GENETIC VARIATIONS OF NINE MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA USING AUTOSOMAL AND Y-CHROMOSOME SHORT TANDEM REPEATS (STRS) ANALYSIS

AEDRIANEE REEZA BINTI ALWI

UNIVERSITI SAINS MALAYSIA

2016

GENETIC VARIATIONS OF NINE MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA USING AUTOSOMAL AND Y-CHROMOSOME SHORT TANDEM REPEATS (STRS) ANALYSIS

by

AEDRIANEE REEZA BINTI ALWI

Thesis submitted in fulfillment of the requirements for the degree of Master of Science

September 2016

ACKNOWLEDGEMENT

In the name of Allah, the Most Beneficent, the Most Merciful.

First of all, I am grateful to Allah for establishing me to complete this long awaited thesis for the requirements for the degree of Master of Science.

My deepest gratitude goes to my main supervisor, Mr. S. Pannerchelvam for his full support, expert guidance, understanding and encouragement during my research and study. He has the incredible patience in me throughout the years in completing this thesis. I would also like to thank Associate Professor Dr Zafarina Zainuddin (Co Supervisor) for her valuable guidance and encouragement extended to me.

I also wish to express my sincere thanks to the Director General of Department of Chemistry Malaysia, Director of Forensic Division and Head of DNA Forensic laboratory for their permission and providing me with all the necessary facilities in the Department to pursue my research. Not forgetting the Vice Chancellor, Director of Institute of Graduate Studies, Dean of School of Health Sciences and the staff of Universiti Sains Malaysia for their support, help and cooperation throughout this academic exploration.

Many thanks go to my numerous friends and colleagues who helped and gave me their full cooperation in completing this thesis. Finally, I would like to thank my husband, daughter, parents and family for their unconditional love and support during my studies. I would not have been able to complete this thesis without their continuous love and encouragement.

TABLE OF CONTENTS

Acknowledgement………………………………………………………………... ii

Table of Contents……………………………………………………………… iii

List of Tables………………………………………………………………...... x

List of Figures………………………………………………………………. xvi

List of Symbols, Abbreviations and Acronymns …………………………….. xix

List of Publications…………………………………………………………. xxi

Abstrak………………………………………………………….............. xxii

Abstract………………………………………………………….............. xxiv

CHAPTER 1- INTRODUCTION…………………………………………… 1

1.1 Deoxyribonucleic Acid (DNA) profiling……………………………….. 1

1.2 Human Genome………………………………………………………...... 2

1.2.1 Nuclear DNA……………………………………………….... 3

1.2.2 Mitochondrial DNA………………………………………… 6

1.3 DNA Based Human Identification Systems……………………………... 8

1.3.1 Restriction Fragment Length Polymorphism (RFLP)……….. 8

1.3.2 Polymerase Chain Reaction (PCR) based typing system……. 9

1.3.2.1 HLA-DQα and PolyMarkers…………………….. 11

1.3.2.2 Short Tandem Repeats (STR)…………………… 12

1.3.2.2.1 Autosomal STR…………………….. 14

1.3.2.2.2 X-chromosome STR………………... 15

1.3.2.2.3 Y-chromosome STR………………... 18

1.4 The AmpF*l*STR® Identifiler® PCR Amplification Kit…………………. 22

1.4.1 D8S1179…………………………………………………... 25

1.4.2 D21S11……………………………………………………. 25

1.4.3 D7S820…………………………………………………… 25

1.4.4 CSF1PO………………………………………………….... 25

1.4.5 D3S1358…………………………………………………... 26

1.4.6 TH01……………………………………………………… 26

1.4.7 D13S317…………………………………………………... 26

1.4.8 D16S539…………………………………………………... 27

1.4.9 D2S1338…………………………………………………... 27

1.4.10 D19S443………………………………………………………... 27
1.4.11 vWA……………………………………………………………. 28
1.4.12 TPOX………………………………………………………….... 28
1.4.13 D18S51………………………………………………………… 28
1.4.14 D5S818………………………………………………………… 29
1.4.15 FGA…………………………………………………………….. 29
1.5 The AmpF*l*STR® Yfiler® PCR Amplification Kit……………………… 30
1.5.1 DYS456………………………………………………………… 33
1.5.2 DYS389I……………………………………………………….. 33
1.5.3 DYS390………………………………………………………… 33
1.5.4 DYS389II………………………………………………………. 34
1.5.5 DYS458………………………………………………………… 34
1.5.6 DYS19………………………………………………………….. 34
1.5.7 DYS385a/b…………………………………………………..... 34
1.5.8 DYS393………………………………………………………… 35
1.5.9 DYS391………………………………………………………… 35
1.5.10 DYS439………………………………………………………… 35
1.5.11 DYS635………………………………………………………… 36
1.5.12 DYS392………………………………………………………… 36
1.5.13 Y GATA H4…………………………………………………… 36
1.5.14 DYS437………………………………………………………… 36
1.5.15 DYS438………………………………………………………… 37
1.5.16 DYS448………………………………………………………… 37
1.6 Population Studies…………………………………………………...... 37
1.6.1 The Malaysian population…………………………………... 37
1.6.2 The history of Malay…….…………………………………... 38
1.6.3 The Malay sub-ethnic groups……………………………….. 41
1.6.4 Population genetics………………………………………….. 49
1.6.5 Population database…………………………………………. 51
1.7 Objectives ……………………………………………………………..... 52

CHAPTER 2- MATERIALS AND METHODS…………………………..... 53
2.1 Biological Samples…………………………………………………..... 53
2.2 Materials and Reagents………………………………………………... 56
2.2.1 FTA® Classic Card……………………………......................... 56
2.2.2 Chemicals………………………………………………………... 58
2.2.3 Reagents……………………………………………………….... 58
2.2.3.1 10 mM NaOH…………………………………….. 58
2.2.3.2 TE Buffer [(10mM Tris-HCl, pH 8.0), (0.1mM EDTA, pH 8.0)]………………………………….... 58
2.2.3.3 The Human Identification PCR Amplification Kit.. 61
2.2.3.3.1 The AmpF*l*STR® Identifiler® PCR Amplification Kit……………………. 61
2.2.3.3.2 The AmpF*l*STR® Yfiler® PCR Amplification Kit……………………. 61
2.2.3.4 The Applied Biosystems® DS-33 Matrix Standard Kit (Dye Set G5)……………………………….... 64
2.2.3.5 Reagents on the Applied Biosystems® 3130*xl* Genetic Analyzer…………………………………. 64
2.3 Instrumentation and Apparatus……………………………………….... 64
2.4 Methods…………………………………………………………......... 69
2.4.1 Purification of bloodstained specimens on the FTA® card…….... 69
2.4.2 Polymerase Chain Reaction (PCR) Amplification……………..... 70
2.4.2.1 AmpF*l*STR® Identifiler® PCR for the autosomal Short Tandem Repeats……………………........ 70
2.4.2.2 AmpF*l*STR® Yfiler® PCR for the Y-Chromosome Short Tandem Repeats…………….. 74
2.4.3 Electrophoresis on the Applied Biosystems® 3130*xl* Genetic Analyzer…………………………………………………………. 78
2.4.3.1 Spectral calibration…………………………….... 78
2.4.3.2 Analysis of the amplified products……………….. 78

2.4.3.2.1 Run module settings on *3130xl* data collection software………………….. 78
2.4.3.2.2 Performing a Fragment Analysis Run.. 79
2.4.3.2.3 Sample preparation………………….. 79

2.4.4 GeneMapper ® ID Software analysis…………………………… 82

2.5 Statistical Analysis Method…………………………………………… 83
2.5.1 Common statistical analysis method for Autosomal STR and Y-STR………………………………………………………………. 83
2.5.1.1 Allele Frequency……………………………….. 83
2.5.1.2 Gene Diversity (GD)…………………………… 84
2.5.1.3 Power of Discrimination (PD)……………………. 85
2.5.1.4 Polymorphism Information Content (PIC)……….. 86
2.5.1.5 Power of exclusion (PE)………………………….. 86
2.5.1.6 Analysis of molecular variance (AMOVA)……… 87
2.5.1.7 F-statistics; Population fixation index (F_{ST}), Inbreeding coefficient (F_{IS}) and Overall fixation index (F_{IT})……………………………………… 90
2.5.1.8 Phylogenetic tree………………………………. 92
2.5.2 Statistical Analysis method specific for Autosomal STR…….…. 93
2.5.2.1 Exact test of Hardy-Weinberg equilibrium (HWE). 93
2.5.2.2 Linkage disequilibrium (LD)……………………... 93
2.5.2.3 Principal component analysis (PCA)……………... 94
2.5.3 Statistical Analysis method specific for Y-STR……………….... 95
2.5.3.1 Haplotype frequency…………………………… 95
2.5.3.2 Haplotype diversity…………………………….. 96
2.5.3.3 Multidimensional Scaling (MDS) from the R_{ST} value………………………………………………. 96

CHAPTER 3- RESULTS & DISCUSSIONS……………………………….. 99
3.1 Statistical Analysis for Autosomal STR and Y-STR……….................... 99
3.1.1 Allele frequency for autosomal STR of the nine (9) Malay sub-ethnic groups and the combined Malay populations (CMP)…...... 99

3.1.1.1 Genetic Locus D8S1179………………………….. 99
3.1.1.2 Genetic Locus D21S11…………………………… 101
3.1.1.3 Genetic Locus D7S820…………………………… 101
3.1.1.4 Genetic Locus CSF1PO…………………………... 104
3.1.1.5 Genetic Locus D3S1358………………………….. 106
3.1.1.6 Genetic Locus TH01……………………………… 106
3.1.1.7 Genetic Locus D13S317………………………….. 106
3.1.1.8 Genetic Locus D16S539………………………….. 110
3.1.1.9 Genetic Locus D2S1338………………………….. 110
3.1.1.10 Genetic Locus D19S443………………………….. 113
3.1.1.11 Genetic Locus vWA……………………………… 113
3.1.1.12 Genetic Locus TPOX……………………………... 116
3.1.1.13 Genetic Locus D18S51…………………………… 116
3.1.1.14 Genetic Locus D5S818…………………………… 120
3.1.1.15 Genetic Locus FGA……………………………… 120

