472	0.50055	0.09564	0.54472	0.50055	0.09534
rs6955448	0.40921	0.40418	0.89706	0.40921	0.40613	0.36772
rs717302	0.45528	0.498	0.12148	0.45528	0.498	0.11463
rs719366	0.37398	0.39316	0.35429	0.37398	0.39316	0.35424
rs722098	0.40379	0.42443	0.38677	0.40379	0.42443	0.38942
rs722290	0.50407	0.49759	0.83275	0.50407	0.49759	0.83457
rs727811	0.44444	0.49873	0.03727	0.44986	0.50302	0.07931
rs729172	0.46612	0.45373	0.6438	0.46612	0.45549	0.38456
rs733164	0.42276	0.44684	0.35089	0.42276	0.44684	0.35325
rs735155	0.50949	0.49185	0.52357	0.51762	0.50302	0.53229
rs737681	0.46883	0.48744	0.51853	0.47154	0.48857	0.3954
rs740598	0.45528	0.49714	0.1155	0.4607	0.50007	0.1171
rs740910	0.38211	0.38154	1	0.5393	0.53103	0.48865
rs763869	0.48238	0.48788	0.83306	0.48238	0.48788	0.83097
rs8037429	0.45799	0.50031	0.11881	0.45799	0.50031	0.11734
rs8078417	0.39024	0.40772	0.44421	0.40108	0.41505	0.72557
rs826472	0.49051	0.46956	0.43639	0.54201	0.51384	0.19061
rs873196	0.39295	0.42861	0.11607	0.39295	0.42861	0.11472
rs876724	0.34959	0.36513	0.47575	0.62331	0.63094	0.66063
rs891700	0.48238	0.49256	0.75204	0.52304	0.53708	0.09635
rs901398	0.42547	0.42443	1	0.42547	0.42443	1
rs907100	0.55827	0.49962	0.02902	0.67751	0.6364	0.13739
rs914165	0.48509	0.50031	0.60446	0.49864	0.51064	0.64606
rs917118	0.45799	0.46679	0.73602	0.45799	0.46679	0.7387
rs938283	0.2981	0.3307	0.07986	0.2981	0.3307	0.0812
rs964681	0.45528	0.47155	0.51167	0.45528	0.47155	0.5112
rs987640	0.46883	0.50062	0.24726	0.52846	0.56609	0.40852
rs9905977	0.40379	0.41794	0.53453	0.58266	0.61827	0.23603
rs993934	0.48509	0.48831	0.91448	0.48509	0.48831	0.91501
rs9951171	0.44986	0.48788	0.13334	0.45528	0.4937	0.32024

Appendix 11. Forensic statistics for the target iiSNPs (iiSNP) and the target iiSNPs with flanking region variants (iiSNP-flank).

locus	Number of samples alleles	Number of observed alleles		Random match probability (RMP)		Power of discrimination (PD)		Power of exclusion (PE)		Typical paternity index (TPI)		Polymorphism information content (PIC)	
		iiSNP	iiSNP-flank	iiSNP	iiSNP-flank	iiSNP	iiSNP-flank	iiSNP	iiSNP-flank	iiSNP	iiSNP-flank	iiSNP	iiSNP-flank
rs826472	738	2	7	0.401	0.345	0.599	0.655	0.179	0.227	0.981	1.092	0.359	0.417
rs1109037	738	2	6	0.365	0.127	0.635	0.873	0.164	0.410	0.946	1.604	0.374	0.668
rs2399332	738	2	6	0.395	0.354	0.605	0.646	0.138	0.153	0.887	0.923	0.351	0.389
rs279844	738	2	6	0.390	0.362	0.610	0.638	0.136	0.149	0.883	0.913	0.353	0.383
rs354439	738	2	6	0.369	0.362	0.631	0.638	0.136	0.141	0.883	0.896	0.364	0.373
rs735155	738	2	6	0.389	0.370	0.611	0.630	0.196	0.203	1.019	1.037	0.371	0.388
rs1336071	738	2	5	0.361	0.318	0.639	0.682	0.162	0.189	0.941	1.003	0.375	0.426
rs2830795	738	2	5	0.459	0.280	0.541	0.720	0.094	0.196	0.788	1.019	0.305	0.471
rs8078417	738	2	5	0.431	0.420	0.569	0.580	0.108	0.115	0.820	0.835	0.324	0.334
rs876724	738	2	5	0.469	0.200	0.531	0.800	0.086	0.320	0.769	1.327	0.298	0.565
rs891700	738	2	5	0.375	0.317	0.625	0.683	0.173	0.208	0.966	1.048	0.371	0.432
rs10776839	738	2	4	0.360	0.215	0.640	0.785	0.160	0.286	0.937	1.238	0.375	0.539
rs1463729	738	2	4	0.376	0.366	0.624	0.634	0.186	0.191	0.997	1.008	0.374	0.384
rs2076848	736	2	4	0.374	0.329	0.626	0.671	0.176	0.205	0.974	1.040	0.372	0.419
rs4530059	738	2	4	0.383	0.335	0.617	0.665	0.134	0.155	0.879	0.927	0.357	0.409
rs4606077	738	2	4	0.399	0.309	0.601	0.691	0.139	0.177	0.891	0.976	0.350	0.440
rs740910	738	2	4	0.456	0.287	0.544	0.713	0.103	0.224	0.809	1.085	0.308	0.466
rs9905977	738	2	4	0.423	0.210	0.577	0.790	0.116	0.271	0.839	1.198	0.330	0.549
rs9951171	738	2	4	0.366	0.358	0.634	0.642	0.147	0.151	0.909	0.918	0.369	0.377
rs1015250	738	2	3	0.437	0.382	0.563	0.618	0.107	0.125	0.816	0.858	0.321	0.377
rs1294331	738	2	3	0.377	0.376	0.623	0.624	0.149	0.151	0.913	0.918	0.364	0.365
rs12997453	738	2	3	0.433	0.366	0.567	0.634	0.103	0.120	0.809	0.846	0.322	0.387
rs13182883	738	2	3	0.380	0.378	0.620	0.622	0.164	0.164	0.946	0.946	0.366	0.367
rs1335873	738	2	3	0.365	0.364	0.635	0.636	0.128	0.130	0.866	0.870	0.365	0.367
rs1360288	738	2	3	0.411	0.406	0.589	0.594	0.139	0.143	0.891	0.900	0.343	0.347
rs1490413	738	2	3	0.372	0.369	0.628	0.631	0.160	0.160	0.937	0.937	0.369	0.371
rs1523537	738	2	3	0.383	0.360	0.617	0.640	0.196	0.211	1.019	1.054	0.374	0.396
rs1821380	738	2	3	0.406	0.401	0.594	0.599	0.138	0.138	0.887	0.887	0.345	0.350
rs1979255	738	2	3	0.403	0.400	0.597	0.600	0.110	0.111	0.824	0.827	0.341	0.343
rs2046361	738	2	3	0.381	0.377	0.619	0.623	0.130	0.134	0.870	0.879	0.357	0.361
rs2056277	738	2	3	0.550	0.481	0.450	0.519	0.064	0.087	0.715	0.772	0.244	0.294
rs2111980	738	2	3	0.365	0.355	0.635	0.645	0.168	0.173	0.956	0.966	0.375	0.387
rs214955	738	2	3	0.367	0.339	0.633	0.661	0.162	0.177	0.941	0.976	0.372	0.400
rs321198	738	2	3	0.388	0.382	0.612	0.618	0.189	0.191	1.003	1.008	0.369	0.374
rs430046	738	2	3	0.382	0.380	0.618	0.620	0.143	0.143	0.900	0.900	0.359	0.361
rs445251	738	2	3	0.392	0.384	0.608	0.616	0.179	0.186	0.981	0.997	0.364	0.373
rs560681	738	2	3	0.377	0.351	0.623	0.649	0.162	0.175	0.941	0.971	0.367	0.392
rs6955448	738	2	3	0.438	0.436	0.562	0.564	0.120	0.120	0.846	0.846	0.322	0.324
rs727811	738	2	3	0.354	0.348	0.646	0.652	0.143	0.147	0.900	0.909	0.374	0.380
rs729172	738	2	3	0.407	0.404	0.593	0.596	0.160	0.160	0.937	0.937	0.350	0.353

