

Mitochondrial DNA Analysis in the United Arab Emirates

Reem M. Almheiri

A thesis submitted of the requirements for the degree of Doctor of Philosophy in
Forensic Science, School of Law and Policing. University of Central Lancashire, 2025.



STUDENT DECLARATION FORM

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution. I declare that no material contained in the thesis, except where otherwise stated, has been used in any other submission for an academic award and is solely my own work.

Reem Matar Khalifa Almazaina Almheiri

Type of Award: Doctor of Philosophy

School: School of Law and Policing

Abstract

Mitochondrial DNA (mtDNA) analysis has emerged as a powerful investigation tool in forensic science over the past few decades. The unique properties of mtDNA, such as maternal inheritance and high copy number, make it particularly useful in cases where nuclear DNA is rapidly degraded or insufficient. Until recently, routine analysis has been limited to the control region; however, recent advances in sequencing technologies have enabled comprehensive analysis of the entire mitochondrial genome. The development of massively parallel sequencing (MPS) technologies has revolutionized the field, allowing for high-throughput sequencing with enhanced discrimination power and cost efficiency.

This study aimed to develop a relevant high quality population database for frequency estimation and to evaluate the implementation of whole mitochondrial genome sequencing in forensic science within a selective population in the United Arab Emirates (UAE).

A population database consisting of 610 whole mitogenome haplotypes was developed, encompassing 510 samples from UAE nationals, 50 samples from Indians, and 50 samples from Pakistanis. The results show that each sample exhibited a set of SNPs variations, which were used to define individual haplotypes and subsequently assign each sample to its closest haplogroup.

The Precision ID Whole mtDNA Genome Panel Kit was utilized to implement the whole mitochondrial genome sequencing. The overall haplotype diversity was comparable to other global populations, with minor improvements in haplotype diversity observed when

compared to control region analysis alone. The total number of haplotypes was 445 for the Emiratis, 49 for the Indians and 50 for the Pakistanis. The unique haplotypes reported 400 individuals out of the 510 Emirati mtDNA sequences. For Indians, 48 individuals expressed unique haplotypes. Last, for Pakistanis, 50 individuals had unique haplotypes. Those numbers represented 78.43% unique haplotypes within the dataset of this study, while the remaining 21.57% were shared haplotypes for Emiratis. For the Indian population, 96% of the individuals had unique haplotypes, with only two individuals having a shared. All the Pakistani samples displayed unique haplotypes, accounting for 100% of their dataset.

Haplogroups for the obtained haplotypes were assigned using various tools, including online EMPOP and Haplogrep, and Converge Software. In the Emirati dataset, the H haplogroup was the most common at 15.49%, followed by Haplogroup U at 13.73%. In the Pakistani dataset, Haplogroup M is predominant, accounting for 50% of the population, followed by Haplogroup U at 22%. Similarly, in the Indian dataset, Haplogroup M is also the most prevalent, comprising 44% of the samples, with Haplogroup U appearing in 22% of the population. Emiratis had 241 different haplogroups, while Pakistanis and Indians showed 42 and 40 haplogroups, respectively.

Initially, a cohort of 93 biological specimens (blood reference samples) underwent sequencing of the mitochondrial control region via the Sanger method. Subsequently, these identical specimens were subjected to MPS utilizing Ion Torrent technologies.

Complete concordance was achieved between the datasets generated by these two sequencing methodologies.

The study was extended to implement this technology to 56 casework samples and 56 bone samples, to evaluate its robustness. Profiles were obtained from all samples, although some were classified as partial.

In conclusion, the results from the study represent a crucial step toward the implementation of mtDNA analysis in forensic science in the UAE, enhancing the capability to analyse previously unusable samples and improving responses to missing persons cases and mass disasters. The successful integration of MPS technologies into forensic workflows will significantly advance our understanding of the application of mitochondrial DNA analysis and its potential in the context of the UAE.

Acknowledgement

"In the name of Allah, the Most Gracious, the Most Merciful. 'And whatever of blessings and good things you have, it is from Allah.' (Quran, Surat al-Nahl: 18)"

I would like to extend my sincere appreciation to all individuals who have played a significant role in the completion of this PhD thesis, without whom this work would not have been possible.

First and foremost, I am deeply grateful to my supervisors, William H. Goodwin and Rashed H. Alghafri, whose guidance, support, and supervision have been invaluable throughout this research journey. Their insights, expertise, and support have been essential in balancing my research responsibilities and professional commitments. Their mentorship has greatly enriched my understanding of the subject matter, and their continuous encouragement has profoundly influenced the direction and quality of this thesis.

I would also like to express my heartfelt thanks to the UCLan faculty members throughout my student years for their valuable contributions. Their input, feedback, and recommendations have greatly enhanced the scholarly rigor of this work.

I am thankful to the participants who provided their samples and consented to be part of this study. Their cooperation was crucial for data collection, advancing mitochondrial DNA analysis, and contributing to the understanding of the United Arab Emirates.

I acknowledge the support and resources provided by the International Center For Forensic Sciences – Dubai Police. Their commitment to research excellence and the facility environment have been instrumental in the successful completion of this thesis.

To my friends, I cannot think of a better group of people to be on this journey with. I am very fortunate to have such a supportive system. Your camaraderie and support have made this journey intellectually stimulating and enjoyable.

Lastly, I express my heartfelt appreciation to my family members and loved ones for their unwavering love, encouragement, and understanding. Their belief in my abilities and unwavering reinforcement have been the driving force behind my pursuit of knowledge.

I extend my gratitude to Thermo Fisher Scientific for their prompt support and technical assistance throughout this research. Their expertise and constant availability were crucial in overcoming challenges and contributed to the success of this thesis.

To all those mentioned above and to others who have contributed in various ways, known or unknown, I am deeply thankful. You have been crucial in shaping my academic and personal growth, and I feel privileged to have had the opportunity to collaborate with such exceptional individuals. Finally, I would like to acknowledge my hard work, perseverance, and dedication throughout this journey. The countless hours of research, writing, and learning have been challenging yet rewarding. I am proud of my commitment and grateful for the growth and knowledge gained. This achievement is a testament to my determination and passion for scientific discovery.