3.1.2 Allele frequency for Y-STR of the nine (9) Malay sub-ethnic groups and the combined Malay populations (CMP)…………… 123

3.1.2.1 Genetic Locus DYS456…………………………... 123
3.1.2.2 Genetic Locus DYS389I………………………….. 123
3.1.2.3 Genetic Locus DYS390…………………………... 126
3.1.2.4 Genetic Locus DYS389II………………………… 126
3.1.2.5 Genetic Locus DYS458…………………………... 127
3.1.2.6 Genetic Locus DYS19……………………………. 131
3.1.2.7 Genetic Locus DYS385a……………………….. 131
3.1.2.8 Genetic Locus DYS385b……………………….. 131
3.1.2.9 Genetic Locus DYS393…………………………... 136
3.1.2.10 Genetic Locus DYS391…………………………... 136
3.1.2.11 Genetic Locus DYS439…………………………... 136
3.1.2.12 Genetic Locus DYS635…………………………... 137
3.1.2.13 Genetic Locus DYS392…………………………... 137
3.1.2.14 Genetic Locus Y GATA H4……………………… 138
3.1.2.15 Genetic Locus DYS437…………………………... 143
3.1.2.16 Genetic Locus DYS438…………………………... 143

3.1.2.17 Genetic Locus DYS448………………………….. 143
3.1.3 Gene Diversity (GD)…………………………………………..... 147
3.1.4 Power of discrimination (PD)…………………………………... 151
3.1.5 Polymorphism Information Content (PIC)……………………..... 156
3.1.6 Power of exclusion (PE)……………………………………….... 160
3.1.7 Analysis of molecular variance (AMOVA) …………………...... 163
3.1.8 F-statistics; Population fixation index (F_{ST}), Inbreeding coefficient (F_{IS}) and Overall fixation index (F_{IT})………………... 165
3.1.9 Phylogenetic tree………………………………………………… 166
3.2 Statistical Analysis for Autosomal STR…………………………………. 173
3.2.1 Exact test of Hardy- Weinberg Equilibrium (HWE)………….…... 173
3.2.2 Linkage disequilibrium………………………………………..…..... 178
3.2.3 Principal component analysis (PCA)………………………….…..... 189
3.3 Statistical Analysis for Y-STR…………………………………………... 192
3.3.1 Haplotype frequency & Haplotype diversity…………………..…..... 192
3.3.2 Multi-dimensional Scaling (MDS) from R_{ST} value……………… 210

CHAPTER 4- SUMMARY AND CONCLUSION………………………… 217
REFERENCE……………………………………………………………… 219
APPENDICES……………………………………………………………... 235
Appendix A: Certificate of Ethical Approval……………………………….. 235
Appendix B: Consent form………………………………………………… 236
Appendix C: Results of 15 STR typing of nine Malay sub-ethnic groups……………………………………………………... 240
Appendix D: Results of 17 Y- STR haplotyping of nine Malay sub-ethnic groups……………………………………………………... 282
Appendix E: Allele frequencies of 15 STRS in the Acheh sub-ethnic groups……………………………………………………... 289
Appendix F: Allele frequencies of 15 STRS in the Champa sub-ethnic groups……….. 291
Appendix G: Allele frequencies of 15 STRS in the Rawa sub-ethnic groups……………………………………………………... 293
Appendix H: Allele frequencies of 15 STRS in the Kedah sub-ethnic groups……………………………………………………..... 295
Appendix I: Allele frequencies of 15 STRS in the Minangkabau sub-ethnic groups….. 297
Appendix J: Allele frequencies of 15 STRS in the Bugis sub-ethnic groups……….. 299

Appendix K: Allele frequencies of 15 STRS in the Kelantan sub-ethnic groups…… 301
Appendix L: Allele frequencies of 15 STRS in the Jawa sub-ethnic groups…… 303
Appendix M: Allele frequencies of 15 STRS in the Banjar sub-ethnic groups…… 305
Appendix N: Allele frequencies of 15 STRS in the nine Malay sub-ethnic groups (CMP)…… 307
Appendix O: Allele frequencies of 17 Y-STRS in the Acheh sub-ethnic groups…… 309
Appendix P: Allele frequencies of 17 Y-STRS in the Champa sub-ethnic groups…… 310
Appendix Q: Allele frequencies of 17 Y-STRS in the Rawa sub-ethnic groups…… 311
Appendix R: Allele frequencies of 17 Y-STRS in the Kedah sub-ethnic groups…… 312
Appendix S: Allele frequencies of 17 Y-STRS in the Minangkabau sub-ethnic groups…… 313
Appendix T: Allele frequencies of 17 Y-STRS in the Bugis sub-ethnic groups…… 314
Appendix U: Allele frequencies of 17 Y-STRS in the Kelantan sub-ethnic groups…… 315
Appendix V: Allele frequencies of 17 Y-STRS in the Jawa sub-ethnic groups…… 316
Appendix W: Allele frequencies of 17 Y-STRS in the Banjar sub-ethnic groups…… 317
Appendix X: Allele frequencies of 17 Y-STRS in the nine Malay sub-ethnic groups (CMP)…… 318

LIST OF TABLES

Page

Table 1.1 : The loci amplified, chromosomal locations with the corresponding fluorescent marker dyes, the alleles contained in the AmpF*l*STR® allelic ladder and the genotype of The AmpF*l*STR® Control DNA 9947A………………………… 23

Table 1.2 : The loci amplified, with the corresponding fluorescent marker dyes, the alleles contained in the AmpF*l*STR® allelic ladder and the haplotype of The AmpF*l*STR® Control DNA 007….. 31

Table 2.1 : The list and number of blood specimen samples from nine different Malay sub-ethnic groups………………………….. 54

Table 2.2 : List of chemicals……………………………………………. 60

Table 2.3 : The AmpF*l*STR® Identifiler® PCR Amplification Kit contents…………………………………………………….. 62

Table 2.4 : The AmpF*l*STR® Yfiler® PCR Amplification Kit contents…………………………………………………….. 63

Table 2.5 : The master mix cocktail of AmpF*l*STR® Identifiler® PCR………………………………………………………. 72

Table 2.6 : The amplification volume of AmpF*l*STR® Identifiler® PCR.. 72

Table 2.7 : The master mix cocktail of AmpF*l*STR® Yfiler® PCR………………………………………………………. 76

Table 2.8 : The amplification volume of AmpF*l*STR® Yfiler® PCR………………………………………………………. 76

Table 2.9 : The run module settings for AmpF*l*STR® Identifiler® and AmpF*l*STR® Yfiler® on the 3130*xl* genetic analyzer……….. 81

Table 3.1 : Allele frequencies of STR D8S1179 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population……………………………………….............. 100

Table 3.2 : Allele frequencies of STR D21S11 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population…………………………………………………... 102

Table 3.3 : Allele frequencies of STR D7S820 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population…………………………………………………….... 103

Table 3.4 : Allele frequencies of STR CSF1PO genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population……………………………………………........ 105

Table 3.5 : Allele frequencies of STR D3S1358 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 107

Table 3.6 : Allele frequencies of STR TH01 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 108

Table 3.7 : Allele frequencies of STR D13S317 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 109

Table 3.8 : Allele frequencies of STR D16S539 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 111

Table 3.9 : Allele frequencies of STR D2S1338 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 112

Table 3.10 : Allele frequencies of STR D19S443 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 114

Table 3.11 : Allele frequencies of STR vWA genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 115

Table 3.12 : Allele frequencies of STR TPOX genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 118

Table 3.13 : Allele frequencies of STR D18S51 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 119

Table 3.14 : Allele frequencies of STR D5S818 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………………… 121

Table 3.15 : Allele frequencies of STR FGA genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 122

Table 3.16 : Allele frequencies of Y-STR DYS456 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 125

Table 3.17 : Allele frequencies of Y-STR DY389I genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 125

Table 3.18 : Allele frequencies of Y-STR DY390 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 128

Table 3.19 : Allele frequencies of Y-STR DY389II genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 129

Table 3.20 : Allele frequencies of Y-STR DY458 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 130

Table 3.21 : Allele frequencies of Y-STR DYS19 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 133

Table 3.22 : Allele frequencies of Y-STR DYS385a genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 134

Table 3.23 : Allele frequencies of Y-STR DYS385b genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 135

Table 3.24 : Allele frequencies of Y-STR DYS393 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 139

Table 3.25 : Allele frequencies of Y-STR DYS391 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 139

Table 3.26 : Allele frequencies of Y-STR DYS439 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………… 140

Table 3.27 : Allele frequencies of Y-STR DYS635 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………… 141

Table 3.28 : Allele frequencies of Y-STR DYS392 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………… 142

Table 3.29 : Allele frequencies of Y-STR Y GATA H4 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population……………………………… 142

Table 3.30 : Allele frequencies of Y-STR DYS437 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………… 145

Table 3.31 : Allele frequencies of Y-STR DYS438 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………… 145

Table 3.32 : Allele frequencies of Y-STR DYS448 genetic loci for the nine (9) different Malay sub-ethnic groups and combined Malay population………………………………………… 146

Table 3.33 : Gene diversity values of STR genetic loci and average for the nine (9) different Malay sub-ethnic groups and combined Malay population……………………………… 149

Table 3.34 : Gene diversity values of Y-STR genetic loci and average for the nine (9) different Malay sub-ethnic groups and combined Malay population……………………………… 150

Table 3.35 : The power of discrimination (PD) of 15 autosomal STR loci in the nine (9) different Malay sub-ethnic groups…………... 154

Table 3.36 : The power of discrimination (PD) of 17 Y-STR loci in the nine (9) different Malay sub-ethnic groups…………………. 155