rs737681	738	2	3	0.374	0.373	0.626	0.627	0.162	0.164	0.941	0.946	0.368	0.370
rs740598	738	2	3	0.359	0.356	0.641	0.644	0.151	0.155	0.918	0.927	0.373	0.377
rs907100	738	2	3	0.410	0.218	0.590	0.782	0.244	0.394	1.132	1.550	0.374	0.564
rs914165	738	2	3	0.368	0.354	0.632	0.646	0.175	0.186	0.971	0.997	0.375	0.391
rs987640	738	2	3	0.361	0.271	0.639	0.729	0.162	0.214	0.941	1.060	0.375	0.472
rs1005533	738	2	2	0.378	0.378	0.622	0.622	0.191	0.191	1.008	1.008	0.375	0.375
rs10092491	738	2	2	0.414	0.414	0.586	0.586	0.177	0.177	0.976	0.976	0.351	0.351
rs1024116	738	2	2	0.368	0.368	0.632	0.632	0.175	0.175	0.971	0.971	0.375	0.375
rs1028528	738	2	2	0.382	0.382	0.618	0.618	0.136	0.136	0.883	0.883	0.358	0.358
rs10488710	734	2	2	0.362	0.362	0.638	0.638	0.145	0.145	0.904	0.904	0.370	0.370
rs10495407	738	2	2	0.413	0.413	0.587	0.587	0.141	0.141	0.896	0.896	0.342	0.342
rs1058083	738	2	2	0.407	0.407	0.593	0.593	0.139	0.139	0.891	0.891	0.345	0.345
rs10773760	738	2	2	0.405	0.405	0.595	0.595	0.184	0.184	0.992	0.992	0.358	0.358
rs13218440	738	2	2	0.365	0.365	0.635	0.635	0.141	0.141	0.896	0.896	0.368	0.368
rs1355366	738	2	2	0.378	0.378	0.622	0.622	0.186	0.186	0.997	0.997	0.373	0.373
rs1382387	738	2	2	0.481	0.481	0.519	0.519	0.057	0.057	0.699	0.699	0.293	0.293
rs1413212	738	2	2	0.413	0.413	0.587	0.587	0.123	0.123	0.854	0.854	0.338	0.338
rs1454361	738	2	2	0.382	0.382	0.618	0.618	0.186	0.186	0.997	0.997	0.371	0.371
rs1498553	738	2	2	0.375	0.375	0.625	0.625	0.168	0.168	0.956	0.956	0.369	0.369
rs1528460	738	2	2	0.385	0.385	0.615	0.615	0.168	0.168	0.956	0.956	0.364	0.364
rs159606	738	2	2	0.402	0.402	0.598	0.598	0.139	0.139	0.891	0.891	0.348	0.348
rs1886510	738	2	2	0.404	0.404	0.596	0.596	0.145	0.145	0.904	0.904	0.348	0.348
rs2040411	738	2	2	0.407	0.407	0.593	0.593	0.214	0.214	1.060	1.060	0.366	0.366
rs2107612	738	2	2	0.390	0.390	0.610	0.610	0.115	0.115	0.835	0.835	0.349	0.349
rs221956	738	2	2	0.385	0.385	0.615	0.615	0.147	0.147	0.909	0.909	0.359	0.359
rs2269355	738	2	2	0.378	0.378	0.622	0.622	0.164	0.164	0.946	0.946	0.367	0.367
rs2342747	738	2	2	0.427	0.427	0.573	0.573	0.143	0.143	0.900	0.900	0.335	0.335
rs251934	738	2	2	0.382	0.382	0.618	0.618	0.141	0.141	0.896	0.896	0.359	0.359
rs2831700	738	2	2	0.379	0.379	0.621	0.621	0.168	0.168	0.956	0.956	0.368	0.368
rs338882	738	2	2	0.361	0.361	0.639	0.639	0.153	0.153	0.923	0.923	0.373	0.373
rs3780962	738	2	2	0.372	0.372	0.628	0.628	0.173	0.173	0.966	0.966	0.372	0.372
rs4364205	738	2	2	0.400	0.400	0.600	0.600	0.198	0.198	1.025	1.025	0.365	0.365
rs576261	738	2	2	0.374	0.374	0.626	0.626	0.162	0.162	0.941	0.941	0.368	0.368
rs6444724	738	2	2	0.379	0.379	0.621	0.621	0.162	0.162	0.941	0.941	0.366	0.366
rs6811238	738	2	2	0.400	0.400	0.600	0.600	0.230	0.230	1.098	1.098	0.375	0.375
rs717302	738	2	2	0.358	0.358	0.642	0.642	0.151	0.151	0.918	0.918	0.374	0.374
rs719366	738	2	2	0.443	0.443	0.557	0.557	0.099	0.099	0.799	0.799	0.316	0.316
rs722098	738	2	2	0.417	0.417	0.583	0.583	0.116	0.116	0.839	0.839	0.334	0.334
rs722290	738	2	2	0.380	0.380	0.620	0.620	0.191	0.191	1.008	1.008	0.373	0.373
rs733164	738	2	2	0.399	0.399	0.601	0.601	0.128	0.128	0.866	0.866	0.347	0.347
rs763869	738	2	2	0.379	0.379	0.621	0.621	0.173	0.173	0.966	0.966	0.369	0.369
rs8037429	738	2	2	0.357	0.357	0.643	0.643	0.153	0.153	0.923	0.923	0.375	0.375
rs873196	738	2	2	0.411	0.411	0.589	0.589	0.110	0.110	0.824	0.824	0.336	0.336
rs901398	738	2	2	0.422	0.422	0.578	0.578	0.130	0.130	0.870	0.870	0.334	0.334
rs917118	738	2	2	0.390	0.390	0.610	0.610	0.153	0.153	0.923	0.923	0.358	0.358
rs938283	738	2	2	0.505	0.505	0.495	0.495	0.063	0.063	0.712	0.712	0.276	0.276
rs964681	738	2	2	0.385	0.385	0.615	0.615	0.151	0.151	0.918	0.918	0.360	0.360
rs993934	738	2	2	0.380	0.380	0.620	0.620	0.175	0.175	0.971	0.971	0.369	0.369

Appendix 12. Pairwise F_{ST} genetic distance.

Appendix 12A. Pairwise F_{ST} distances, based on the genotypes of 87 iiSNPs among individuals from Kuwait (KUW), Saudi Arabia (KSA) and seven global population groups: Africa (AFR), America (AMR), Central South Asia (CSA), the Middle East (ME), Europe (EUR), East Asia (EAS) and Oceania (OCE), grouped according to the 51 HGDP-CEPH population panel.

	AFR	EUR	ME	CSA	EAS	OCE	AMR	KUW	KSA
AFR	0								
EUR	0.10553	0							
ME	0.08722	0.00891	0						
CSA	0.09227	0.01893	0.01124	0					
EAS	0.11088	0.04472	0.03572	0.02583	0				
OCE	0.1537	0.10835	0.09147	0.08475	0.10116	0			
AMR	0.15093	0.06932	0.06831	0.06017	0.05896	0.14658	0		
KUW	0.08026	0.00945	0.00183	0.00722	0.03345	0.08358	0.0657	0	
KSA	0.10018	0.0124	0.00284	0.01454	0.0367	0.09534	0.07205	0.00348	0

Appendix 12B. Pairwise F_{ST} distances based on the genotypes of 87 iiSNPs among individuals from Kuwaiti (KUW), Saudi Arabian (KSA), Palestinian, Druze, Bedouin, and Mozabite populations.

	Mozabite	Bedouin	Druze	Palestinian	Kuwait	KSA
Mozabite	0					
Bedouin	0.01624	0				
Druze	0.00873	0.00935	0			
Palestinian	0.00916	0.00775	0.00435	0		
Kuwait	0.00992	0.00403	0.00442	0.00391	0	
KSA	0.01244	0.00452	0.00701	0.00313	0.00348	0

Appendix 13. The default stutter filter percentage used for the Y and X-STR markers in the UAS.