List of Abbreviations

°C: Degrees Celsius

xg: Force of Gravity

A: Adenine

ABI: Applied Biosystem

AMOVA: Analysis of Molecular Variance

ATP: Adenosine Triphosphate

BAM: Binary Alignment Map

BCE: Before Common Era

bp: Base Pair

BTA: Buffer TE with Albumin

C: Cytosine

CE: Capillary Electrophoresis

CR: Control Region

CRS: Cambridge Reference Sequence

D.F: Degrees of Freedom

DNA: Deoxyribonucleic acid

D-loop: Displacement loop

DTT: Dithiothreitol

EDTA: Ethylenediaminetetraacetic Acid

EMPOP: EDNAP (European DNA Profiling Group) Mitochondrial DNA Population Database

FBI: Federal Bureau of Investigation

FSS: Forensic Science Service

FTA: Flinders Technology Associates

G: Guanine

GEDNAP: German DNA Profiling Group

H-strand: Heavy Strand

Hg: Haplogroup

HgD: Haplogroup Diversity

HV1: Hypervariable segment I

HV2: Hypervariable segment II

HV3: Hypervariable segment III

HWE: Hardy-Weinberg Equilibrium

ICRC: International Committee of the Red Cross

IGV: Integrative Genomic Viewer

INDEL: insertion-Deletion Polymorphisms

IPC: Internal Positive Control

ISFG: International Society for Forensic Genetics

ISPs: Ion Sphere Particles

kb: Kilobases/Kilobase-pairs

KYA: Thousand Years Ago

L: Liter

L-strand: Light Strand

LH: Length Heteroplasmy

LHON: Leber's Hereditary Optic Neuropathy

M: Molar

Mb: Megabase-Pairs

MDS: Multidimensional Scaling

mg: Milligram

min: Minute(s)

μL: Microliter

MLP: MultiLocus Probes

mm: Millimeter

mM: Millimolar

MOU: Memorandum of Understanding

MPS: Massive Parallel Sequencing

mtDNA: Mitochondrial DNA

NDIS: National DNA Index System

NFW: Nuclease Free Water

ng: Nanogram

NGS: Next Generation Sequencing

NL: Noise Level

NTC: Non Template Control

NUMTs: Nuclear Mitochondrial DNA Segments

OLA: Oligonucleotide Ligation Assay

OXPHOS: Oxidative Phosphorylation

PC: Promega Corporation

PCR: Polymerase Chain Reaction

PD: Power of Discrimination

PE: Power of Exclusion

pGEM®: plasmid vectors by Promega Corporation

pH: Potential of Hydrogen

PHP: Point Heteroplasmy

PK: Proteinase K

PM: Probability of Match

Pmol: Picomoles

Pg: Picogram

PPE: Personal Protective Equipment

QC: Quality Control

RFLP: Restriction fragment Length Polymorphism

RMP: Random Match Probability

RNA: Ribonucleic Acid

ROS: Reactive Oxygen Species

rpm: Revolutions Per Minute

s: Second(s)

SBE: Single-Base Extension

SNH: Single Nucleotide Heteroplasmy

SNP: Single Nucleotide Polymorphism

SOP: Standard Operating Procedures

STR: Short Tandem Repeat

TE: Tris-Ethylenediaminetetraacetic acid

TFS: Thermo Fisher Scientific

Tris-HCl: Tris(hydroxymethyl)aminomethane hydrochloride

TSC: The SNP Consortium

TSS: Torrent Suite™ Software

VNTR: Variable Number Tandem Repeat

STUDENT DECLARATION FORM.....	i
Abstract	ii
Acknowledgement	v
List of Abbreviations.....	vii
List of Figures	xvi
List of Tables.....	xxiv
Chapter 1.....	27
1. Introduction	27
1.1. A Historical Perspective of Forensic Science and Its Current Applications	27
1.2. The Arabian Peninsula	31
1.3. The formation of the United Arab Emirates	35
1.4. A brief history of Forensic Genetics.....	40
1.4.1. DNA Profiling.....	42
1.4.2. STRs.....	42
1.5. Mitochondria	44
1.5.1. Mitochondrial DNA	45
1.5.2. Inheritance pattern of mtDNA.....	48
1.5.3. Copy Number	50
1.5.4. Heteroplasmy.....	50
1.5.5. Nuclear Mitochondrial DNA Segments (NUMT)	52
1.6. Analysis methods development.....	54
1.6.1. Restriction Fragment Length Polymorphism (RFLP)	54
1.6.2. Polymerase Chain Reaction (PCR).....	55
1.6.3. RFLP-PCR Assays.....	55
1.6.4. Sanger Sequencing.....	56
1.7. Comparison of Nuclear DNA (nDNA) and mtDNA	58
1.8. mtDNA Population Genetics	58
1.8.1. Haplotypes	59
1.8.2. mtDNA Haplogroups and Phylogenetic Tree	60
1.8.3. Forensic Genetics.....	66
1.9. Whole mtDNA genome	67
1.10. Mitochondrial DNA Analysis in UAE Populations	68

1.11.	Mitochondrial DNA Analysis in Middle East Populations	72
1.12.	Mitochondrial DNA Analysis Sub-Asian Population.....	79
1.12.1.	Indians	79
1.12.2.	Pakistanis	82
1.13.	Aim	86
1.14.	Objectives.....	87
Chapter 2.....		89
2.	Materials and Methods	89
2.1.	Materials	89
2.2.	Samples from United Arab Emirates Population	92
2.3.	Amplification of mtDNA Control Region	93
2.3.1.	Primers Selection	93
2.3.2.	Primers Quality Control	94
2.4.	Quality Control.....	94
2.5.	DNA Extraction and Purification	95
2.5.1.	Population Samples Extraction	96
2.5.2.	Casework Samples	97
2.5.3.	Bone Samples.....	97
2.5.4.	Direct PCR Extraction	98
2.5.5.	DNA quantitation	98
2.6.	The PCR Amplification conditions	101
2.6.1.	Primers Selections.....	102
2.6.2.	Control Region Amplification	102
2.6.3.	Gel Electrophoresis	103
2.7.	Sanger Sequencing.....	103
2.7.1.	PCR Product Purification	103
2.7.2.	BigDye® Sequencing Reaction	104
2.7.3.	BigDye® Sequencing Product Purification	104
2.7.4.	Sequence Detection and Analysis	105
2.8.	Whole mtDNA Sequencing	106
2.9.	Commercially Available Whole mtGenome MPS workflow	107
2.10.	Ion Torrent workflows	109