Table 3.37 : The Polymorphism Information Content (PIC) of 15 autosomal STR loci in the nine (9) different Malay sub-ethnic groups……………………………………………... 158

Table 3.38 : The Polymorphism Information Content (PIC) of 17 Y-STR loci in the nine (9) different Malay sub-ethnic groups……… 159

Table 3.39 : The power of exclusion (PE) of 15 autosomal STR loci in the nine (9) different Malay sub-ethnic groups……………... 162

Table 3.40 : The AMOVA analysis for the 15 autosomal STR loci in the nine (9) different Malay sub-ethnic groups…………………. 164

Table 3.41 : The AMOVA analysis for the 17 Y-STR loci in the nine (9) different Malay sub-ethnic groups………………………….. 164

Table 3.42 : The exact test of HWE values of the 15 Autosomal STR loci calculated for Acheh, Champa, Rawa and Kedah Malays……………………………………………………. 175

Table 3.43 : The exact test of HWE values of the 15 Autosomal STR loci calculated for Minangkabau, Bugis, Kelantan and Jawa Malays……………………………………………………. 176

Table 3.44 : The exact test of HWE values of the 15 Autosomal STR loci calculated for Banjar Malays………………………......... 177

Table 3.45 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Acheh Malays………………………. 180

Table 3.46 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Champa Malays…………………… 181

Table 3.47 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Rawa Malays………………………. 182

Table 3.48 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Kedah Malays……………………... 183

Table 3.49 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Minangkabau Malays………………. 184

Table 3.50 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Bugis Malays………………………. 185

Table 3.51 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Kelantan Malays……………………. 186

Table 3.52 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Jawa Malays………………………... 187

Table 3.53 : Table of significant linkage disequilibrium for all pairs of 15 autosomal STR loci in Banjar Malays……………………... 188

Table 3.54 : Haplotype frequency distributions of 17 Y-STR loci in Acheh Malay sub-ethnic group (*n=7*)………………………. 194

Table 3.55 : Haplotype frequency distributions of 17 Y-STR loci in Champa Malay sub-ethnic group (*n=13*)…………………… 195

Table 3.56 : Haplotype frequency distributions of 17 Y-STR loci in Rawa Malay sub-ethnic group (*n=11*)……………………… 196

.

Table 3.57	:	Haplotype frequency distributions of 17 Y-STR loci in Kedah Malay sub-ethnic group (*n=9*)……………………….	197
Table 3.58	:	Haplotype frequency distributions of 17 Y-STR loci in Minangkabau Malay sub-ethnic group (*n=23*)……………...	198
Table 3.59	:	Haplotype frequency distributions of 17 Y-STR loci in Bugis Malay sub-ethnic group (*n=15*)………………………	199
Table 3.60	:	Haplotype frequency distributions of 17 Y-STR loci in Kelantan Malay sub-ethnic group (*n=43*)…………………...	200
Table 3.61	:	Haplotype frequency distributions of 17 Y-STR loci in Jawa Malay sub-ethnic group (*n=14*)……………………………..	202
Table 3.62	:	Haplotype frequency distributions of 17 Y-STR loci in Banjar Malay sub-ethnic group (*n=18*)……………………...	203
Table 3.63	:	Haplotype frequency distributions of 17 Y-STR loci in the nine Malay sub-ethnic groups (*n=153*)……………………...	204
Table 3.64	:	Pairwise genetic distance based on R_{ST} value of the haplotypes observed in the nine (9) Malays sub-ethnic groups for the 17 Y-STR loci………………………………..	213

LIST OF FIGURES

			Page
Figure 1.1	:	The organization of human genome………………………	3
Figure 1.2	:	Structure of DNA strands…………………………………..	4
Figure 1.3	:	mtDNA schematic diagram……………………………...	7
Figure 1.4	:	The Ideogram of the X-chromosome describing the location of STR markers…………………..……………	17
Figure 1.5	:	A typical inheritance pattern for the Y chromosome………	20
Figure 1.6	:	The Y-STR positions along the Y-chromosome…………...	21
Figure 1.7	:	Representative electropherogram of the AmpF*l*STR® control DNA 9947A………………………………………..	24
Figure 1.8	:	Representative electropherogram of the AmpF*l*STR® Control DNA 007 in the AmpF*l*STR® Yfiler® kit………..	32
Figure 1.9	:	Map showing the state boundries in West and East Malaysia …………………………………………….…..	46
Figure 1.10	:	The Minangkabau people in their traditional costume……..	47
Figure 1.11	:	A Bugis lady of Malaysia…………………………………..	47
Figure 1.12	:	The Banjar ladies of Malaysia……………………………...	48
Figure 1.13	:	The ladies of Champa Malay……………………………...	48
Figure 2.1	:	The geographical location of Malay sub-ethnic groups in Peninsular Malaysia where the samples were collected……	55
Figure 2.2	:	The FTA® Classic Card (by Whatman™)……………….....	57
Figure 2.3	:	Harris Micro Punch™……………………………………...	57
Figure 2.4	:	Performance Optimized Polymer 4 (POP-4™ Polymer), Hi-Di™ Formamide, 10X Buffer with EDTA and GeneScan™ LIZ 500™………………………………….....	66
Figure 2.5	:	9700 GeneAmp® Thermalcycler……………………......	66
Figure 2.6	:	3130*xl* Genetic Analyzer…………………………………...	67

Figure 2.7	:	3130*xl* Capillary Array 36cm………………………………	67
Figure 2.8	:	MicroAmp® Optical 96-well Reaction Plate, 96-well plate septa and 96-well plate retainer…………………………....	68
Figure 2.9	:	The AmpF*l*STR® Identifiler® PCR Amplification Kit……	71
Figure 2.10	:	The amplification conditions of AmpF*l*STR® Identifiler® PCR…………………………………………………......	73
Figure 2.11	:	The AmpF*l*STR® Yfiler® PCR Amplification Kit………..	75
Figure 2.12	:	The amplification conditions of AmpF*l*STR® Yfiler® PCR……………………………………………………...	77
Figure 2.13	:	The schematic diagram of the methodology for the autosomal STR and Y-STR analysis…..…………………...	98
Figure 3.1	:	Neighbor-Joining Phylogram among the the nine (9) sub-ethnic groups constructed based on genetic distances calculated using the allele frequencies of the 15 Autosomal STR genetic loci…………………………………………....	169
Figure 3.2	:	Neighbor-Joining Cladogram between the combined Malay Population (CMP=Malay) with other populations constructed based on genetic distances calculated using the allele frequencies of the 15 Autosomal STR genetic loci.....	170
Figure 3.3	:	Neighbor-Joining Cladogram between the combined Malay Population (CMP=Malay), nine (9) Malay sub-ethnic groups and with other populations constructed based on genetic distances calculated using the allele frequencies of the 15 Autosomal STR genetic loci…………………….…...	171
Figure 3.4	:	Neighbor-Joining Phylogram among the the nine (9) sub-ethnic groups constructed based on genetic distances calculated using the allele frequencies of the 17 Y-STR genetic loci…………………………………………………	172
Figure 3.5	:	Principal Component Analysis (PCA) using the 15 autosomal STR loci allele frequencies for the nine (9) Malay sub-ethnic groups, CMP and other world populations..	191
Figure 3.6	:	MDS plot based on population pairwise R_{ST} value of the haplotypes observed in the nine (9) Malays sub-ethnic groups……………………………………………………..	214

Figure 3.7 : MDS plot based on population pairwise R_{ST} value of the haplotypes observed in the nine (9) Malays sub-ethnic groups and other populations…………………………….... 215

Figure 3.8 : MDS plot based on population pairwise R_{ST} value of the haplotypes observed in the Combined Malay Population and other populations…………………………………….... 216

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMNS

θ	Theta
°C	Degree celcius
μl	microliter
Å	Angstrom
AMOVA	Analysis of molecular variance
ATP	Adenosine Triphosphate
bp	base pair
CE	Capillary Electrophoresis
CMP	Combined Malay population
DNA	Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic acid
F_{IS}	Inbreeding coefficient
F_{IT}	Overall fixation index
F_{ST}	Population fixation index
g	gram
GD	Gene Diversity
HET	Heterozygosity
HLA	Human Leukocyte Antigen
HV	Hypervariable region
HWE	Hardy-Weinberg Equilibrium
KV	Kilovolt
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
MDS	Multidimensional Scalling
ml	mililter
mM	milimolar
mtDNA	Mitochondrial DNA
NaOH	Natrium hydroxide
ng	nanogram

NRC	National Research Council
PCA	Principal Component analysis
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Power of Exclusion
PIC	Polymorphism Information Content
rCRS	Revised Cambridge Reference Sequence
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
rpm	revolution per minute
sec	second
STR	Short Tandem Repeats
SWGDAM	Scientific Working Group on DNA Analysis Methods
TE	Tris-EDTA
UV	Ultra violet
VNTR	Variable Number Tandem Repeats
YHRD	Y-STR Haplotype Reference Database

LIST OF PUBLICATIONS

Aedrianee, R.A, Paneerchelvam, S. & Zafarina, Z. Genetic variations of nine Malay sub-ethnic groups in Peninsular Malaysia using autosomal and Y-chromosome short tandem repeats (STRS) analysis. Poster presentation at the 1st International Conference on Forensic Science, Forensic Medicine & Criminology held on 19th & 20th April 2016 at the Management & Science University, Shah Alam, Selangor.