Loci	% Stutter	% Analytical	% Interpretation
DYS19	< 15	> 1.5	> 4.5
DYS385a-b	< 20	> 1.5	> 4.5
DYF387S1	< 20	> 1.5	> 4.5
DYS389I	< 20	> 1.5	> 4.5
DYS389II	< 35	> 5	> 15
DYS390	< 15	> 1.5	> 4.5
DYS391	< 20	> 1.5	> 4.5
DYS392*	< 30	> 1.5	> 4.5
DYS437	< 45	> 1.5	> 4.5
DYS438	< 15	> 1.5	> 4.5
DYS439	< 15	> 1.5	> 4.5
DYS448	< 15	> 3.3	> 10
DYS460	< 15	> 1.5	> 4.5
DYS481	< 50	> 1.5	> 4.5
DYS505	< 15	> 1.5	> 4.5
DYS522	< 15	> 1.5	> 4.5
DYS533	< 15	> 1.5	> 4.5
DYS549	< 22	> 1.5	> 4.5
DYS570	< 22	> 1.5	> 4.5
DYS576	< 15	> 1.5	> 4.5
DYS612 ¹	< 35	> 1.5	> 4.5
DYS635	< 15	> 3.3	> 10
DYS643	< 20	> 1.5	> 4.5
Y-GATA-H4 ^{2,3}	< 35	> 1.5	> 4.5
HPRTB	< 15	> 1.5	> 4.5
DXS7132	< 22	> 1.5	> 4.5
DXS7423	< 15	> 1.5	> 4.5
DXS8378	< 15	> 1.5	> 4.5
DXS10074	< 25	> 1.5	> 4.5
DXS10103	< 22	> 1.5	> 4.5
DXS10135	< 22	> 1.5	> 4.5

Frequencies of length-based, repeat sequence-based, and flanking sequence-based alleles among 365 Kuwaiti individuals for seven X-STRs of the ForenSeq™ DNA Signature Prep Kit. All the sequences shown are as reported by the UAS flanking region report. To avoid confusion with the flanking region sequence, STR repeat region sequences are coloured red. Flanking variants are highlighted in bold. Allele nomenclature is in line with ISFG recommendations. Novel alleles observed by STRait Razor are highlighted in green. Sequence not reported by UAS are highlighted in grey.

257

DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTAT ATTTTTATT TATTTATTTTTTGAGTTGGAGTTTCGCTCTTGTGCCCCAGGCTGGAGTTCAG TGGACCATCTCAGCTCACT	480	DXS8378 [CE 6]-GRCh38-ChrX-9402182-9402373 [ATAG]6	6	1	0.002		6	1	0.002
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCGCTCTTGTGCCCCAGGCTGG AGTTCAAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 8]-GRCh38-ChrX-9402182-9402373 [ATAG]8	8	1	0.002		8	1	0.002
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCGCTCTTGTGCCCCAGG CTGGAGTTCAAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 9]-GRCh38-ChrX-9402182-9402373 [ATAG]9	9	5	0.010		9	5	0.010
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCGCTCTTGTGCCCC AGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 10]-GRCh38-ChrX-9402182-9402373 [ATAG]10	10	156	0.325		10	156	0.325
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCGCTCTTGTG CCCCAGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 11]-GRCh38-ChrX-9402182-9402373 [ATAG]11	11	186	0.388		11	186	0.388
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCGCTCT TGTGCCCCAGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 12]-GRCh38-ChrX-9402182-9402373 [ATAG]12	12	114	0.238		12	114	0.238
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCG CTCTTGTGCCCCAGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 13]-GRCh38-ChrX-9402182-9402373 [ATAG]13	13	15	0.031		13 (a)	13	0.027
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGA C AGGTCA CTATCTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTAT ATTTTATTTATTTATTTTTGAGTTGGAGTTTCG CTCTTGTGCCCCAGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 13]-GRCh38-ChrX-9402182-9402373 [ATAG]13 9402309 C>T					13 (b)	2	0.004
DXS8378	TGTTTCATTCCATCCTGGGACAGTTCACCCCCTCCTTAGGCAACCCGGTG GCCCGTGTCTGGCACAGGTCA CTATCTATCTATCTATCTATCTATCTATCTAT CTATCTATCTATCTATCTAT CTATATTTTATTTATTTATTTTTGAGTTGGAGTT TCGCTCTTGTGCCCCAGGCTGGAGTTCAGTGGCACCATCTCAGCTCACT	480	DXS8378 [CE 14]-GRCh38-ChrX-9402182-9402373 [ATAG]14	14	2	0.004		14	2	0.004
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAGAAAGAAGAAAGAAAGAGAATAGAAAAAGAGAAGAGAAAGAGAAAAAG AAAAAGAAAAGAAA	489	DXS10135 [CE 15]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]11 AAAG	15	3	0.006	15	3	0.006	
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAGAAAGAAGAAAGAAAGAAAGAGAATAGAAAAAGAGAAGAGAAAGAGAA AAGAAAAAGAAAAGAAA	489	DXS10135 [CE 16]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]12 AAAG	16	1	0.002	16	1	0.002	
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAAGAAAGAAAGAAAGAAAGGAAAAAGAGAATAGAAAGAGAAGAGAAAGAG AAAAGAAAAAGAAAAGAAA	489	DXS10135 [CE 16.2]-GRCh38-ChrX-9338302-9338453 [AAGA]16 AA GAAA rs201630737	16.2	1	0.002	16.2	1	0.002	
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAGAAAGAAGAAAGAAAGAAAGAGAATAGAAAGAAAGAGAAGAGAAAGAG AGAAAAGAAAAAGAAAAGAAA	489	DXS10135 [CE 17]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]13 AAAG	17	12	0.025	17 (a)	10	0.020	
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAGAAAGAAGAAAGAAAGAAAGAGAATAGAAAGAAAGAGAAGAGAAAGAG AGAAAAGAAAAAGAAAAGAAA	489	DXS10135 [CE 17]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]11 AAGG AAGA AAAG				17 (b)	2	0.004	
DXS10135	AAGAAAGAAGAGAAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG AAGAAAGAAGAAAGAAAGAAAGAAAGAGAATAGAAAGAAAGAGAAGAGAAAG GAGAAAAGAAAAAGAAAAGAAA	489	DXS10135 [CE 17.1]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]14 AAAG rs201630737	17.1	5	0.010	17.1	5	0.010	

	AAAGAAAGAAAAGAGAATAGAAAAGAGAAGAGAAAAGAGAAAAGAAAAAG AAAAGAAA								
DXS10135	AAGAAAAGAAAGAGAAAGGAAAGAA								

	AAAAGAGAATAGAAAAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAAAG AAA									
DXS10135	AAGAAAGAAAGAGAAAGGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAGAAAGAAAGAAAGAA AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AAAAGAGAATAGAAAAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAAAGA AA	489	DXS10135 [CE 37.2]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]17 GA [AAGA]16 AAAG	37.2	2	0.004	37.2	2	0.004	

Appendix 15- population statistics for 7 X-STRs in the Kuwaiti population.

Appendix 15A- Exact differentiation test between males and females.

	Locus	P-value Among samples	P-value Between all pairs of samples
1	HPRTB	0.52035	0.4782
2	DXS8378	0.9531	0.95425
3	DXS7423	0.37685	0.3678
4	DXS7132	0.8934	0.85615
5	DXS10135	0.95995	0.9691
6	DXS10074	0.5706	0.5503
7	DXS10103	0.1375	0.15265

Appendix15B- Results of the Hardy-Weinberg equilibrium (HWE) tests.

Observed and expected heterozygosity and p-values obtained from the Fisher's Exact test.