2.10.1.	Qubit® 3 Fluorometer Quantification	110
2.10.2.	Library Preparation	114
2.10.2.1.	Ion Chef Library Preparation	115
2.10.2.2.	Manual Library Preparation	115
2.10.3.	DNA Sequencing.....	124
2.10.3.1.	Initialization.....	125
2.10.3.2.	Sequencing.....	126
2.10.4.	Statistical data Analysis.....	126
Chapter 3.....		128
3.	Sanger Sequencing Analysis of the Control Region	128
3.1.	Introduction	128
3.2.	Chapter Aim	129
3.3.	Chapter Objectives.....	129
3.4.	Methods	129
3.4.1.	Sanger sequencing Optimization	129
3.4.2.	DNA Extraction.....	130
3.4.3.	Control Region Amplification	131
3.4.4.	BigDye® Terminator Sequencing	134
3.5.	Populations Analysis using Sanger sequencing	139
3.6.	Discussion.....	152
3.7.	Conclusion	158
Chapter 4.....		159
4.	Massive Parallel Sequencing of Whole mtDNA Genome	159
4.1.	Introduction	159
4.2.	Chapter Aim	160
4.3.	Chapter Objectives.....	161
4.4.	Methods	162
4.5.	Massive Parallel Sequencing Emirati Reference Data	162
4.6.	Whole mtDNA Variant Calling Sequencing Analysis	163
4.6.1.	DNA sequences Reconstruction.....	165
4.6.2.	Alignment and Reference Genome	165
4.7.	Quality assessment and Output Files	166

4.8.	Mitochondrial genome coverage.....	173
4.9.	MPS Data Generation	180
4.10.	Mixtures Detection	180
4.11.	Validation of Results Using Standards	180
4.12.	Proficiency Testing	189
4.13.	Discussion.....	192
4.14.	Conclusion	193
Chapter 5.....		194
5.	Massive Parallel Sequencing Data Analysis	194
5.1.	Introduction	194
5.2.	Chapter Objectives.....	195
5.3.	Methods	195
5.4.	Results: Section 1	195
5.5.	MPS Data Analysis: Emiratis Samples Set	195
5.5.1.	Haplotypes Generating and Assessment	196
5.5.2.	Haplogroups Assignments	202
5.6.	Results: Section 2	231
5.7.	MPS Data Analysis: Indians Samples Set	231
5.7.1.	Haplotypes generating.....	231
5.7.2.	Haplogroup Assignments	236
5.8.	Results: Section 3	237
5.9.	MPS Data Analysis: Pakistanis Samples Set	237
5.9.1.	Haplogroup Assignment.....	243
5.9.2.	Haplogroup Diversity	244
5.10.	Concordance between Sanger Sequencing and MPS	244
5.11.	Discussion.....	245
5.12.	Conclusion	250
Chapter 6.....		253
6.	Casework Samples	253
6.1.	Introduction	253
6.2.	Chapter Aim	254
6.3.	Chapter Objectives.....	254

6.4.	Methods	254
6.5.	Casework Samples Sequencing	255
6.6.	Results	260
6.7.	Discussion.....	262
6.8.	Conclusion	265
Chapter 7.....		266
7.	Bone and High Primate Samples	266
7.1.	Introduction	266
7.2.	Chapter Aim	266
7.3.	Chapter Objectives.....	267
7.4.	Methods	267
7.5.	Solid Tissues Sequencing	267
7.6.	High Primates Samples	271
7.7.	Results	271
7.8.	Discussion.....	275
7.9.	Conclusion	278
Chapter 8.....		280
8.	General Discussion.....	280
8.1.	Introduction	280
8.2.	Mitochondrial DNA Analysis Techniques	282
8.3.	Population-Specific Findings in the UAE	283
8.4.	Comparison Across Populations (Emirati, Indian, Pakistani)	284
8.5.	Forensic Casework Samples	285
8.6.	Challenges in mtDNA Sequencing and Heteroplasmy	286
8.7.	Beyond Forensic Significance.....	287
8.8.	Human population genetics.....	289
8.9.	Conclusions	291
8.10.	Scope for Future Studies	294
8.11.	Limitations of this Study	295
8.12.	Recommendations	298
References.....		303
Appendices.....		328

List of Figures

Figure 1.1.A map showing the borders of each city in the UAE (Taken from Abed and Hellyer, 2001).	39
Figure 1.2. Geographical location of the UAE (red) on the world map relative to Middle East and East Asia, highlighting Pakistan and India (Taken from Grubb, 2007).	40
Figure 1.3. Illustrative image showing the position of mitochondria inside the biological cell and their general structure (National Human Genome Research Institute, 2023). ...	45
Figure 1.4. Diagram showing mitochondrial circular DNA structure (Zeviani et al. 2007).	48
Figure 1.5. Expanded control region with hypervariable region indicated (Lutz et al., 2000).	48
Figure 1.6. Worldwide spread of human population migrations pattern and major mtDNA haplogroups (Taken from Nature Reviews Genetics 2015, 16 (9): 530-542).	62
Figure 1.7. Human mitochondrial DNA phylogenetic tree. (Wikitree n.d.)	64
Figure 1.8. Haplogroup T2b7a1 distribution on a world map obtained from EMPOP.	65
Figure 1.9. Representation of haplogroups in the Dubai dataset derived from control region sequencing (Alshamali et al., 2007).	70
Figure 1.10. Representation of haplogroups frequencies in a whole mtDNA pie chart from the study by Aljasmi et al., 2020 (N=232).	72
Figure 1.11. Representation of haplogroups in the Egyptian population mtDNA control region (Saunier et al., 2009).	73
Figure 1.12. Pie chart representation of Kuwaiti population haplogroups of mtDNA CR (Scheible et al., 2011)	74
Figure 1.13. Bahraini population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)	77
Figure 1.14. Lebanese population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)	78
Figure 1.15. Jordanian population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)	78