VARIASI GENETIK BAGI SEMBILAN SUB ETNIK MELAYU DI SEMENANJUNG MALAYSIA MENGGUNAKAN ANALISIS SHORT TANDEM REPEATS (STRS) AUTOSOM DAN KROMOSOM Y

ABSTRAK

Populasi Melayu di Semenanjung Malaysia terdiri daripada pelbagai kumpulan sub-etnik disebabkan oleh proses immigrasi dan asimilasi daripada pelbagai etnik dan puak-puak di sekitar perairan di Asia Tenggara sejak dahulu kala. Di dalam kajian ini, sebanyak 290 sampel darah telah diambil daripada 9 kumpulan sub- etnik Melayu di Semenanjung Malaysia yang terdiri daripada Acheh, Champa, Rawa, Kedah, Minangkabau, Bugis, Jawa, Banjar dan Kelantan dan telah dipalitkan ke atas kad FTA. Variasi genetik pada 15 STRs autosom telah dijalankan. Sampel DNA lelaki yang diperolehi selanjutnya dianalisa pada 17 STRs kromosom Y. Alel yang diperolehi adalah polimorfik dan purata kepelbagaian gen (GD) bagi 15 STRs autosom di kalangan sembilan kumpulan sub- etnik Melayu adalah berjulat di antara 0.7681 hingga 0.8028, manakala bagi 17 STRs pada kromosom Y berjulat di antara 0.6144 hingga 0.7165. Nilai kuasa diskriminasi gabungan bagi 15 STRs autosom untuk Acheh, Champa, Rawa, Kedah, Minangkabau, Bugis , Kelantan, Jawa dan Banjar adalah masing-masing 1 dalam 4.809×10^{13}, 1 dalam 4.352×10^{14}, 1 dalam 4.871×10^{15}, 1 dalam 2.493×10^{16}, 1 dalam 8.439×10^{16}, 1 dalam 1.910×10^{15}, 1 dalam 1.877×10^{17}, 1 dalam 5.519×10^{15} dan 1 dalam 9.020×10^{16}. Nilai kandungan informasi polimorfisme (PIC) daripada 15 STRs austosom berjulat di antara 0.40 hingga 0.86 dan 0.10 hingga 0.87 bagi 17 Y-STRs. Tiada penyimpangan daripada keseimbangan Hardy- Weinberg (HWE) yang signifikan di kalangan sembilan

kumpulan sub-etnik Melayu. Bagi analisis STR pada kromosom Y, sebanyak 138 haplotip telah ditemui dan di antaranya 125 haplotip adalah unik. Kepelbagaian haplotip yang dikira bagi sembilan kumpulan sub-etnik Melayu adalah 0.99845, ini menunjukkan bahawa haplotip yang ditemui di dalam kajian ini mempunyai variasi dan polimorfik yang tinggi. Carta *Phylograms, Principal Coordinate Analysis* (PCA) dan *Multi-Dimentional Scalling* (MDS) yang dihasilkan juga menunjukkan kelompok yang sama bagi kesembilan kumpulan sub-etnik Melayu dan mempunyai hubungan genetik yang rapat dengan Indonesia, Thailand dan Filipina. Lokus STR autosom dan kromosom Y yang digunakan adalah sangat utuh dan sesuai untuk proses individualisasi dalam analisis forensik memandangkan ia mempunyai kuasa diskriminasi yang tinggi antara individu.

GENETIC VARIATIONS OF NINE MALAY SUB-ETHNIC GROUPS IN PENINSULAR MALAYSIA USING AUTOSOMAL AND Y-CHROMOSOME SHORT TANDEM REPEATS (STRS) ANALYSIS

ABSTRACT

The Malay population of Peninsular Malaysia consists of various sub-ethnic groups due to years of immigration and assimilation of various ethnicity and tribes within maritime of Southeast Asia. In this study, a total of 290 blood samples were collected from 9 Malay sub-ethnic groups in Peninsular Malaysia consisting of Acheh, Champa, Rawa, Kedah, Minangkabau, Bugis, Jawa, Banjar and Kelantan and spotted onto FTA® Classic Card. The genetic variations for 15 autosomal STRs were performed. The male DNA samples were further analyzed on the 17 Y-chromosome STRs. The alleles observed were polymorphic and the average gene diversity (GD) for the 15 autosomal STRs across the nine Malay sub-ethnic groups ranging from 0.7681 to 0.8028 where as the 17 Y-STRs gene diversity ranging from 0.6144 to 0.7165. The combined power of discrimination of the 15 autosomal STRs for the Acheh, Champa, Rawa, Kedah, Minangkabau, Bugis, Kelantan, Jawa and Banjar were 1 in 4.809 x 10^{13}, 1 in 4.352 x 10^{14}, 1 in 4.871 x 10^{15}, 1 in 2.493 x 10^{16}, 1 in 8.439 x 10^{16}, 1 in 1.910 x 10^{15}, 1 in 1.877 x 10^{17}, 1 in 5.519 x 10^{15} and 1 in 9.020x 10^{16}, respectively. The values of the polymorphic information content (PIC) of the 15 autosomal STRs ranging from 0.40 to 0.86 and 0.10 to 0.87 for the 17 Y-STRs. There is no significant departure from Hardy-Weinberg Equilibrium (HWE) among the nine Malay sub-ethnic groups. For the Y-STRs analysis, 138 haplotypes were observed and among them 125 haplotypes were unique. The haplotype diversity

calculated for the nine Malays sub-ethnic groups was 0.99845, thus indicating that the haplotypes found in this study is highly variable and polymorphic. The phylograms, Principal Component Analysis (PCA) and Multi- Dimensional Scaling (MDS) plots generated also indicated similar population cluster of the nine Malay sub-ethnic groups and indicated a close genetic relationship with the Indonesia, Thailand and Philippines. The autosomal STR and Y-STR markers used are very robust and suitable for individualization in forensic case work analysis since it has high discrimination power to discriminate between individuals.

CHAPTER 1

INTRODUCTION

1.1 Deoxyribonucleic Acid (DNA) profiling

Human identification using specific regions of deoxyribonucleic acid (DNA) within the human genome has emerged as a powerful evidentiary tool in the criminal justice system. The well developed techniques enable individualization of human especially in forensic cases done in a fast and accurate manner.

DNA contains the blueprint of life called the genetic code that stores information necessary for passing down genetic attributes to future generations. Alec Jeffreys found that certain regions of DNA contained DNA sequences that were repeated over and over again next to each other and he described this discovery as DNA fingerprinting or DNA profiling as it is now known (Butler, 2010).

Based on this tremendous forensic DNA analysis technique, forensic scientists can use DNA obtained in blood, semen, saliva, skin or hair found at a crime scene to uniquely identify a matching DNA of an individual, let it be the victim or the perpetrator. This analysis also helps to prove guilt or innocence in the criminal law.

For the past three decades, forensic DNA analysis has evolved and many public and private laboratories all around the world provide this service to the law enforcement community. Growth in the usage of computer databases containing DNA profiles

from crime scene samples or convicted offenders were also observed within the years to solve criminal cases around the world.

A statistical evaluation of the evidence obtained from DNA analysis shall be calculated so that the significance of the match can be ascertained (Risch & Devlin, 1992). The examination of DNA can become compromised when environmental factors intervene, of when human error results in incorrect conclusions. These factors are crucial in determining what weight to give DNA evidence (Kobilinsky *et al.,* 2005).

1.2 Human Genome

The complete set of genetic information (DNA content) of a human diploid cell is called human genome. The total length of the human genome is over 3 billion base pairs. The human genome comprises two genomes; nuclear genome which accounts 99.9995% of the total genetic information and mitochondrial genome which accounts for 0.0005% (Figure 1.1). The nuclear genome provides the great bulk of essential genetic information encoded as DNA sequences with the 23 chromosome pairs in cell nuclei, where as only small DNA molecule found within individual mitochondria genome.

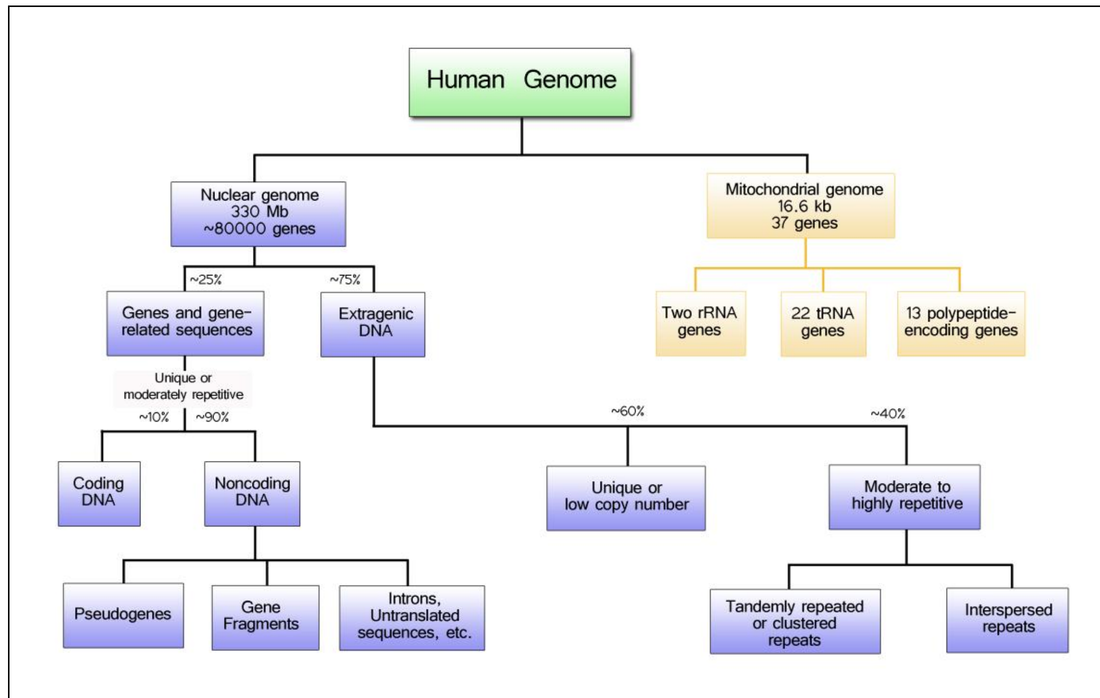

Figure 1.1: The organization of human genome

(Modified from: http://biology.kenyon.edu/courses/biol63/genome/genome.htm)

1.2.1 Nuclear DNA

Nuclear DNA is a nucleic acid that contained within a nucleus of eukaryotic organisms. It has a double helix structure with two strands wound around each other as first described by Francis Crick and James D. Watson in 1953 (Butler, 2010). DNA is composed of nucleotide made up of five-carbon sugar, phosphate group and an organic base. The nucleobase imparts the variation in each nucleotide unit, while the phosphate and sugar form the backbone structure of the DNA molecule. Nucleotides are distinguished by their bases consisting of the purines, large bases which include A (Adenine) and G (Guanine), and the pyrimidines, small bases which

include T (Thymine) and C (Cytosine) (Butler, 2010). The two DNA strands are linked together through a hybridization process where the nucleotides pair up with their complementary bases through hydrogen bonds (Figure 1.2).