Locus	Genotypes	Observed Heterozygosity	Expected heterozygosity	P-value
HPRTB	118	0.81356	0.7679	0.56298
DXS8378	118	0.68644	0.69492	0.9862
DXS7423	118	0.68644	0.68619	0.96327
DXS7132	118	0.76271	0.75838	0.39537
DXS10135	118	0.94068	0.94638	0.78199
DXS10074	118	0.80508	0.85377	0.06794
DXS10103	118	0.69492	0.72413	0.28914

Appendix 16- Haplotype frequencies for four linkage groups in the Kuwaiti population.

LG1-Haplotype	LG1-Count	LG1-Frequency	LG2-Haplotype	LG2-Count	LG2-Frequency	LG3-Haplotype	LG3-Count	LG3-Frequency	LG4-Haplotype	LG4-Count	LG4-Frequency
10 17	2	0.0083	12 13	2	0.0083	10 16	1	0.0041	12	1	0.0041
10 18	3	0.0124	12 14	2	0.0083	10 20	1	0.0041	13	15	0.0622
10 19	3	0.0124	12 15	3	0.0124	11 16	7	0.029	14	99	0.4108
10 19.1	2	0.0083	12 16	4	0.0166	11 18	9	0.0373	15	85	0.3527
10 20	5	0.0207	12 17	12	0.0498	11 19	9	0.0373	16	33	0.1369
10 20.1	2	0.0083	12 18	6	0.0249	11 20	5	0.0207	17	8	0.0332
10 21	3	0.0124	12 7	1	0.0041	12 16	2	0.0083			
10 21.1	5	0.0207	12 8	2	0.0083	12 17	10	0.0415			
10 22	6	0.0249	13 13	1	0.0041	12 18	14	0.0581			
10 22.1	1	0.0041	13 14	3	0.0124	12 19	51	0.2116			
10 23	8	0.0332	13 15	5	0.0207	12 20	10	0.0415			
10 24	6	0.0249	13 16	11	0.0456	13 16	5	0.0207			
10 24.1	1	0.0041	13 17	10	0.0415	13 17	14	0.0581			
10 25	2	0.0083	13 18	12	0.0498	13 18	15	0.0622			
10 26	6	0.0249	13 19	8	0.0332	13 19	32	0.1328			
10 26.1	1	0.0041	13 7	1	0.0041	13 20	9	0.0373			
10 27	5	0.0207	13 8	14	0.0581	13 21	2	0.0083			
10 28	2	0.0083	13 9	1	0.0041	14 16	4	0.0166			
10 29	4	0.0166	13.3 17	1	0.0041	14 17	6	0.0249			
10 30	2	0.0083	14 10	1	0.0041	14 18	4	0.0166			
10 32	1	0.0041	14 12	1	0.0041	14 19	14	0.0581			
10 37	1	0.0041	14 14	4	0.0166	14 20	4	0.0166			
11 15	1	0.0041	14 15	8	0.0332	14 21	2	0.0083			
11 17	1	0.0041	14 16	10	0.0415	15 17	1	0.0041			
11 18	1	0.0041	14 17	14	0.0581	15 18	3	0.0124			
11 18.1	2	0.0083	14 18	20	0.083	15 19	2	0.0083			
11 19	3	0.0124	14 19	2	0.0083	15 20	1	0.0041			
11 19.1	1	0.0041	14 20	2	0.0083	15 21	1	0.0041			
11 20	10	0.0415	14 7	2	0.0083	16 17	1	0.0041			
11 21	13	0.0539	14 8	17	0.0705	16 19	1	0.0041			
11 21.1	1	0.0041	14 9	2	0.0083	7 19	1	0.0041			
11 22	3	0.0124	15 13	1	0.0041						
11 22.1	3	0.0124	15 14	1	0.0041						
11 23	10	0.0415	15 15	1	0.0041						
11 24	7	0.029	15 16	9	0.0373						
11 24.2	1	0.0041	15 17	12	0.0498						
11 24.3	1	0.0041	15 18	6	0.0249						
11 25	7	0.029	15 19	6	0.0249						
11 26	6	0.0249	15 20	1	0.0041						
11 27	6	0.0249	15 7	2	0.0083						
11 28	7	0.029	15 8	8	0.0332						

11 29	6	0.0249	15 9	1	0.0041						
11 30	3	0.0124	16 14	1	0.0041						
11 31	3	0.0124	16 15	1	0.0041						
11 32	3	0.0124	16 16	2	0.0083						
11 33	2	0.0083	16 17	1	0.0041						
12 17	1	0.0041	16 18	4	0.0166						
12 17.1	2	0.0083	16 20	1	0.0041						
12 18	2	0.0083	17 19	1	0.0041						
12 18.1	1	0.0041									
12 19.1	2	0.0083									
12 20	4	0.0166									
12 21	4	0.0166									
12 22	4	0.0166									
12 22.1	1	0.0041									
12 23	2	0.0083									
12 23.1	1	0.0041									
12 24	4	0.0166									
12 25	3	0.0124									
12 26	1	0.0041									
12 26.1	1	0.0041									
12 27	7	0.029									
12 28	6	0.0249									
12 29	4	0.0166									
12 30	2	0.0083									
12 31	4	0.0166									
12 32	1	0.0041									
12 37.2	1	0.0041									
13 21	1	0.0041									
13 21.1	1	0.0041									
13 22	1	0.0041									
13 23.1	1	0.0041									
13 24	1	0.0041									
13 25	1	0.0041									
13 29	1	0.0041									
9 16.2	1	0.0041									
9 19	1	0.0041									
9 23	1	0.0041									
9 31	1	0.0041									

A total of 27 Y-STR loci were typed in 248 Kuwaiti males. Loci highlighted in grey are generated by the ForenSeq™ DNA Signature Prep kit (MPS; length-based) and those highlighted in orange are generated by the PowerPlex® Y23 kit (CE). The first 19 loci are shared between the two kits, aligned next to each other for direct comparison. The black highlighted cells indicate all missing haplotypes (in MPS data); green highlighted cells indicate di/tri allelic patterns; red highlighted cells are MPS dropouts; yellow highlighted cells indicate null alleles (CE data). The last three columns are: major haplogroup prediction for the Kuwaiti individuals, prediction likelihood (%) and prediction probability fitness. The probabilities highlighted in pink were excluded due to the high chance of misclassification.