Figure 1.16. Graphical representation of frequencies of the mtDNA haplogroup composition of the 100 sampled Makrani from Pakistan (Taken from Siddiqi et al., 2015)	84
Figure 1.17. Demonstration of the relative proportions of the observed haplogroups in Sindhi population (Yasmin et al., 2017)	85
Figure 2.1. A Schematic diagram showing the primer chosen to amplify mtDNA control region	94
Figure 2.2. Calibration graph of Qubit® quantification assays	113
Figure 2.3. Library dilutions to target concentration 30 pM	121
Figure 2.4. Ion Chef™ Instrument deck (Illustration obtained from Precision ID mtDNA Panels with the HID Ion S5™/HID Ion GeneStudio™ S5 System Application Guide)	122
Figure 3.1. A captured run of samples on 2% agarose gel electrophoresis. HyperLadder™ 1 kb is in lane M (far right and left of the sample wells). Lane 1 is the negative control. Lanes 2-5, 6-9, 10-13, and 14-17 corresponds to female blood samples 1, 2, 3, and 4 respectively, with four replicates (A-D). See Table 3.1 for sample details.	134
Figure 3.2. An electropherogram showing the mtDNA control region sequence with smooth baseline using forward primer (L15879) using precipitation purification method with ethanol. In this and subsequent figures, electropherogram is typical of four different experiments	136
Figure 3.3. An electropherogram of the mtDNA control region forward primer (L15879) sequence with high baseline noise using BigDye XTerminator® Purification Kit.	136
Figure 3.4. An electropherogram showing the mtDNA control region sequence with smooth baseline using reverse primer (H727) using the precipitation purification methods with ethanol	136
Figure 3.5. An electropherogram displaying the mtDNA CR sequence with noticeable baseline noise using the reverse primer (H727), following the BigDye XTerminator® Purification Kit	137
Figure 3.6. BioEdit™ software result showing an example of sequence quality generated for one sample using three primers F15975 (top), R240 (middle), and R635 (bottom).	138

Figure 3.7. Electropherogram of a single random sample showing the dropping effect in sequencing the control region using a single forward primer (F15975).....	138
Figure 3.8. Electropherogram of a single random sample capturing the sequence drop of the control region generated using a single reverse primer (R635) (Sequence viewed without the reverse option).....	139
Figure 3.9. Results of the mtDNA control region analysis performed using the mtProfiler tool (http://mtprofiler.yonsei.ac.kr)	146
Figure 3.10. HaploGrep results screen of a batch entry for Emirati samples, showing haplogroups, clusters, and quality control with Haplogroup H as the most prevalent (Parson and Bandelt, 2007). The Figure also highlights warnings and failed samples. Warnings may indicate issues such as undetermined variants, global private mutations unknown to Phylotree, or local private mutations associated with other haplogroups. Failed samples are marked in red and generally occur when the detected haplogroup quality is low (quality $\leq 80\%$) or when the sample is missing more than two expected polymorphisms.....	148
Figure 3.11. Pie chart showing haplogroup clusters of the Emiratis set.....	156
Figure 3.12. Pie chart showing haplogroup clusters of the Pakistanis set.....	157
Figure 3.13. Pie chart showing haplogroup clusters of the Indians set	157
Figure 4.1. Circular plot of the mtGenome sequencing coverage for the forward and reverse pools summary of extracted reference blood samples developed by TSS.	167
Figure 4.2. A closer image of the mtGenome sequencing coverage chart, the green colored squares flags confirmed variants in the sequence. The blue color indicates forward coverage in sequencing data. It shows how many reads from the sequencing process cover each position in the forward direction of the DNA strand. Conversely, the red color represents reverse coverage. The kilobase (kB) marks indicates the positions along the mtDNA that is around 16.5 kB in length, which help in pinpointing specific locations on the mtDNA. The numeral intervals (2000, 4000, 6000, 8000) are marks to provide visual assess of the coverage reads measurement.....	168
Figure 4.3. An example of the 'Variant' sheet in the `variants_colored.xlsx` file.	169
Figure 4.4. Colour codes used in the 'Variant' sheet in the `variants_colored.xlsx` and the different scores for the listed columns.	171

Figure 4.5. ISP density colour gradient, where Red indicates areas of high ISP loading, while blue is the lowest.	174
Figure 4.6. Read Length Histogram.....	174
Figure 4.7. Summary of Ion Sphere Particle (ISP) Sequencing Results.....	176
Figure 4.8. Emirati samples Chip 1 TSS quality check.....	176
Figure 4.9. Emirati samples Chip 2 TSS quality check.....	177
Figure 4.10. Emirati samples Chip 3 TSS quality check	177
Figure 4.11. Pakistani samples Chip 1 TSS quality check	178
Figure 4.12. Pakistani samples Chip 2 TSS quality check	178
Figure 4.13. Indian samples Chip 1 TSS quality check.....	179
Figure 4.14. Indian samples Chip 2 TSS quality check.....	179
Figure 4.15. Deletion at position 3107 with 98.8% in Promega 9947 standard.....	184
Figure 4.16. Position 14470C in Component CHR (A) with 100% confirmed cytosine.	186
Figure 4.17. Position 309.2C of Component GM09947A (B) with 80.6% confirmed insertion.	186
Figure 5.1. An overview of 19 haplogroups in 510 samples of the Emirati population, detailing sample counts, frequencies, and percentages. The coloured bar illustrates population frequencies for variants observed across different global populations. Each colour corresponds to a specific population category as follows: Purple: African/African American, Light Blue: European (non-Finnish), Dark Blue: European (Finnish), Teal: Amish, Lavender: Ashkenazi Jewish, Gold: Middle Eastern, Orange: South Asian, Green: East Asian, Red: Latino/Admixed American and Grey: Other. This representation was generated using Haplogrep 3 (Schönherr et al., 2023).....	203
Figure 5.2. The distribution of haplogroup H within the population is illustrated in the accompanying chart. The colour-coded haplogroups are as follows: H1+152 (red), H13b1+200 (dark brown), H1a (brown), H1e1a4 (blue-grey), H2+152+16311 (purple), H2a* (light beige), H2a2a1 (white), H2a2a1c (light grey), H2a2a2 (pink), H3h7 (dark brown), H57 (brown), H6 (green), HV14 (golden yellow), HV2a (red-orange), H1+16355 (deep blue), H14a (teal), H1ah2 (cyan), H1e1a6 (grey-blue), H2a1 (dark green), H2a2a1d (olive green), H32 (light grey), H5 (light brown), H5a1+152 (brown), H6b2 (yellow-	