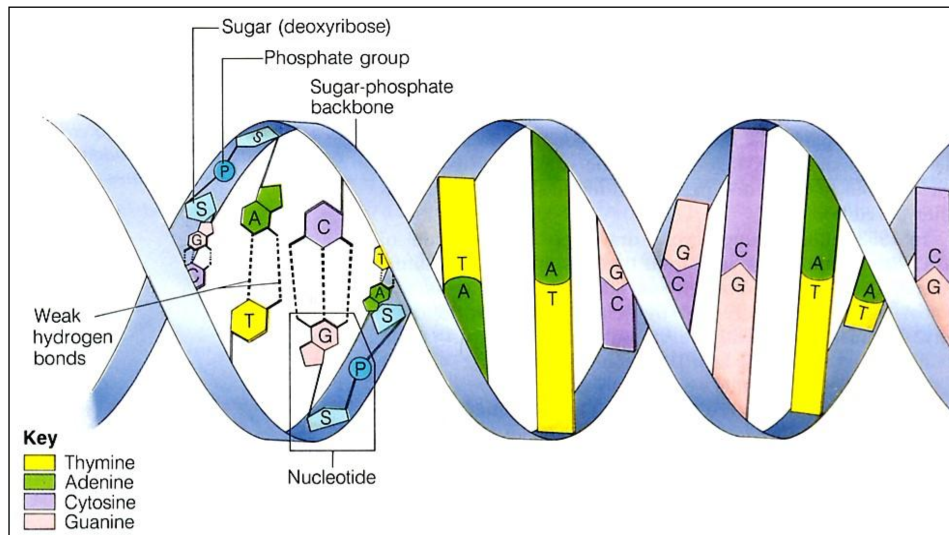

Figure 1.2: Structure of DNA strands

(Source: http://ehrig-privat.de/ueg/images/dna-structure.jpg)

The Nuclear DNA adheres to Mendelian inheritance, with information coming from two parents, one male and one female. DNA found in the nucleus of the cell is divided into chromosomes, which are dense packets of DNA and protection protein called histones (Butler, 2010). The nuclear DNA consists of 22 pairs of autosomal chromosomes and two sex-determining chromosomes. Chromosomes in all body (somatic) cells are in diploid state (contained two sets of each chromosome), unlike gametes (sperm or egg) are in haploid state (single set of chromosome).

The DNA material in chromosomes is composed of coding (~10%) and non-coding regions (~90%). The coding regions which contain information for a cell to make

proteins are known as genes which only make up ~5% of human genomic DNA. The regions which do not code for proteins have been known as 'junk' DNA. Surprisingly, these are the regions where the markers for human identity testing are found either between genes or within genes. Various polymorphic markers among individuals were found throughout the non-coding regions of the human genome. The chromosomal position or location of a gene or a DNA marker in this region is commonly referred to as locus.

The alternative possibilities of a gene or genetic locus are termed alleles. Differences in alleles at a particular locus are essential for human identification. Characterization of alleles present at a genetic locus is known as genotype, thus a DNA profile is a combination of genotypes obtained for multiple loci.

The genomes of any two people are more than 99% the same. Still, the tiny fraction of the genome that varies among humans is very important. DNA variations are part of what make each of us unique. The nuclear genome is full of repeated DNA sequences clusters which come in all types and length. These regions can be referred to as satellite DNA, mini-satellite DNA and micro-satellite DNA. Satellite DNA consisted of a long repeat unit (hundred to several thousand bases). The medium-length repeat referred to mini-satellite or VNTR (variable number tandem repeats), is ranging from 8 to 100 bases in length. DNA regions with short repeat units that are 2 to 7 bp in length are called microsatellite or better known as short tandem repeats (STR). The numbers of repeats in STR markers are highly variable among individuals thus making it robust for human identification purposes (Butler, 2006).

1.2.2 Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a circular DNA located in mitochondria. Mitochondria exist in every cell that converts chemical energy to adenosine triphosphate (ATP). Each cell may contain hundreds of mtDNA molecules that enable greater success with biological samples that may have been damaged or degraded.

Mitochondrial DNA has approximately 16,569 base pairs and posses 37 'genes' that code for products used in the oxidative phosphorylation process or cellular energy production (Anderson *et al*., 1981). Figure 1.3 is a schematic diagram showing the mtDNA genome that consisted of the transcribed 'genes', ribosomal ribonucleic acid (RNA) and transfer RNA.

mtDNA is maternally inherited unlike the nuclear DNA. Mitochondrial DNA in the spermatozoa is almost never permitted to enter the ovum during the fertilization process. It is passed along from mother to offspring (male/ female) and in turn from daughter to male or female offspring. mtDNA passes unchanged to offspring unless of spontaneous mutations. In human, it has been identified that at the D-loop region of the mtDNA consists of three hypervariable region 1 (HV1), hypervariable region 2 (HV2) and hypervariable region 3 (HV3) (Anderson *et al*., 1981; Haslindawaty *et al*., 2010), which exhibit multiple variations between individuals.

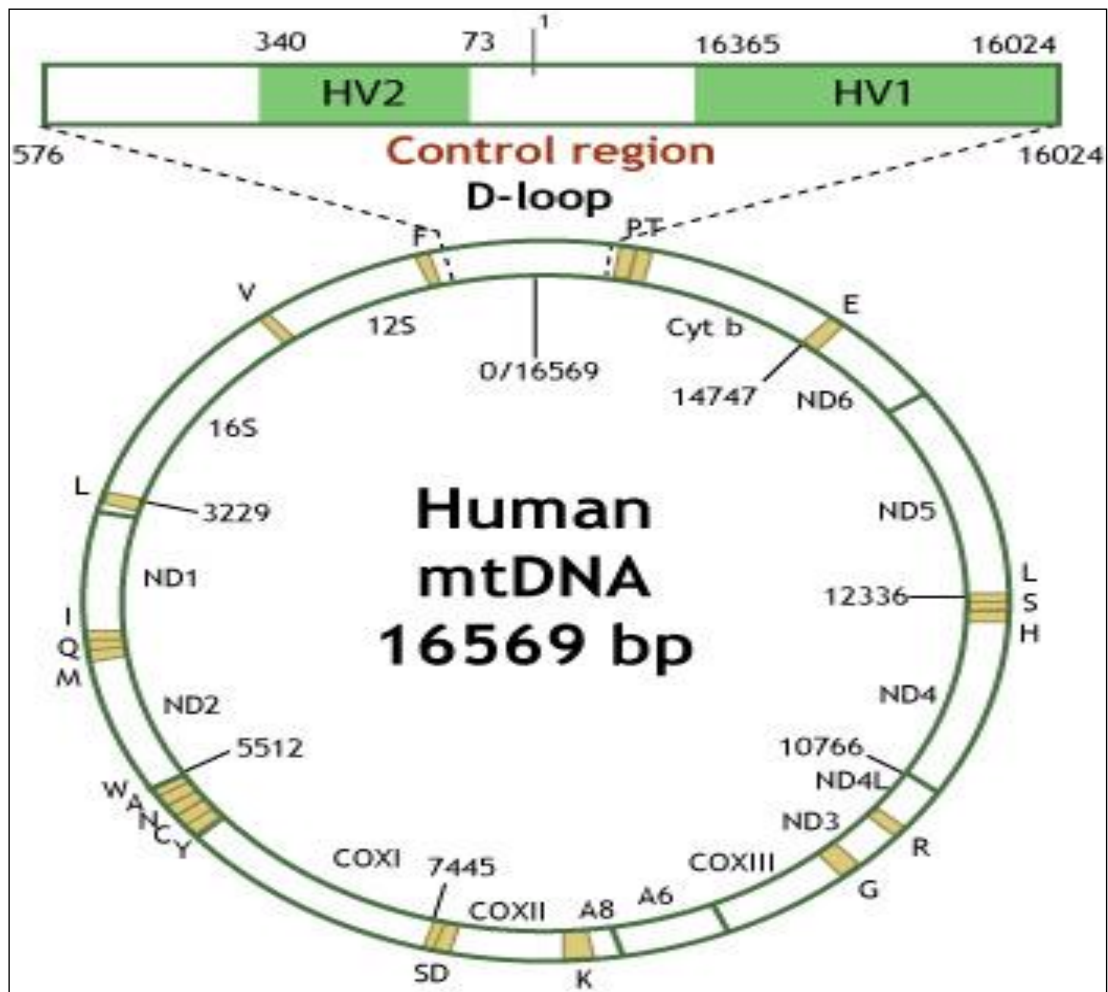

Figure 1.3: mtDNA schematic diagram

(Source: http://projects.nfstc.org/pdi/Subject09/pdi_s09_m02_01_a.htm)

Human mtDNA was first sequenced in 1981 (Anderson *et al.*, 1981). The results of mtDNA are commonly reported in comparison to the revised Cambridge reference sequence (rCRS). The chances of getting DNA typing result from mtDNA is higher than nuclear DNA, particularly in cases where the amount of extracted DNA is very small or samples are quite old and badly degraded, often bone, teeth and hair.

mtDNA is quite useful in forensic investigations of remains recovered of a missing person or mass disaster.