272

19 loci shared between the ForenSeq and the Y23																																	Y23 loci				ForenSeq loci												
ID	DYS28	DYS16	DYS13	DYS13	DYS33H	DYS33H	DYS33H	DYS33H	DYS30	DYS30	DYS31	DYS31	DYS30	DYS32	DYS42T	DYS42T	DYS43	DYS43	DYS49	DYS49	DYS29	DYS29	DYS35	DYS35	DYS44	DYS44	DYS46	DYS46	YGATAH4	YGATAH4	DYS38	DYS38	DYS30	DYS46	DYS2B1S1	DYS46	DYS2C1	DYS2C1	Haplotype	Prediction likelihood (%)	Fitness prediction								
115	17	17	14	14	13	13	30	30	23	23	11	11	11	11	14	14	10	10	11	11	20	20	26	26	12	12	13	13	18	18	21	21	9	9	11	11	13.18	13.18	12	20.2	14	37.37	10	13	12	31	J1a	Probability = 25.00%	Fitness=58.6 [1.00]
116	19	19	14	14	13	13	31	31	24	24	10	10	11	11	14	14	10	10	11	11	20	20	24	24	10	10	12	12	11	11	11	11	17.17	17.17	13	18	16	39.40	10	13	12	28	E1b1b	Probability = 99.88%	Fitness=77.1 [1.13]				
117	16	16	14	14	13	13	29	29	23	23	10	10	11	11	14	14	10	10	10	10	20	20	25	25	10	10	13	13	18	18	20	20	9	9	11	11	12.17	12.17	12	17.2	15	37.37	11	12	12	30	J1a	Probability = 93.23%	Fitness=64.54 [1.22]
118	18	18	14	14	13	13	29	29	23	23	11	11	11	11	14	14	10	10	11	11	20	20	25	25	11	11	12	12	18	18	21	21	9	9	12	12	13.19	13.19	12	18.2	14	37.37	12	14	13	30	J1a	Probability = 48.58%	Fitness=62.12 [1.18]
119	19	19	16	16	13	13	30	30	23	23	11	11	13	13	15	15	11	11	12	12	19	19	23	23	12	12	13	13	17	17	23	23	10	10	12	12	12.12	12.12	13	15	37.41	11	12	10	32	R1a	Probability = 100.00%	Fitness=60.0 [0.72]	
120	16	16	14	14	14	14	30	30	23	23	10	10	11	11	14	14	10	10	13	13	19	19	25	25	11	11	12	12	18	18	20	20	10	10	11	11	11.18	11.18	12	18.2	15	36.39	11	13	12	30	J1a	Probability = 99.53%	Fitness=54.04 [0.85]
121	18	18	13	13	13	13	30	30	23	23	10	10	11	11	15	15	9	9	12	12	20	20	21	21	11	11	12	12	16	16	21	21	10	10	12	12	13.19	13.19	12	15	40.41	10	12	11	30	J2	Probability = 94.87%	Fitness=46.86 [1.03]	
122	18	18	14	14	13	13	29	29	23	23	10	10	11	11	14	14	10	10	11	11	20	20	27	27	11	11	12	12	17	17	19	19	9	9	11	11	14.19	14.19	12	18.2	15	37.38	11	13	12	32	J1a	Probability = 99.99%	Fitness=71.22 [1.13]
125	17	17	14	14	13	13	30	30	23	23	11	11	11	11	14	14	10	10	11	11	20	20	26	26	11	11	13	13	18	18	21	21	9	9	11	11	13.19	13.19	12	19.2	14	37.38	10	13	13	31	J1a	Probability = 47.92%	Fitness=76.22 [1.31]
126	17	17	15	15	13	13	29	29	23	23	9	9	11	11	15	15	9	9	11	11	21	21	22	22	12	12	13	13	16	16	21	21	11	11	11	11	13.16	13.16	12	15	39.39	10	12	11	32	J2	Probability = 100.00%	Fitness=72.13 [1.48]	
127	18	18	16	16	13	13	30	30	25	25	10	10	11	11	14	14	11	11	10	10	19	19	21	21	12	12	12	12	17	17	23	23	11	11	12	12	11.14	11.14	13	16	36.38	11	12	10	30	R1a	Probability = 100.00%	Fitness=43.11 [1.02]	
128	18	18	14	14	13	13	30	30	24	24	11	11	11	11	14	14	10	10	11	11	20	20	26	26	12	12	13	13	18	18	21	21	9	9	11	11	13.19	13.19	12	18.2	14	36.37	10	13	12	30	J1a	Probability = 77.47%	Fitness=70.12 [1.26]
130	18	18	15	15	14	14	30	30	22	22	10	10	11	11	15	15	9	9	12	12	21	21	24	24	11	11	13	13	18	18	21	21	10	10	12	12	13.17	13.17	12	16	39.39	9	10	12	29	J2	Probability = 76.29%	Fitness=43.65 [0.81]	
131	17	17	15	15	13	13	29	29	23	23	10	10	11	11	14	14	10	10	11	11	21	21	25	25	12	12	13	13	16	16	21	21	12	12	11	11	12.17	12.17	13	16	39.39	10	11	12	32	J2	Probability = 99.00%	Fitness=72.13 [1.48]	
134	18	18	14	14	14	14	31	31	24	24	10	10	11	11	14	14	10	10	11	11	21	21	25	25	10	10	13	13	19	19	21	21	12	12	12	12	16.18	16.18	14	15	36.37	10	11	12	32	E1b1b	Probability = 62.69%	Fitness=95.81 [0.31]	
135	16	16	15	15	13	13	28	28	23	23	10	10	11	11	15	15	9	9	12	12	22	22	22	22	11	11	13	13	17	17	21	21	10	10	12	12	13.16	13.16	13	15	40.40	11	11	10	31	J2	Probability = 100.00%	Fitness=62.59 [1.09]	
139	19	19	14	14	13	13	30	30	23	23	10	10	11	11	14	14	10	10	12	12	20	20	22	22	12	12	13	13	20	20	20	20	12	12	12	12	16.18	16.18	13	15	36.38	10	13	11	31	E1b1b	Probability = 62.66%	Fitness=44.24 [1.11]	
142	18	18	14	14	13	13	30	30	23	23	12	12	11	11	14	14	10	11	11	20	20	26	26	12	12	12	12	18	18	21	21	9	9	11	11	13.18	13.18	12	19.2	14	36.37	10	14	12	30	J1a	Probability = 48.66%	Fitness=60.28 [1.09]	
144	17	17	14	14	13	13	30	30	23	23	10	10	11	11	15	15	9	9	12	12	21	21	22	22	12	12	13	13	17	17	21	21	11	11	12	12	13.15	13.15	12	14	38.40	11	11	12	32	J2	Probability = 70.57%	Fitness=39.63 [0.81]	
145	18	18	15	15	14	14	31	31	24	24	11	11	13	13	14	14	12	12	12	12	19	19	26	26	12	12	12	12	18	18	23	23	11	11	12	12	13.15	13.15	12	16	35.36	11	12	10	33	R1b	Probability = 0.26%	Fitness=25.65 [0.51]	
146	15	15	15	15	14	14	30	30	24	24	9	9	11	11	15	15	9	9	10	10	19	19	21	21	11	11	11	11	15	15	20	20	10	10	11	11	13.15	13.15	12	19	37.38, 39	10	12	11	30	J2	Probability = 0.13%	Fitness=22.12 [0.45]	
149	19	19	14	14	13	13	31	31	25	25	11	11	11	11	14	14	10	10	13	13	19	19	23	23	12	12	13	13	18	18	22	22	12	12	11	11	17.18	17.18	13	14	34.37	11	13	13	30	E1b1b	Probability = 65.95%	Fitness=33.42 [0.84]	
150	18	18	16	16	13	13	30	30	23	23	10	10	11	11	14	14	10	10	11	11	18	18	25	25	11	11	13	13	19	19	21	21	9	9	11	11	13.17	13.17	12	18.2	15	36.38	10	13	13	31	J1a	Probability = 93.42%	Fitness=62.19 [1.19]
151	19	19	14	14	14	14	30	30	24	24	10	10	11	11	15	15	9	9	12	12	21	21	23	23	11	11	10	10	21	21	20	20	10	10	12	12	15.19	15.19	12	16	38.41	11	11	11	31	J2	Probability = 0.14%	Fitness=20.99 [0.44]	
152	16	16	14	14	12	12	27	27	22	22	10	10	14	14	15	15	9	9	12	12	19	19	23	23	12	12	10	10	15	15	23	23	11	11	13	13	12.17	12.17	11	15	34.41	10	12	12	30	J2	Probability = 100.00%	Fitness=38.20 [0.79]	
154	18	18	15	15	13	13	27	27	25	25	10	10	11	11	13	13	13	13	20	20	28	28	11	11	11	11	11	11	16	16	21	21	10	10	10	10	10.12	10.12	13	15	34.37	11	13	12	31	O2a2 F52	Probability = 0.00%	Fitness=61.0 [0.26]	
155	17	17	14	14	14	14	31	31	24	24	9	9	11	11	14	14	10	10	12	12	20	20	25	25	11	11	12	12	18	18	22	22	13	13	12	12	16.17	16.17	13	15	37.38	12	13	13	30	E1b1b	Probability = 99.94%	Fitness=58.76 [1.36]	
156	17	17	14	14	13	13	30	30	23	23	10	10	11	11	14	14	10	10	11	11	20	20	25	25	11	11	12	12	14	14	21	21	10	10	11	11	12.17	12.17	13	14	37.38	11	12	10	31	G2	Probability = 99.99%	Fitness=63.40 [1.00]	
166	18	18	14	14	14	14	32	32	23	23	10	10	11	11	14	14	10	10	13	13	20	20	24	24	11	11	12	12	17	17	21	21	9	9	11	11	13.17	13.17	12	18.2	15	36.38	11	13	12	29	J1a	Probability = 14.20%	Fitness=50.27 [0.87]
167	17	17	14	14	13	13	30	30	23	23	11	11	11	11	14	14	10	10	11	11	20	20	26	26	12	12	13	13	18	18	21	21	9	9	11	11	13.19	13.19	12	18.2	14	37.37	10	13	13	30	J1a	Probability = 51.74%	Fitness=74.45 [1.34]
168	19	19	14	14	13	13	29	29	23	23	12	12	14	14	15	15	12	12	12	12	19	19	22	22	12	12	12	12	16	16	23	23	10	10	13	13	12.14	12.14	11	16	35.36	11	12	10	30	R1b	Probability = 99.63%	Fitness=25.03 [0.60]	
169	18	18	14	14	13	13	29	29	23	23	10	10	11	11	14	14	10	10	11	11	20	20	24	24	11	11	12	12	18	18	21	21	9	9	11	11	13.18	13.18	12	18.2	14	37.37	11	13	13	29	J1a	Probability = 72.52%	Fitness=74.00 [1.40]