brown), HV15 (bright green), HV6 (lime green), H11a (light yellow-green), H14b1 (pale yellow), H1b (light grey), H1e+16129 (dark grey), H1q3 (pink), H2a2a1f (green), H3c2 (light olive), H5'36 (brownish pink), H5a5 (light pink), H5r (hot pink), H8+(114) (light brown), HV1 (magenta), HV15*1 (soft pink), and HV21 (navy blue).205

Figure 5.3. The distribution of haplogroup U within the population is illustrated in the accompanying chart. The colour-coded haplogroups are as follows: U1a (light blue), U2b2 (lavender), U2e1'2'3 (sky blue), U3a (green), U5a1+@16192 (pale green), U5b2a1a+16311 (dark green), U7 (white), U9a (pink), U1a3 (orange), U2c'd (light pink), U2e1f (gold), U3a2a1 (yellow), U5a2a (mustard yellow), U5b2b4a (olive), U2a (blue), U2b (dark blue), U2d (pale pink), U2e3 (orange), U3b3 (yellow-orange), U5b (red), U6 (purple), U7a3b (pale green), U9a1 (yellow), U9b1 (navy blue), U2e1 (light grey), U3 (light green), U4b1+146+152 (dark green), U5b1d1a (brown), U6a+16189+(103) (grey), and U7a4 (bright red).207

Figure 5.4. The distribution of haplogroup J within is illustrated in the chart, generated using Haplogrep 3 software. The colour-coded haplogroups are: J2a2b (dark blue), J1c5 (light blue), J1b1a1 (dark orange), J1c (peach), J1c2 (dark green), J1c4 (pale green), J1c3 (mustard yellow), J2a2 (pale yellow), J1b (teal), J1b* (light turquoise), J2a2e (red), J1b1b3 (light pink), J1d1a1 (grey), J1b2 (light grey), J1b1b1 (dark pink), J1d1a (light pink), J2a2a1 (lavender), J1b3a (pale purple), JT (light brown), J (pale beige), J1d (royal blue), J1d3a (sky blue).209

Figure 5.5. The distribution of haplogroup M illustrated in the chart. The colour-coded haplogroups are: M (bright red), M1 (violet), M1a1 (dark gray), M1a5 (light sky blue), M18'38 (medium gray), M23 (lavender), M2b (cyan), M2b1a (light cyan), M3 (dark green), M30 (light brown), M30*1 (peach), M30+16234 (teal), M30b (dark olive), M30c1 (sand brown), M30d (warm brown), M30g (chestnut brown), M33a1b (deep red-brown), M33a2a (salmon pink), M3a1+204 (light orange-brown), M36a (peach-orange), M3c2 (deep purple), M38+195 (brick red), M3d (dark purple), M40a1a (royal blue), M45a (mustard yellow), M49 (deep forest green), M4a (sea green), M5a1b* (olive green), M5a2a1a (teal green), M5a2a2* (lime green), M5a2a4 (bright yellow-green), M5b2 (golden yellow), M57+152 (spring green), M6a1a (crimson red), and M65b (orange). 211

Figure 5.6. The distribution of R subhaplogroups illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: R0a1+152 (dark blue), R0a2 (light blue), R5a2b (dark orange), R (peach), R0a2d (dark green), R0a1a1a (pale green), R0a1a4 (mustard yellow), R0a2f (pale yellow), R0a2h (teal), R0 (light turquoise), R0a1a (red), R0a (light pink), R30b2a (grey), R0a1a1 (light grey), R2d (dark pink), R30a1c (light pink), R2a (lavender), R2c (pale purple), R30 (light brown), R9b1 (pale beige), R9 (royal blue), R7 (sky blue), R5a1a (orange), R6a2 (pale orange), R5a2 (dark green).....213

Figure 5.7. The distribution of N subhaplogroups within the population is illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: N (dark blue), N2a2 (light blue), N1b1a2 (dark orange), N1a1a3 (peach), N1a3a (dark green), N1b1a+16129 (pale green), N1a1b1 (mustard yellow).....214

Figure 5.8. The distribution of T subhaplogroups illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: T1a (dark blue), T1a1 (light blue), T1a2a (dark orange), T2g (peach), T2b (dark green), T2e (pale green), T1a1m1 (mustard yellow), T2d2 (pale yellow), T2b4+152 (teal), T2c1a2 (light turquoise).....215

Figure 5.9. The distribution of K subhaplogroups within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: K1a4c1 (dark blue), K1a4f (light blue), K2b1b (dark orange), K1a1a1 (peach), K1a2a (dark green), K1a3a (pale green), K1a4c (mustard yellow), K2a2a (pale yellow), K2a5b (teal), K1a4 (light turquoise).....216

Figure 5.10. The distribution of subhaplogroups of macrohaplogroup L within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: L1c3a (dark blue), L1b1a8 (light blue), L3e2b1a2 (dark orange), L0d2c1a (peach), L2b1a (dark green), L1 (pale green), L3e2b (mustard yellow), L3e1e1 (pale yellow), L3e1b2 (teal), L3f1b4a1 (light turquoise), L3f1b41 (red), L3d1a1a (light pink), L2a1h* (grey), L3b1a1a (light grey), L2a1a2 (dark pink), L1c3b1a (light pink), L1c2a1a (lavender), L2a1b1a (pale purple), L1c2b2 (light brown), L3b1a+152 (pale beige), L3h2 (royal blue), L2a1+143 (sky blue), L2a1+143+16189+16192 (orange), L0a2a2a (pale orange), L0a1a2 (dark green), L1ba3 (yellow green).....218