1.3 DNA Based Human Identification Systems

Alec Jeffreys discovered that certain regions of DNA contained DNA sequences that were repeated over and over again next to each other in tandem array while studying the gene coding for myoglobin (Jeffreys *et al.,* 1985). He found that the number repeated sections in a sample differ from individual to individual. He then created the technique to examine the length of these DNA repeat sequences to perform human identification tests. Technologies used for performing forensic DNA analysis to identify human varies depending on its ability to obtain results. History has marked that the human identification testing community has used various techniques for the past 20 years and now the evolution of the technologies being seen worldwide.

1.3.1 Restriction Fragment Length Polymorphism (RFLP)

RFLP was the first DNA fingerprinting technique used in human identification. It is a technique that treats DNA samples with restriction enzymes, which cleave DNA to produce fragments that contain a repeating segment. The number of repeats varies significantly between individuals. The chopped-up DNA fragments were then separated by fragment size on an agarose gel electrophoresis and transferred to a membrane via the southern blot procedure. The length of the fragments is determined based on the complementary hybridization of the membrane to a labeled DNA probe.

Each fragment length is considered as an allele and informative as to the genetic analysis (Kobilinsky *et al.,* 2005).

RFLP consisted of single locus probe (SLP) and multi-locus probe (MLP). The SLP only detects one or two alleles at a locus and it is more sensitive and easier to interpret in forensic samples, where there is possibility of mixture from multiple contributors (Kobilinsky *et al.,* 2005). Moreover, data can be obtained even when the DNA is degraded. This technique is rather slow in process and tedious. It requires large amount of DNA. RFLP cannot be applied on degraded DNA samples.

1.3.2 Polymerase Chain Reaction (PCR) based typing system

PCR is a technique that can amplify a specific DNA segment generating thousands to millions copies of DNA (Saiki *et al*, 1985; Mullis & Faloona, 1987). Polymerase is an enzyme responsible for DNA synthesis while "Chain Reaction" means exponential growth (at the rate of 2^n where n denotes the number of generation). Thus, PCR is an enzymatic technique used to amplify a specific DNA segment *in vitro* using two site specific primers that hybridise to complimentary DNA strands. It produces large amount of target DNA in a single enzymatic reaction within a matter of hours. This method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA (Rahman *et al,* 2013; Abalaka & Henry, 2011).

PCR is commonly carried out in a reaction volume 10-200µl in small reaction tubes in a thermal cycler. Thermal cycler heats and cools the reaction tubes to achieve the

temperature required at each step of the reaction. Typically, PCR consists of a series of 20-40 repeated temperatures changes, called cycles, with each cycle consisting of 2-3 discrete temperature steps. The PCR process consists of series of 20-35 cycles. Each consists of three steps:

a) Denaturation – unwind and separate the two strands of the double helix DNA at 94-98°C where the hydrogen bonds between complementary bases are disrupted.
b) Annealing – the reaction temperature is lowered to 50- 65 °C for 20-40 seconds allowing the primers to anneal at the single stranded DNA template.
c) Extension – the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction. Commonly temperature of 72 °C is used.

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of specific region of DNA. PCR is important for genetic fingerprinting and paternity test as this technique can identify a person through different PCR-based methods. Some PCR fingerprint methods have high discriminative power and can be used to identify genetic relationships between individuals. PCR also can be used to analyze critical small amounts of sample in forensic analysis where trace amounts of DNA that may be present in blood, hair or tissues can be copied so that there is enough DNA for analysis (Bartlet & Stirling, 2003).

1.3.2.1 HLA-DQα and PolyMarkers

The first PCR based commercial DNA typing kit used in forensic analysis was the *AmpliType HLA-DQα* kit which detected sequence variation at the Human Leukocyte Antigen (HLA) DQA1 gene found on chromosome 6 (Kobilinsky *et al.*, 2005). The loci involved in this system are all sequence polymorphisms that are detected by hybridization to allele-specific oligonucleotide (ASO) probes bound at specific locations on a test strip composed of nylon membrane.

The first DQα kit could only distinguish 6 different alleles and only possible to define 21 genotypes. The discriminatory power of this system is much more limited than RFLP analysis (approximately 1:2000) because only a limited number of alleles (and therefore, a limited number of genotypes) exist in the population. Nevertheless, the system is useful when evidence samples yield limited amounts of DNA and/or DNA that is too degraded for RFLP analysis.

Recognizing the need for a greater power of discrimination, the AmpliType pm +DQA1 kit was developed by Perkin-Elmer/ Roche (Kobilinsky *et al.*, 2005). Referred as PolyMarker, this kit coamplified a portion of the HLA DQα gene along with five other DNA segments. The AmpliType pm +DQA1 system tests for polymorphisms at 6 loci namely, Low Density Lipoprotein Receptor (LDLR), Glycophorin A (GLYPA), Hemoglobin G Gamma Globulin (HBGG), D7S8, Group-Specific Component (GC) and HLA-DQA1 (Kobilinsky *et al.*, 2005). However, the

power of discrimination was only in the range of 1 in 10,000 unrelated individuals and still they were insufficient to differentiate people on larger scale.

1.3.2.2 Short Tandem Repeats (STR)

Microsatellites or better known today as short tandem repeats (STR) were discovered in the late 1980s. STR based DNA profiling technique has become the method of choice in forensic laboratories since it has a greater capabilities than RFLP, HLA-DQA1 and AmpliType PM to analyze DNA. The technique uses PCR to amplify small region of genomic DNA, known as STR loci. These STRs have shorter repeat units in the range of 2 to 7 bp compared to the approximately 8 to 100 bp repeat units found in minisatellite genetic markers (Butler, 2010). Thousands of polymorphic microsatellites have been characterized in human genome and there may be more than a million microsatellite loci present depending on how they are counted (Kobilinsky *et al*., 2005). STR markers are scattered throughout the genome and occur on average every 10,000 nucleotides (Butler, 2010).

STRs have become popular DNA markers because they are easily amplified by the PCR. The number of repeats in STR markers can be highly variable among individuals, which make these STRs effective for human identification purposes (Butler, 2005). A number of STR loci had been discovered from genomic sequencing work and were being widely utilized for gene mapping studies.

STR repeat sequences are named by the length of the repeat unit. Two nucleotides repeats are called dinucleotide, three nucleotides in the repeat unit called trinucleotides, tetranucleotide have four, pentanucleotides have five and

hexanucleotides have six repeat units in the core repeat. For human identification, tetranucleotide repeats STRs have been chosen and validated. STRs are often divided into simple repeats which contain units of identical length and sequence; compound repeats comprise two or more adjacent simple repeats and complex repeats which may contain several repeat blocks variable unit length as well as variable intervening sequences (Kobilinsky *et al.*, 2005).

A number of multiplex systems have been designed to allow multiple STR loci to be amplified simultaneously. The loci chosen for the multiplexing also will have different size range so that one marker will not overlap another. The alleles at each locus must be measured in length after each multiplex. In most forensic DNA multiplex kits, STR locus generally differs by 4 nucleotide base pairs. However, not all alleles for STR locus contain complete repeat units. Microvariants are alleles that contain incomplete repeat units (Kobilinsky *et al.*, 2005).

For human identification, STR loci with a high degree of heterozygosity and a high level of polymorphism are preferred in order to obtain the ability to discriminate between individuals. Most forensically useful STR loci have fewer than 30 different alleles. The smaller the size of the STR loci makes them more suitable for amplification. PCR amplification of STRs can be used when there is only small amount of DNA available for testing or when the DNA obtained is degraded (Kobilinsky *et al.*, 2005; Butler, 2005).

1.3.2.2.1 Autosomal STR

Autosomal STR utilizes DNA from the 22 pairs of autosomal chromosomes. Autosomal DNA is inherited from both of our parents, thus STR DNA profiling may be taken by either a male or a female. STR markers on the autosomal chromosomes have been used in population genetics studies and for extended forensic and sibship studies. In the past two decades, a number of autosomal STR markers have been used in forensic DNA typing. They are chosen from separate chromosome or are widely spaced on the same chromosome to avoid any problems with linkage between the markers (Butler & Hill, 2012).

A common set of standardized markers must be used to make DNA typing to be effective. In 1997, the Federal Bureau of Investigation (FBI) Laboratory has chosen 13 STR core loci to be the basis of the Combined DNA Index System (CODIS) in United States. They are CSF1PO, FGA, THO1, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S211. The European also has their own standard set called European Standard Set (ESS) with 7 STRs core loci. The European Network of Forensic Science Institute (ENFSI) however in 2009 voted to adopt another five additional STR loci to the ESS. In April 2011, the FBI laboratory proposed and expanded set of core STR loci for the United States in order to reduce the likelihood of adventitious matches as the number of profiles stored in the U.S national DNA database increases, increase international compatibility to

assist law enforcement data-sharing efforts and increase the discrimination power to assist missing-persons cases (Kobilinsky *et al*., 2005).

Forensic scientists often purchased quality-controlled validated commercial kits for DNA typing work. These commercial kits come with allelic ladders, positive control and premixed reagents including the fluorescently labeled oligonucleotides (primers) that target the specific locations in the human genome to be PCR amplified. Many of these kits enable simultaneous, multicolour flourescene detctection of 15 autosomal STRs and the sex-typing marker amelogenin in a single reaction. For example, both Identifiler® and PowerPlex® 16 kits enable typing of the U.S core 13 STRs plus two additional loci. The sex determination amelogenin gene is often included in this commercial kit to provide information regarding gender. Amelogenin gene is located on both X and Y chromosome. Amelogenin is not considered as a STR locus.

1.3.2.2.2 X-chromosome STR

The analysis of STR markers on the X-chromosome has been used in population genetic and extended forensic and sibling ship studies. The X-STR markers are all located on the same chromosome unlike the autosomal STR markers. These X-STRs also have been proven to be useful in cases of deficiency paternity testing and in effective mother-son kinship and father-daughter testing (Barbaro & Cormaci, 2006). Males will inherit exclusively the X-chromosome from the mother and females would have two X-chromosome that they inherit from both parents.