[illegible]

19 loci shared between the ForeSeq and the Y23																																														Y23 loci				ForeSeq loci				Haplotype	Prediction likelihood (%)	Fitness prediction																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																								
ID	DYS25	DYS26	DYS27	DYS28	DYS29	DYS30	DYS31	DYS32	DYS33	DYS34	DYS35	DYS36	DYS37	DYS38	DYS39	DYS40	DYS41	DYS42	DYS43	DYS44	DYS45	DYS46	DYS47	DYS48	DYS49	DYS50	DYS51	DYS52	DYS53	DYS54	DYS55	DYS56	DYS57	DYS58	DYS59	DYS60	DYS61	DYS62	DYS63	DYS64	DYS65	DYS66	DYS67	DYS68	DYS69	DYS70	DYS71	DYS72	DYS73	DYS74	DYS75	DYS76	DYS77				DYS78	DYS79	DYS80	DYS81	DYS82	DYS83	DYS84	DYS85	DYS86	DYS87	DYS88	DYS89	DYS90	DYS91	DYS92	DYS93	DYS94	DYS95	DYS96	DYS97	DYS98	DYS99	DYS100	DYS101	DYS102	DYS103	DYS104	DYS105	DYS106	DYS107	DYS108	DYS109	DYS110	DYS111	DYS112	DYS113	DYS114	DYS115	DYS116	DYS117	DYS118	DYS119	DYS120	DYS121	DYS122	DYS123	DYS124	DYS125	DYS126	DYS127	DYS128	DYS129	DYS130	DYS131	DYS132	DYS133	DYS134	DYS135	DYS136	DYS137	DYS138	DYS139	DYS140	DYS141	DYS142	DYS143	DYS144	DYS145	DYS146	DYS147	DYS148	DYS149	DYS150	DYS151	DYS152	DYS153	DYS154	DYS155	DYS156	DYS157	DYS158	DYS159	DYS160	DYS161	DYS162	DYS163	DYS164	DYS165	DYS166	DYS167	DYS168	DYS169	DYS170	DYS171	DYS172	DYS173	DYS174	DYS175	DYS176	DYS177	DYS178	DYS179	DYS180	DYS181	DYS182	DYS183	DYS184	DYS185	DYS186	DYS187	DYS188	DYS189	DYS190	DYS191	DYS192	DYS193	DYS194	DYS195	DYS196	DYS197	DYS198	DYS199	DYS200	DYS201	DYS202	DYS203	DYS204	DYS205	DYS206	DYS207	DYS208	DYS209	DYS210	DYS211	DYS212	DYS213	DYS214	DYS215	DYS216	DYS217	DYS218	DYS219	DYS220	DYS221	DYS222	DYS223	DYS224	DYS225	DYS226	DYS227	DYS228	DYS229	DYS230	DYS231	DYS232	DYS233	DYS234	DYS235	DYS236	DYS237	DYS238	DYS239	DYS240	DYS241	DYS242	DYS243	DYS244	DYS245	DYS246	DYS247	DYS248	DYS249	DYS250	DYS251	DYS252	DYS253	DYS254	DYS255	DYS256	DYS257	DYS258	DYS259	DYS260	DYS261	DYS262	DYS263	DYS264	DYS265	DYS266	DYS267	DYS268	DYS269	DYS270	DYS271	DYS272	DYS273	DYS274	DYS275	DYS276	DYS277	DYS278	DYS279	DYS280	DYS281	DYS282	DYS283	DYS284	DYS285	DYS286	DYS287	DYS288	DYS289	DYS290	DYS291	DYS292	DYS293	DYS294	DYS295	DYS296	DYS297	DYS298	DYS299	DYS300	DYS301	DYS302	DYS303	DYS304	DYS305	DYS306	DYS307	DYS308	DYS309	DYS310	DYS311	DYS312	DYS313	DYS314	DYS315	DYS316	DYS317	DYS318	DYS319	DYS320	DYS321	DYS322	DYS323	DYS324	DYS325	DYS326	DYS327	DYS328	DYS329	DYS330	DYS331	DYS332	DYS333	DYS334	DYS335	DYS336	DYS337	DYS338	DYS339	DYS340	DYS341	DYS342	DYS343	DYS344	DYS345	DYS346	DYS347	DYS348	DYS349	DYS350	DYS351	DYS352	DYS353	DYS354	DYS355	DYS356	DYS357	DYS358	DYS359	DYS360	DYS361	DYS362	DYS363	DYS364	DYS365	DYS366	DYS367	DYS368	DYS369	DYS370	DYS371	DYS372	DYS373	DYS374	DYS375	DYS376	DYS377	DYS378	DYS379	DYS380	DYS381	DYS382	DYS383	DYS384	DYS385	DYS386	DYS387	DYS388	DYS389	DYS390	DYS391	DYS392	DYS393	DYS394	DYS395	DYS396	DYS397	DYS398	DYS399	DYS400	DYS401	DYS402	DYS403	DYS404	DYS405	DYS406	DYS407	DYS408	DYS409	DYS410	DYS411	DYS412	DYS413	DYS414	DYS415	DYS416	DYS417	DYS418	DYS419	DYS420	DYS421	DYS422	DYS423	DYS424	DYS425	DYS426	DYS427	DYS428	DYS429	DYS430	DYS431	DYS432	DYS433	DYS434	DYS435	DYS436	DYS437	DYS438	DYS439	DYS440	DYS441	DYS442	DYS443	DYS444	DYS445	DYS446	DYS447	DYS448	DYS449	DYS450	DYS451	DYS452	DYS453	DYS454	DYS455	DYS456	DYS457	DYS458	DYS459	DYS460	DYS461	DYS462	DYS463	DYS464	DYS465	DYS466	DYS467	DYS468	DYS469	DYS470	DYS471	DYS472	DYS473	DYS474	DYS475	DYS476	DYS477	DYS478	DYS479	DYS480	DYS481	DYS482	DYS483	DYS484	DYS485	DYS486	DYS487	DYS488	DYS489	DYS490	DYS491	DYS492	DYS493	DYS494	DYS495	