Figure 5.11. The distribution of subhaplogroups of macrohaplogroup X within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: X2+225+@16223 (dark blue), X2c1b (light blue), X2o1 (dark orange), X2m'n (peach), X2d1 (dark green), X (pale green).	220
Figure 5.12. The distribution of subhaplogroups of macrohaplogroup I within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: I1a1 (dark blue), I6 (light blue), I* (dark orange), I1 (peach), I1c1 (dark green), I5a3 (pale green).	221
Figure 5.13. The distribution of subhaplogroups of macrohaplogroup W within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: W3a1 (dark blue), W6a (light blue), W1+119 (dark orange), W6b1 (peach), W1 (dark green), W+194 (pale green).	222
Figure 5.14. The distribution of subhaplogroups of macrohaplogroup B within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: B5b1c (dark blue), B4a1a (light blue).	223
Figure 5.15. The distribution of subhaplogroups of macrohaplogroup D within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: D4j (dark blue), D4i (light blue).	224
Figure 5.16. The distribution of subhaplogroups of macrohaplogroup F within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: F1a1a1 (dark blue), F3b1b1 (light blue).	225
Figure 5.17. An online alignment and haplogroup assignment using EMPOP platform of the haplogroup H6b as a heat map.	226
Figure 5.18. Comparative graph of haplogroup percentages (B, D, F, H, HV, I, J, K, L0, L1, L2, L3, L4, M, N, R, T, U, W, X) between three studies on the UAE population. The studies include Aljasmi et al. (2020) with n=232 samples (blue), Alshamali et al. (2008) with n=249 samples (green), and the current study with n=510 samples (orange).	229

Figure 5.19. Whole mtDNA pie chart representation of the haplogroups frequencies identified in the study conducted by Aljasmi et al., 2020 (n=232). (Chapter 1)	230
Figure 5.20. Whole mtDNA pie chart representation of the haplogroups frequencies identified by the current study conducted (n=510).....	230
Figure 5.21. An overview of 9 haplogroups in 50 samples of the Indians population, detailing sample counts, frequencies, and percentages.	236
Figure 5.22. Whole mtDNA pie chart representation of the Indians haplogroups frequencies identified by the current study conducted.....	237
Figure 5.23. An overview of 8 macro-haplogroups in 50 samples of the Pakistani population, detailing sample counts, frequencies, and percentages.....	243
Figure 5.24. Whole mtDNA pie chart representation of the Pakistanis haplogroups frequencies identified by the current study conducted.....	244
Figure 6.1. Casework samples chip 1 run summary from the TSS (n=28).....	257
Figure 6.2. Casework samples chip 2 run summary from the TSS (n=28).....	257
Figure 6.3. Comparison of the casework samples to the three populations data sets. ..	264
Figure 7.1. Solid tissue samples chip 1 run summary from the TSS (n=28).....	269
Figure 7.2. Solid tissue samples chip 2 run summary from the TSS (n=31).....	270
Figure 7.3. A partial screenshot from TSS output file of the male Chimpanzee results.	274
Figure 7.4. Comparison of the samples of solid tissues to the three populations data sets.	277

List of Tables

Table 1.1 Characteristics of nDNA and mtDNA (Butler, 2012).....	58
Table 2.1. Forward and reverse primers obtained from the online protocol (Yonsei University, n.d.) for the amplification of the mtDNA CR.....	94
Table 2.2. Different extractions kits used in the studied samples.....	96
Table 2.3. Primers for PCR and Sequencing.....	102
Table 2.4. PCR conditions for amplification of mtDNA control region	103
Table 2.5. Sequencing PCR cycles conditions.....	104
Table 2.6. Qubit® working solution preparation per sample (n)	111
Table 2.7. Qubit® dsDNA HS Assay samples' preparation.....	111
Table 2.8. Target amplification PCR mix preparation.....	117
Table 2.9. PCR conditions for whole mtDNA target amplification.....	117
Table 2.10. PCR conditions for digestion of target amplicons.....	118
Table 2.11. Perform the ligation reaction.....	118
Table 2.12. Thermal cycler parameters of ligation reaction	118
Table 2.13. Three 10-fold serial dilutions of the E. coli DH10B Control Library.....	120
Table 2.14. QuantStudio™ 5 Real-Time PCR run parameters	121
Table 2.15. Torrent Suite™ Software Plan screen setup.....	123
Table 3.1. Four samples used for comparison for FTA® card and PrepFiler® extraction, and two master mixes Platinum® and Reddymix™	132
Table 3.2. Control region sequences for 33 Emiratis samples with the haplotypes nomenclatures.....	141
Table 3.3. Mitochondrial DNA control region polymorphic variants in the three ethnic groups.....	150
Table 3.4. Number of Haplotypes (Ht), number of Haplogroups (Hg), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for the three populations sets.....	154
Table 4.1. Torrent Suite™ Software (TSS) default analysis parameters obtained from the MITO Genotyper.....	163
Table 4.2. Detailed list of the columns of the 'Variant' sheet in the `variants_colored.xlsx` file.....	170

Table 4.3. Quality Control Results of UAE samples	177
Table 4.4. Quality Control Results of Pakistanis population samples.....	178
Table 4.5. Quality Control Results of Indian population samples	179
Table 4.6. Comparison of the obtained sequence of 2800M promega to the reference	183
Table 4.7. Comparison of Sequencing Results with Promega 9947A Standard	184
Table 4.8. Comparison of Sequencing Results with Origene 9947A Standard.....	185
Table 4.9. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component CHR (A).....	187
Table 4.10. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component GM09947A (B).....	188
Table 4.11. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component CHR clone (C).....	189
Table 4.12. Comparison between the examination reference and the results of GEDNAP 66 sequences of Person A, Person B, and Person C, obtained using Ion Torrent Technologies and Precision ID Whole mtDNA genome panel.....	191
Table 4.13. Comparison between the examination reference and the results of GEDNAP 67 sequences of Person A, Person B, and Person C, obtained using Ion Torrent Technologies and Precision ID Whole mtDNA genome panel.....	191
Table 5.1. Whole mtDNA genome sequences for 33 Emirati samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing (as shown in Table 3.2) are indicated in italics, and insertions/deletions (indels) are highlighted in red.....	197
Table 5.2. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for the three populations sets.	202
Table 5.3. Whole mtDNA genome sequences for 30 indian samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing are indicated in italics, and insertions/deletions (indels) are highlighted in red.	232
Table 5.4. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for Indian samples.	236