X-STR data worldwide is rising slowly as only few scientists are dealing with this area. Commercial kits available for x-linked STR markers (Figure 1.4) rise in the year 2003 with Mentype® Argus X-UL PCR Amplification Kit from Biotype AG.

The kit amplifies the four X-chromosomal STR markers DXS7132, DXS7423, DXS8378 and HPRTB as well as amelogenin for gender determination simultaneously. In the mid 2005, Mentype® Argus X-8 PCR Amplification Kit was available and able to analyze eight X-chromosomal STR markers. The latest x-chromosome kit available is the Investigator Argus X-12 PCR Amplification Kit by Qiagen which amplifies simultaneously the 12 markers DXS7132, DXS7423, DXS8378, DXS10074, DXS10079, DXS10101, DXS10103, DXS10134, DXS10135, DXS10146, DXS10148 and HPRTB. A polymerase chain reaction in house multiplex called X-Decaplex was optimized in order to type ten X-chromosome short tandem repeats (STRs) in a single reaction, including: DXS8378, DXS9902, DXS7132, DXS9898, DXS6809, DXS6789, DXS7133, GATA172D05, GATA31E08, and DXS7423 by a collaborative work carried out by the Spanish and Portuguese ISFG Working Group (GEP-ISFG) (Gusmão *et.al*, 2009).

Over the last decade, a number of publications (Barbaro & Cormaci, 2006; Tillmar *et al.* 2008; Szibor *et al.*, 2003) have described the typing of X-chromosome markers and also the application of these markers in forensic casework especially in the DNA analysis of multiple females under the hypothesis that they share the same father. This analysis of X-STR can complement the analysis of autosomal markers for solving complex kinship cases as well as human identification.

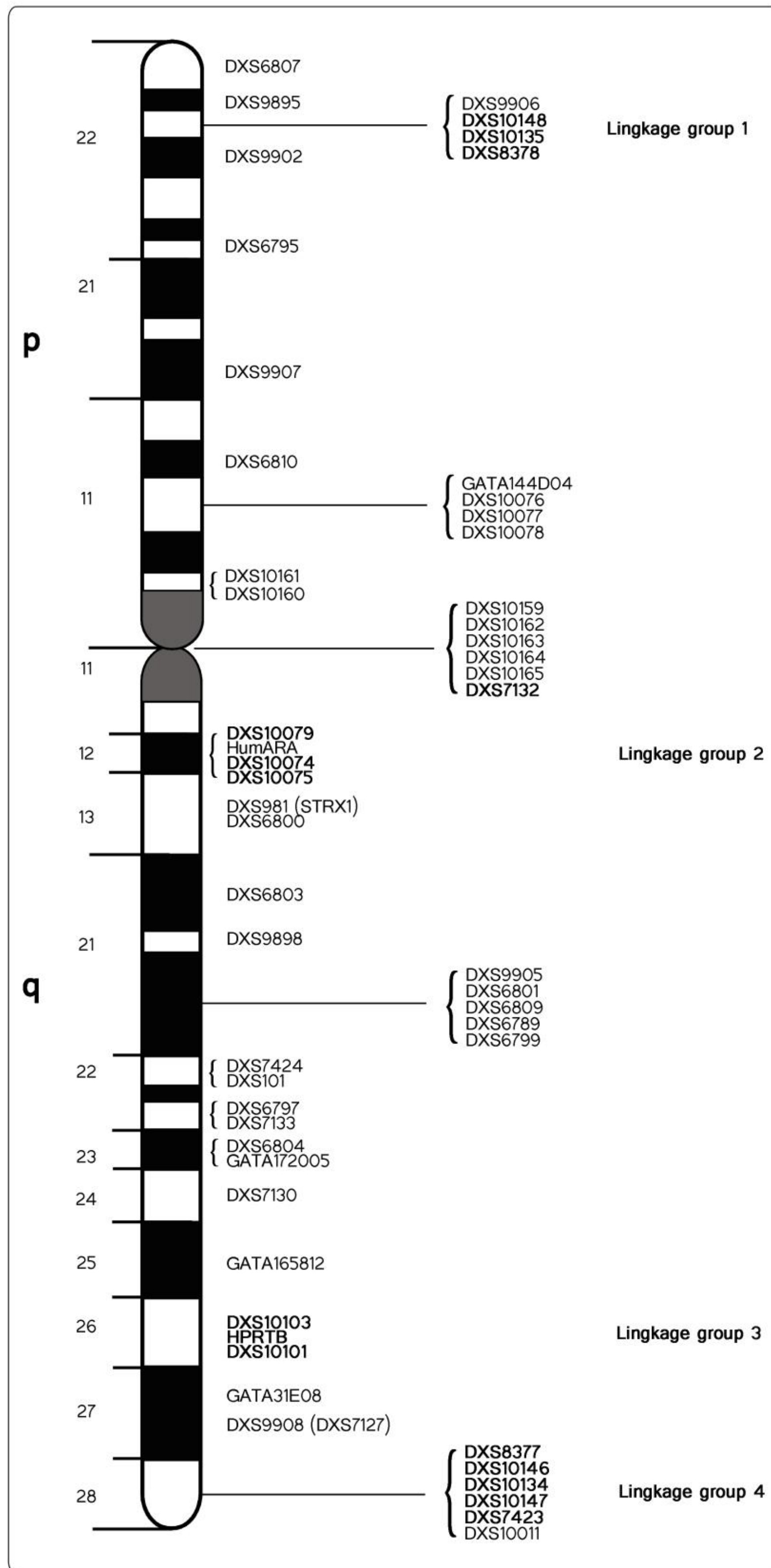

Figure 1.4: The Ideogram of the X-chromosome describing the location of STR markers.
(Modified from: http://www.chrx-str.org/)

1.3.2.2.3 Y-chromosome STR

The Y-STR is the short tandem repeats that occur on the Y chromosome. The value of the Y chromosome in forensic DNA testing is that it is found only in males (Butler, 2005). Unlike the autosomal DNA markers which are subjected to recombination with each generation because half of an individual's genetic information comes from his/her father and half from his/ her mother. The Y chromosome is passed down from generation to generation without changing (except for mutation). The paternal lineages with Y chromosome markers are shown in Figure 1.5 where it describes how the Y chromosome is inherited. The Y chromosome contains 59 million base pairs and only 161 coding genes which are vital for sex determination, spermatogenesis and some other male-related functions. With lineage markers, the genetic information obtained from Y chromosome STR is referred as haplotype rather than genotype because the STRs are inherited together as a single unit (Liu *et al.*, 2007).

Y-chromosome STRs have been used widely in forensic casework especially in sexual assault cases, verification of amelogenin Y deficient in males, paternity testing, missing person investigations and human migration and evolutionary studies (Gusmão *et al.*, 2006; Zhang *et al.*, 2014). Most of the non coding Y-STRs loci are located on the long arm (*q*) while the rest are found on the short arm (*p*) (Figure 1.6). These Y-STRs are polymorphic although less than the autosomal STRs.

The Y-STR markers used in human identity testing has increased tremendously since 1990. With the first discovery of DYS19 locus, more than 44 Y-STR loci have been identified (Roewer *et.al*, 1992). Since the Y-STR loci are all located on the same chromosome, they are said to be linked. The Y-STR markers also being developed in commercial kits with minimal haplotype sets by the European forensic community consisting of core Y-STR loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS385a/b. The YCAIIa/b loci were added as optional markers to create and extended haplotype. In early 2003, the Scientific Working Group on DNA Analysis Methods (SWGDAM) selected a core set of markers that includes the 9 markers in the minimal haplotype plus DYS438 and DYS439 (Butler, 2003). The available commercial Y-STR kits are mostly PowerPlex® Y by Promega Corporation and AmpF*l*STR® Yfiler™ by Applied Biosystems.

The statistical analysis of Y-STR evidence sample has to be determined to illustrate how common or how rare the profile is in the relevant population. The largest database of Y chromosome STR haplotypes in different populations was created by Lutz Roewer and a colleague at Humbolt University of Berlin, Germany and it is available online via https://yhrd.org. As of July 2015, the Y-STR Haplotype Reference Database (YHRD) comprised of 154,329 haplotypes with 991 populations in 129 countries is represented (Willuweit & Roewer, 2015).

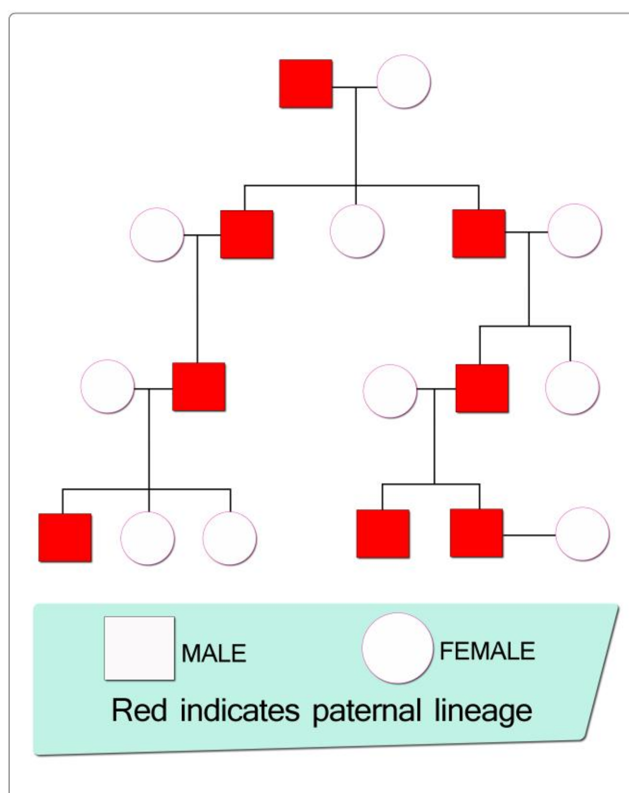

Figure 1.5: A typical inheritance pattern for the Y chromosome.