DYS496	DYS497	DYS498	DYS499	DYS500	DYS501	DYS502	DYS503	DYS504	DYS505	DYS506	DYS507	DYS508	DYS509	DYS510	DYS511	DYS512	DYS513	DYS514	DYS515	DYS516	DYS517	DYS518	DYS519	DYS520	DYS521	DYS522	DYS523	DYS524	DYS525	DYS526	DYS527	DYS528	DYS529	DYS530	DYS531	DYS532	DYS533	DYS534	DYS535	DYS536	DYS537	DYS538	DYS539	DYS540	DYS541	DYS542	DYS543	DYS544	DYS545	DYS546	DYS547	DYS548	DYS549	DYS550	DYS551	DYS552	DYS553	DYS554	DYS555	DYS556	DYS557	DYS558	DYS559	DYS560	DYS561	DYS562	DYS563	DYS564	DYS565	DYS566	DYS567	DYS568	DYS569	DYS570	DYS571	DYS572	DYS573	DYS574	DYS575	DYS576	DYS577	DYS578	DYS579	DYS580	DYS581	DYS582	DYS583	DYS584	DYS585	DYS586	DYS587	DYS588	DYS589	DYS590	DYS591	DYS592	DYS593	DYS594	DYS595	DYS596	DYS597	DYS598	DYS599	DYS600	DYS601	DYS602	DYS603	DYS604	DYS605	DYS606	DYS607	DYS608	DYS609	DYS610	DYS611	DYS612	DYS613	DYS614	DYS615	DYS616	DYS617	DYS618	DYS619	DYS620	DYS621	DYS622	DYS623	DYS624	DYS625	DYS626	DYS627	DYS628	DYS629	DYS630	DYS631	DYS632	DYS633	DYS634	DYS635	DYS636	DYS637	DYS638	DYS639	DYS640	DYS641	DYS642	DYS643	DYS644	DYS645	DYS646	DYS647	DYS648	DYS649	DYS650	DYS651	DYS652	DYS653	DYS654	DYS655	DYS656	DYS657	DYS658	DYS659	DYS660	DYS661	DYS662	DYS663	DYS664	DYS665	DYS666	DYS667	DYS668	DYS669	DYS670	DYS671	DYS672	DYS673	DYS674	DYS675	DYS676	DYS677	DYS678	DYS679	DYS680	DYS681	DYS682	DYS683	DYS684	DYS685	DYS686	DYS687	DYS688	DYS689	DYS690	DYS691	DYS692	DYS693	DYS694	DYS695	DYS696	DYS697	DYS698	DYS699	DYS700	DYS701	DYS702	DYS703	DYS704	DYS705	DYS706	DYS707	DYS708	DYS709	DYS710	DYS711	DYS712	DYS713	DYS714	DYS715	DYS716	DYS717	DYS718	DYS719	DYS720	DYS721	DYS722	DYS723	DYS724	DYS725	DYS726	DYS727	DYS728	DYS729	DYS730	DYS731	DYS732	DYS733	DYS734	DYS735	DYS736	DYS737	DYS738	DYS739	DYS740	DYS741	DYS742	DYS743	DYS744	DYS745	DYS746	DYS747	DYS748	DYS749	DYS750	DYS751	DYS752	DYS753	DYS754	DYS755	DYS756	DYS757	DYS758	DYS759	DYS760	DYS761	DYS762	DYS763	DYS764	DYS765	DYS766	DYS767	DYS768	DYS769	DYS770	DYS771	DYS772	DYS773	DYS774	DYS775	DYS776	DYS777	DYS778	DYS779	DYS780	DYS781	DYS782	DYS783	DYS784	DYS785	DYS786	DYS787	DYS788	DYS789	DYS790	DYS791	DYS792	DYS793	DYS794	DYS795	DYS796	DYS797	DYS798	DYS799	DYS800	DYS801	DYS802	DYS803	DYS804	DYS805	DYS806	DYS807	DYS808	DYS809	DYS810	DYS811	DYS812	DYS813	DYS814	DYS815	DYS816	DYS817	DYS818	DYS819	DYS820	DYS821	DYS822	DYS823	DYS824	DYS825	DYS826	DYS827	DYS828	DYS829	DYS830	DYS831	DYS832	DYS833	DYS834	DYS835	DYS836	DYS837	DYS838	DYS839	DYS840	DYS841	DYS842	DYS843	DYS844	DYS845	DYS846	DYS847	DYS848	DYS849	DYS850	DYS851	DYS852	DYS853	DYS854	DYS855	DYS856	DYS857	DYS858	DYS859	DYS860	DYS861	DYS862	DYS863	DYS864	DYS865	DYS866	DYS867	DYS868	DYS869	DYS870	DYS871	DYS872	DYS873	DYS874	DYS875	DYS876	DYS877	DYS878	DYS879	DYS880	DYS881	DYS882	DYS883	DYS884	DYS885	DYS886	DYS887	DYS888	DYS889	DYS890	DYS891	DYS892	DYS893	DYS894	DYS895	DYS896	DYS897	DYS898	DYS899	DYS900	DYS901	DYS902	DYS903	DYS904	DYS905	DYS906	DYS907	DYS908	DYS909	DYS910	DYS911	DYS912	DYS913	DYS914	DYS915	DYS916	DYS917	DYS918	DYS919	DYS920	DYS921	DYS922	DYS923	DYS924	DYS925	DYS926	DYS927	DYS928	DYS929	DYS930	DYS931	DYS932	DYS933	DYS934	DYS935	DYS936	DYS937	DYS938	DYS939	DYS940	DYS941	DYS942	DYS943	DYS944	DYS945	DYS946	DYS947	DYS948	DYS949	DYS950	DYS951	DYS952	DYS953	DYS954	DYS955	DYS956	DYS957	DYS958	DYS959	DYS960	DYS961	DYS962	DYS963	DYS964	DYS965	DYS966	DYS967	DYS968	DYS969	DYS970	DYS971	DYS972	DYS973	DYS974	DYS975	DYS976	DYS977	DYS978	DYS979	DYS980	DYS981	DYS982	DYS983	DYS984	DYS985	DYS986	DYS987	DYS988	DYS989	DYS990	DYS991	DYS992	DYS993	DYS994	DYS995	DYS996	DYS997