Table 5.5. Whole mtDNA genome sequences for 30 Pakistani samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing are indicated in italics, and insertions/deletions (indels) are highlighted in red.	238
Table 5.6. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for Pakistani samples.	243
Table 5.7. Number of Haplogroups (Hg) and calculated Haplogroup Diversity (HgD) for the three populations in this study.	244
Table 5.8. Comparison of Sanger Sequencing and MPS.	246
Table 6.1. Casework samples obtained from Forensic Biology section – Dubai Police (n=56).	255
Table 6.2. Quality Control Results of casework samples (n=56).	257
Table 6.3. Sequence data summary for casework samples in comparison with traditional profiling results (n=56).	259
Table 7.1. Solid Tissue samples obtained from Forensic Biology section – Dubai Police (n=56).	267
Table 7.2. Quality Control Results of solid tissue samples (n=56).	270
Table 7.3. Sequence data summary for the solid tissue samples (bones and teeth) in comparison with traditional profiling results	272

Chapter 1

1. Introduction

1.1. A Historical Perspective of Forensic Science and Its Current Applications

Forensic science has ancient origins, evolving significantly from early rudimentary methods to modern, scientifically-driven practices. Initially used without much understanding of its underlying principles, forensic science has become an indispensable tool in both crime-solving and distinguishing human populations through genetic analysis. Its evolution demonstrates the continuous development of legal and scientific practices, bridging the gap between ancient knowledge and modern innovation. This journey underscores the critical role forensic science plays in ensuring justice and understanding human history.

Forensic science dates back to ancient civilizations, with the first recorded case occurring in Ancient Rome in 44 BCE. The Roman physician Antistius performed an autopsy on Julius Caesar, revealing that out of the 23 stab wounds, only one was fatal. This incident is recognized as the earliest recorded use of forensic pathology, laying the foundation for expert testimony in legal matters. The term “forensic” itself is derived from the Latin word *forensis*, meaning “before the forum,” emphasizing the significance of presenting expert knowledge in legal contexts (Hemanth et al., 2020).

Other ancient civilizations, such as the Egyptians, made notable contributions. Their practice of mummification, dating back to 3000 BCE, was an early attempt to understand the decomposition process and preserve bodies. This knowledge of anatomy and decomposition indirectly contributed to what we now recognize as forensic science (Hemanth et al., 2020). The Greeks further advanced the field by examining poisons and their effects on the human body, marking a precursor to forensic toxicology (Maras & Miranda, 2014).

An interesting case from the Middle Bronze Age involves Prophet Joseph (Yusuf), lived approximately between the 18th and 17th centuries BCE, and his brothers, recorded in the Qur'an. Driven by jealousy, his brothers faked his death to deceive their father, Prophet Jacob (Yaqub). The brothers lured Joseph into a secluded area under the guise of a simple outing. Once there, they threw him into a dry well, intending to leave him for dead. And was later found by passing traders heading to Egypt and sold him to the King of Egypt. To cover up their actions, the brothers resorted to fabricating evidence of Joseph's death by presenting his shirt, stained with goat's blood, as false evidence of his death. This use of blood-stained clothing was an early example of forensic manipulation—misleading visual evidence intended to convince an observer of a fabricated event. As recounted in Surah Yusuf (12:16-18), they claimed a wolf had devoured him. This story highlights an early attempt at manipulating forensic evidence through visual misdirection, which, if subjected to modern forensic science, could have been disproven by analyzing the blood on Joseph's shirt. This is an example of early forensic misdirection that mirrors how

physical evidence has been used historically to deceive. As recounted in Surah Yusuf (12:16-18), the brothers said, "O our father, indeed we went racing each other and left Joseph with our possessions, and a wolf devoured him. But you would not believe us, even if we were truthful."

As forensic knowledge grew, new methodologies emerged during the 19th century, particularly in fingerprint analysis and ballistics. Sir Francis Galton developed the first fingerprint classification system, proving that fingerprints are unique to each individual. This breakthrough laid the groundwork for fingerprint analysis to be used as crucial legal evidence (Galton, 1892). Similarly, Henry Goddard's work in ballistics linked a bullet to a murder weapon for the first time, advancing criminal investigations by tying physical evidence directly to perpetrators (Turvey, 2022).

In contemporary forensic science, DNA analysis stands out as one of the most groundbreaking developments. It allows for precise identification of suspects by analyzing biological materials such as blood, hair, and skin cells. DNA profiling has revolutionized criminal justice by enabling the identification of perpetrators and the exoneration of wrongly accused individuals (Butler, 2015). This technique has been particularly effective in solving cold cases and even historical crimes that had long been considered unsolvable. Edmond Locard's Exchange Principle, formulated in the early 20th century, plays a central role in modern forensic investigations. Locard proposed that "every contact leaves a trace," emphasizing the importance of collecting and analyzing trace evidence such as fibers, hair, and skin cells from crime scenes. This principle laid the foundation for modern

evidence collection techniques and continues to influence crime scene reconstruction and perpetrator identification through the examination of minute traces (Turvey, 2022).

In addition to solving crimes, forensic science now plays a significant role in distinguishing between human populations. Mitochondrial DNA (mtDNA) and autosomal DNA markers are crucial tools for tracing ancestry and understanding population genetics. These techniques are particularly useful in forensic anthropology for identifying human remains, especially in mass disaster situations, where establishing the victims' origins is vital for identification purposes (Budowle et al., 2003).