(Modified from: http://www.dnacenter.com/dna-testing/paternal-lineage.html)

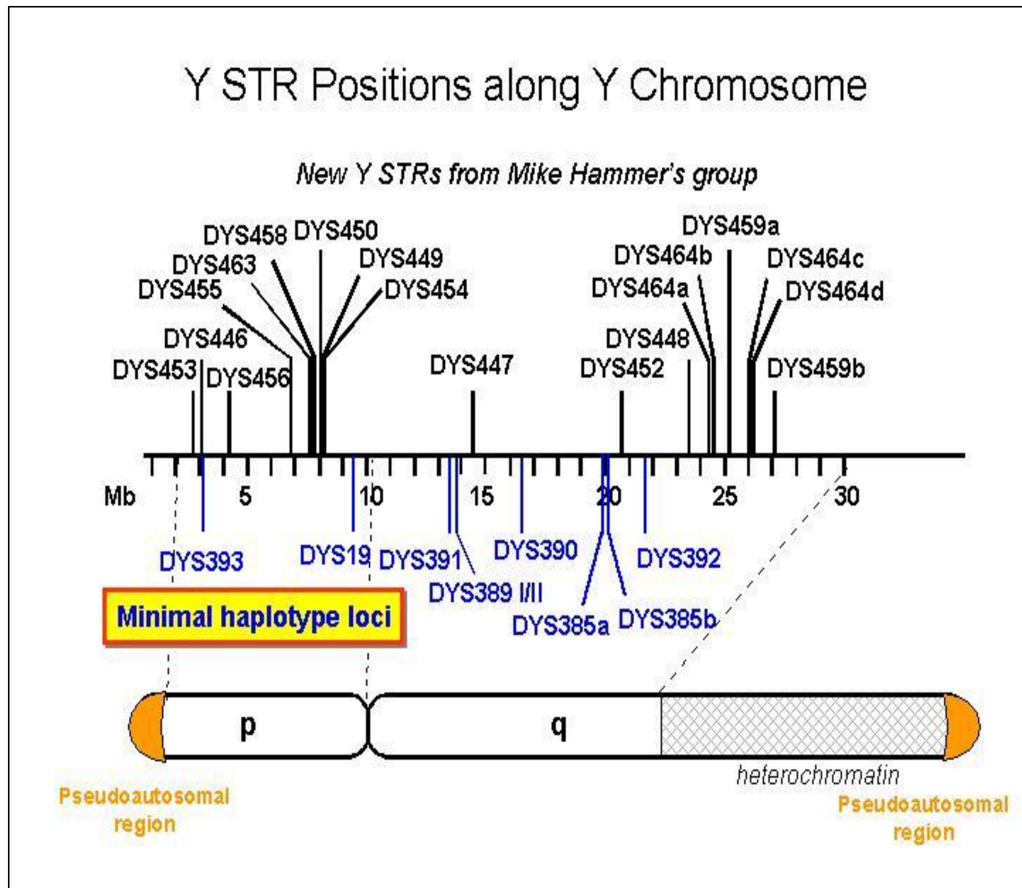

Figure 1.6: The Y-STR positions along the Y-chromosome

(Source: http://www.cstl.nist.gov/strbase/ystrpos1.htm)

1.4 The AmpF*l*STR® Identifiler® PCR Amplification Kit

The AmpF*l*STR® Identifiler® PCR Amplification Kit is a widely used STR based kit for human identification (Applied Biosystems, 2006). The kit amplifies 15 loci and amelogenin in a single tube and provides loci consistent with major world-wide STR databasing standards. It also includes the 13 core loci required for sample entry in CODIS (Budowle *et al.*, 1999) and two additional loci D2S1338 and D19S443.

The AmpF*l*STR® Identifiler® PCR Amplification Kit uses the combination of a five-dye fluorescent system and the inclusion of non-nucleotide linkers which allows for simultaneous amplification and efficient separation of the 15 autosomal STR loci and amelogenin during automated DNA fragment analysis. Table 1.1 shows the loci amplified, their chromosomal locations, fluorescent dyes, the alleles contained in the allelic ladder and the genotype of AmpF*l*STR® control DNA 9947A. Figure 1.7 demonstrate the electropherogram of the AmpF*l*STR® control DNA 9947A.

Table 1.1: The loci amplified, chromosomal locations with the corresponding fluorescent marker dyes, the alleles contained in the AmpF*l*STR® allelic ladder and the genotype of The AmpF*l*STR® Control DNA 9947A.

Locus Designation	Chromosome location	Alleles in the Allelic Ladder	Dye Label	Control DNA 9947A Genotype
D8S1179	8	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19	6-FAM™	13
D21S11	21q11.2-q21	24, 24.2, 25, 26, 27, 28, 28.2, 29, 29.2, 30, 30.2, 31, 31.2, 32, 32.2, 33, 33.2, 34, 34.2, 35, 35.2, 36, 37, 38		30
D7S820	7q11.21-22	6, 7, 8, 9, 10, 11, 12, 13, 14, 15		10,11
CSF1PO	5q33.3-34	6, 7, 8, 9, 10, 11, 12, 13, 14, 15		10,12
D3S1358	3p	12, 13, 14, 15, 16, 17, 18, 19	VIC®	14,15
TH01	11p15.5	4, 5, 6, 7, 8, 9, 9.3, 10, 11, 13.3		8,9.3
D13S317	13q22-31	8, 9, 10, 11, 12, 13, 14, 15		11
D16S539	16q24-qter	5, 8, 9, 10, 11, 12, 13, 14, 15		11,12
D2S1338	2q35-37.1	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28		19,23
D19S433	19q12-13.1	9, 10, 11, 12, 12.2, 13, 13.2, 14, 14.2, 15, 15.2, 16, 16.2, 17, 17.2	NED™	14,15
vWA	12p12-pter	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24		17,18
TPOX	2p23-2per	6, 7, 8, 9, 10, 11, 12, 13		8
D18S51	18q21.3	7, 9, 10, 10.2, 11, 12, 13, 13.2, 14, 14.2, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27		15,19
Amelogenin	X: p22.1-22.3 Y: p11.2	X, Y	PET®	XX
D5S818	5q21-31	7, 8, 9, 10, 11, 12, 13, 14, 15, 16		11
FGA	4q28	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 26.2, 27, 28, 29, 30, 30.2, 31.2, 32.2, 332, 42.2, 43.2, 44.2, 45.2, 46.2, 47.2, 50.2, 51.2		23,24

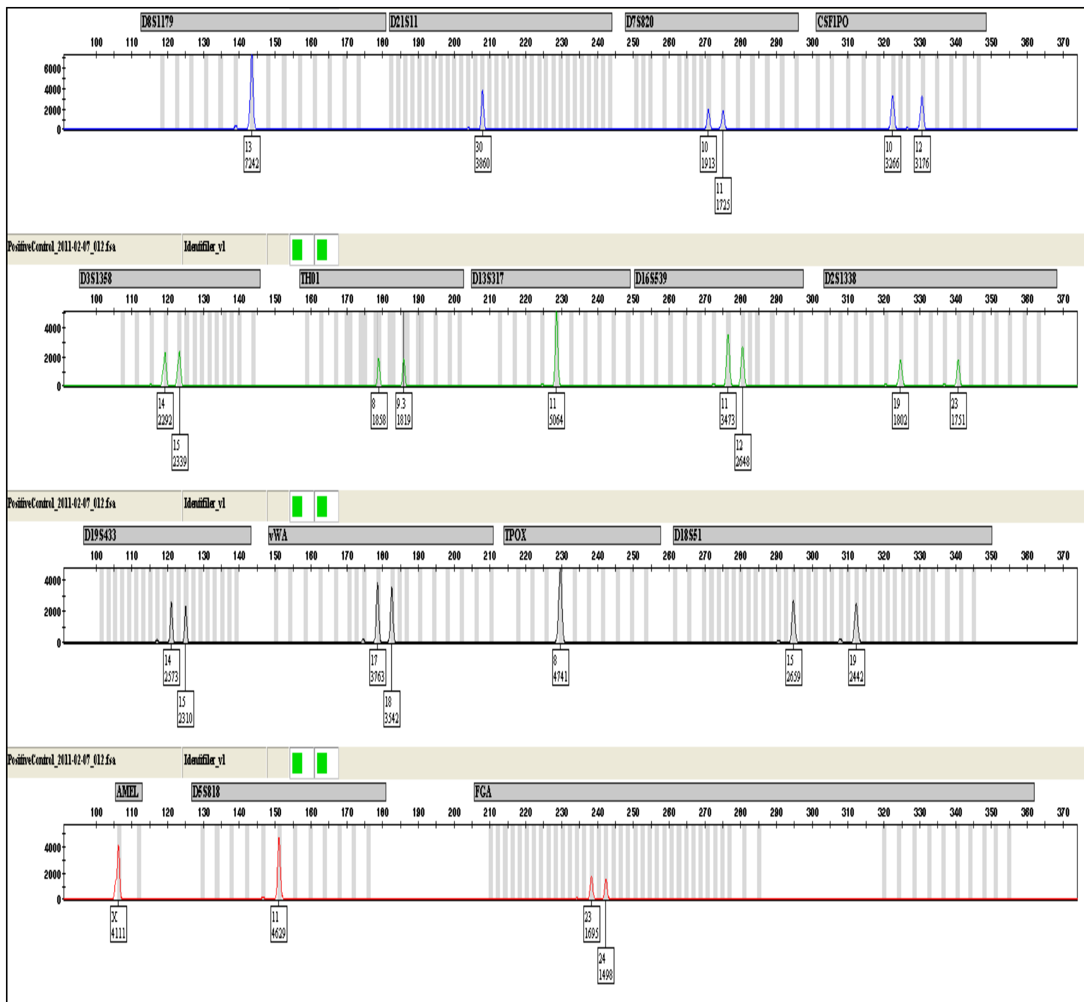

Figure 1.7: Representative Electropherogram of the AmpF*l*STR® control DNA 9947A.