ID	19 loci shared between the ForeSnp and the Y23																																								Y23 loci				ForeSnp loci				Haplotype		Prediction likelihood (%)		Fitness prediction	
	DYS576	DYS578	DYS518	DYS519	DYS398	DYS399	DYS390	DYS391	DYS392	DYS393	DYS407	DYS408	DYS409	DYS444	DYS445	DYS414	DYS415	DYS530	DYS533	DYS540	DYS544	DYS570	DYS575	DYS535	DYS536	DYS434	DYS444	YGATAH1	YGATAH4	DYS385	DYS386	DYS390	DYS456	DYS459	DYS387S1	DYS406	DYS408	DYS395	DYS522	DYS412														
302	19	19	14	14	13	13	30	30	23	23	11	11	11	11	14	14	10	10	11	11	19.2	19.2	26	26	11	11	13	13	18	18	21	21	9	9	11	11	13.20	13.20	12	19.2	14	36.37	11	13	30	J1a	Probability = 56.78%	Fitness=60.01 [1.08]						
303	18	18	14	14	12	12	28	28	23	23	10	10	10	10	15	15	9	9	11	11	19	19	21	21	12	12	12	12	19	19	21	21	9	9	11	11	12.19	12.19	13	15	12	38.40	11	13	33	J2	Probability = 94.98%	Fitness=61.05 [0.75]						
304	19	19	14	14	13	13	29	29	23	23	11	11	11	11	15	15	9	9	12	12	20	20	26	26	11	11	12	12	13	13	23	23	11	11	12	12	13.15	13.15	12	20	15	39.39	9	11	11	29	J2	Probability = 0.06%	Fitness=73.00 [0.41]					
305	18	18	15	15	12	12	28	28	23	23	11	11	11	11	16	16	10	10	11	11	21	21	23	23	10	10	13	13	20	20	21	21	11	11	12	12	13.15	13.15	14	17	14	36.36.3	10	11	12	30	G2	Probability = 78.54%	Fitness=39.06 [1.01]					
306	17	17	15	15	13	13	29	29	25	25	10	10	10	10	11	11	14	14	11	11	10	10	20	20	23	23	12	12	13	13	18	18	23	23	8	8	12	12	11.4	11.14	13	15	16	37.39	11	12	10	32	R1a	Probability = 100.00%	Fitness=44.77 [1.06]			
307	17	17	15	15	12	12	28	28	23	23	10	10	11	11	14	14	11	11	12	12	14	14	11	11	20	20	24	24	11	11	13	13	17	17	20	20	12	12	11	11	8.12	8.12	13	16	13	34.38.2	11	13	12	29	B	Probability = 100.00%	Fitness=69.04 [1.48]	
308	18	18	14	14	14	14	31	31	23	23	10	10	10	10	11	11	14	14	10	10	11	11	22	22	26	26	11	11	12	12	18	18	21	21	9	9	10	10	13.18	13.18	12	17.2	15	36.37	11	13	12	29	J1a	Probability = 100.00%	Fitness=64.87 [1.28]			
309	18	18	17	17	13	13	29	29	25	25	11	11	11	11	14	14	11	11	10	10	20	20	24	24	13	13	13	13	19	19	23	23	10	10	12	12	11.14	11.14	13	15	15	37.38	11	12	10	30	R1a	Probability = 100.00%	Fitness=46.81 [1.11]					
310	18	18	12	12	13	13	30	30	24	24	10	10	10	10	11	11	14	14	10	10	13	13	20	20	27	27	10	10	11	11	18	18	13	17	15	15	39.39	11	13	13	32	E1b1b	Probability = 86.31%	Fitness=33.91 [0.81]										
311	17	17	15	15	13	13	32	32	23	23	10	10	10	10	11	11	16	16	9	9	12	12	21	21	23	23	11	11	13	13	17	17	22	22	10	10	12	12	14.17	14.17	12	15	16	37.39	10	11	11	30	J2	Probability = 63.24%	Fitness=35.14 [0.84]			
312	18	18	14	14	13	13	29	29	23	23	11	11	11	11	14	14	10	10	11	11	20	20	23	23	11	11	13	13	17	17	21	21	9	9	11	11	13.18	13.18	12	18.2	14	38.38	11	13	12	31	J1a	Probability = 32.57%	Fitness=58.96 [1.12]					
313	17	17	15	15	12	12	28	28	23	23	10	10	11	11	14	14	11	11	12	12	20	20	24	24	11	11	13	13	17	17	20	20	12	12	11	11	8.12	8.12	13	16	13	34.38.2	11	13	12	29	B	Probability = 100.00%	Fitness=69.04 [1.48]					
314	17	17	15	15	14	14	30	30	23	23	11	11	13	13	14	14	9	9	12	12	19	19	23	23	12	12	12	12	17	17	21	21	10	10	11	11	14.17	14.17	13	19	16	37.37	10	10	12	30	T	Probability = 96.61%	Fitness=40.48 [0.85]					
315	18	18	14	14	13	13	29	29	24	24	10	10	13	13	15	15	12	12	13	13	19	19	24	24	12	12	13	13	18	18	24	24	11	11	12	12	12.15	12.15	12	16	16	35.36	11	13	11	30	R1b	Probability = 99.99%	Fitness=35.10 [0.84]					
316	16	16	16	16	13	13	29	29	25	25	11	11	11	11	14	14	11	11	10	10	21	21	24	24	12	12	12	12	20	20	23	23	10	10	12	12	11.12.14.11.12.14	13	15	16	36.37	12	12	10	31	R1a	Probability = 100.00%	Fitness=37.31 [0.89]						
317	18	18	14	14	9	9	25	25	23	23	11	11	12	12	16	16	10	10	11	11	21	21	24	24	11	11	12	12	17	17	20	20	11	11	11	11	14.15	14.15	14	15	17	37.40	9	12	11	31	G1	Probability = 100.00%	Fitness=52.00 [1.35]					
319	19	19	16	16	14	14	31	31	25	25	11	11	11	11	14	14	11	11	9	9	20	20	23	23	12	12	12	12	17	17	23	23	11	11	14	14	11.14	11.14	13	16	16	37.38	12	13	11	31	R1a	Probability = 100.00%	Fitness=37.37 [0.89]					
321	16	16	15	15	13	13	32	32	24	24	11	11	11	11	16	16	9	9	12	12	20	20	22	22	11	11	13	13	14	14	23	23	10	10	12	12	13.16	13.16	12	15	15	38.40	10	11	11	27	J2	Probability = 95.97%	Fitness=35.48 [0.84]					
325	19	19	15	15	13	13	29	29	23	23	10	10	10	10	11	11	15	15	10	10	11	11	21	21	21	21	10	10	13	13	17	17	21	21	11	11	12	12	13.16	13.16	13	16	15	38.38	10	11	11	31	J2	Probability = 4.50%	Fitness=30.86 [0.58]			
330	18	18	16	16	14	14	31	31	21	21	10	10	11	11	14	14	10	10	11	11	20	20	24	24	11	11	11	11	17	17	21	21	13	13	11	11	17.17	17.17	14	17	16	36.39	13	11	32	E1b1a	Probability = 100.00%	Fitness=45.79 [1.20]						
333	18	18	15	15	14	14	31	31	24	24	10	10	11	11	14	14	11	11	10	10	20	20	27	27	12	12	12	12	17	17	23	23	10	10	12	12	11.14	11.14	13	16	16	36.37.39	11	12	10	32	E1b1b	Probability = 100.00%	Fitness=48.18 [1.06]					
334	18	18	13	13	13	13	30	30	24	24	10	10	10	10	11	11	14	14	10	10	20	20	22	22	13	13	12	12	12	12	22	22	12	12	12	12	16.17	16.17	13	15	15	38.40	9	13	12	32	E1b1b	Probability = 99.92%	Fitness=49.62 [1.06]					
335	18	18	14	14	13	13	31	31	25	25	10	10	10	10	14	14	10	10	12	12	20	20	25	25	10	10	12	12	19	19	21	21	12	12	11	11	16.17	16.17	13	18	16	38.39	10	13	12	29	E1b1b	Probability = 100.00%	Fitness=61.78 [1.47]					
336	18	18	14	14	13	13	29	29	22	22	11	11	11	11	14	14	10	10	11	11	20	20	25	25	11	11	12	12	17	17	21	21	9	9	12	12	13.19	13.19	12	19.2	13	36.38	12	13	30	J1a	Probability = 93.39%	Fitness=62.20 [1.18]						
337	19	19	15	15	14	14	31	31	24	24	11	11	11	11	14	14	11	11	10	10	20	20	23	23	12	12	13	13	19	19	23	23	10	10	12	12	11.14	11.14	12	16	16	37.39	11	12	11	30	R1a	Probability = 100.00%	Fitness=42.76 [1.02]					
338	13	13	15	15	13	13	31	31	21	21	10	10	10	10	14	14	11	11	12	12	21	21	28	28	11	11	11	11	17	17	21	21	12	12	12	12	15.18	15.18	13	17	15	36.39	10	13	11	30	E1b1a	Probability = 93.18%	Fitness=33.70 [0.78]					
339	15	15	17	17	12	12	28	28	24	24	10	10	12	12	15	15	10	10	12	12	20	20	21	21	11	11	11	11	19	19	20	20	13	13	10	10	14.15	14.15	13	15	14	38.39	11	12	12	30	I2a2b	Probability = 1.46%	Fitness=22.64 [0.42]					
340	18	18	14	14	13	13	29	29	23	23	10	10	10	10	11	11	14	14	10	10	11	11	20	20	24	24	11	11	12	12	18	18	22	22	9	9	11	11	14.18	14.18	12	18.2	14	37.37	11	13	32	J1a	Probability = 33.65%	Fitness=59.54 [1.13]				
341	16	16	15	15	14	14	30	30	22	22	10	10	10	10	11	11	14	14	9	9	11	11	21	21	20	20	11	11	14	14	18	18	21	21	10	10	12	12	13.18	13.18	12	16	15	38.41	10	11	11	32	J2	Probability = 100.00%	Fitness=68.48 [1.22]			
342	19	19	14	14	13	13	31	31	23	23	10	10	10	10	11	11	14	14	10	10	11	11	20	20	24	24	11	11	12	12	16	16	20	20	9	9	11	11	13.17	13.17	13	18.2	15	36.37	10	13	14	30	J1a	Probability = 31.98%	Fitness=68.05 [1.29]			
343	16	16	15	15	12	12	28	28	22	22	10	10	10	10	11	11	16	16	10	10	11	11	21	21	21	21	9	9	12	12	17	17	21	21	12	12	11	11	12.15	12.15	13	19	15	38.38	10	11	11	30	G2	Probability = 99.73%	Fitness=53.85 [1.06]			
344	19	19	17	17	14	14	31	31																																														