Forensic genetics also contributes to studies on human migration and genetic diversity, helping to track the movements and interactions of civilizations over time. By analyzing ancient DNA and identifying unique genetic markers, forensic scientists reconstruct population histories, revealing relationships between ancient and modern groups. These genetic distinctions are pivotal for understanding how populations evolved, migrated, and mixed over millennia (Slatkin & Racimo, 2016).

Forensic science has come a long way since its early applications in ancient Rome, Greece, and Egypt (McCrery, 2013). Its evolution from basic techniques to cutting-edge DNA analysis and advanced forensic methodologies reflects the field's growing importance in both criminal justice and anthropology. Modern forensic science not only solves crimes but also aids in distinguishing human populations, shedding light on our collective history. With continuous advancements in technology, forensic science is poised to become even

more accurate and effective in solving crimes and contributing to the understanding of human evolution.

The present doctoral study is related mitochondrial DNA Analysis in selected populations in the United Arab Emirates and as such it is important to introduce the geography of the Arabian Peninsula.

1.2.The Arabian Peninsula

The history of Arab tribes is extensive and encompasses a wide range of events and developments over the centuries (Teebi and Teebi, 2005). It is very relevant to understand the historical events that contributed to shape the genealogy of the Arabian Peninsula and the genetic foundation of the United Arab Emirates.

During the pre-Islamic era, the Arabian Peninsula was home to various tribes. These tribes had diverse social structures and were involved in trade, agriculture, and pastoralism. Among them were the Thamud, 'Ad, Lihyan, and many others. Migration was common among Arab tribes, with some venturing beyond the Peninsula while others settled in oases, establishing prosperous trading hubs. Notably, Mecca and Medina were initially founded by different Arab tribes (Hoyland, 2001).

The advent of Islam in the 7th century brought significant changes to the Arabian Peninsula. Prophet Muhammad, a member of the Quraysh tribe from Mecca, played a pivotal role in establishing the Islamic faith. The Quraysh, holding control over the Kaaba, held a prominent position during this time (Holland, 2012). Following Muhammad's death, the Rashidun Caliphate emerged, led by the four Rightly Guided Caliphs: Abu Bakr, Umar,

Uthman, and Ali. Under their leadership, Arab Muslim armies expanded their territories, assimilating vast regions into the Islamic realm (Lapidus, 2002).

In 661 CE, the Umayyad Caliphate shifted the Islamic capital to Damascus. The Umayyads extended their influence across the Arabian Peninsula, North Africa, Spain, and parts of Central Asia. They promoted Arabic as the official language and facilitated the dissemination of Arab culture (Bauer, 2010). However, in 750 CE, the Abbasid Caliphate overthrew the Umayyads and established their capital in Baghdad. This era witnessed a flourishing of Arab-Islamic civilization, marked by advancements in science, literature, and administration. Arab tribes continued to hold significance within the Abbasid Empire, maintaining their tribal identities and social structures (Starr, 2013).

As the Islamic empires declined, tribal confederations led by influential families emerged. These tribes often engaged in conflicts and alliances with neighbouring tribes as well as regional powers.

In the 14th century, Arab genealogists divided Arabs into three main groups. The first group, known as Al-Arab al-Ba'ida or "The Extinct Arabs," consisted of ancient tribes that dated back to prehistoric times (Rogan, 2009). Examples of these tribes included 'Ād, Thamud, Tasm, Jadis, Imlaq (including branches of Banu al-Samayda), and others. Historical records suggest that the Jadis and Tasm tribes were reportedly wiped out through acts of genocide. Archaeological excavations have uncovered inscriptions referring to 'Iram, a once prominent city of the 'Aad tribe (Ibn Jarir al-Tabari, 1987). The second group, called Al-Arab al-Ariba or "The Pure Arabs," were the descendants of

Qahtanite Arabs. Qahtan is regarded as the ancestor of the Southern Arabian tribes. Qahtan's lineage is traced back to the biblical figure Joktan (Yoktan), who was a descendant of Prophet Noah (Firestone, 1990). The Southern Arabian tribes, such as the tribes of Yemen and Oman, consider themselves to be the descendants of Qahtan. Qahtan's descendants are believed to have migrated from Southern Arabia to various parts of the Arabian Peninsula, including the southern and eastern regions (Larsson, 2003). The third group, known as Al-Arab al-Mustarabah or "The Arabized Arabs," were also referred to as the Adnanite Arabs. Adnan is believed to be the ancestor of the northern Arabian tribes, including those who settled in the region known as Bilad al-Sham (Greater Syria) and the Northern areas of the Arabian Peninsula. Adnan is considered a direct descendant of Ismail (Ishmael), the son of Prophet Ibrahim (Abraham) from his second wife, Hagar. According to the genealogical accounts, Adnan is the father of Ma'ad, who is considered the progenitor of many Arab tribes, including the Quraysh tribe to which Prophet Muhammad belonged. They traced their lineage back to Ismail (Ishmael), the firstborn son of the patriarch Abraham. The lineage of the Adnanites is often traced through biblical genealogy. According to Arab tradition, the Adnanites are called Arabized because it is believed that Ishmael initially spoke Aramaic and Egyptian before learning Arabic from a Qahtanite Yemeni woman whom he married. Thus, the Adnanites are considered descendants of Abraham (Parolin, 2009).

The genealogy of the Arabian Peninsula is a complex and fascinating subject, closely tied to the region's history, culture, and tribal identities. Arab genealogy traditionally traces

ancestral lines through male patrilineal descent, emphasizing the importance of tribal affiliations. While genealogy holds cultural and social significance, it is important to note that the accuracy of ancient lineages can be challenging to establish due to the lack of written records and the fluid nature of tribal identities (Nebel, 2002). Previously mentioned Adnanite and Qahtanite Divisions, Arab tribes are further organized into clans and sub-tribes, each tracing their lineage back to a common ancestor. These subdivisions often possess their own customs, traditions, and territorial boundaries.

The genetic makeup of the Arabian Peninsula is diverse due to historical interactions, migrations, and trade routes. The Arabian Peninsula has been a crossroads of human migration and trade for thousands of years. Historical influences from neighbouring regions such as the Levant, Mesopotamia, Persia, and Africa have contributed to the genetic diversity in the area. While Arab tribes have generally maintained a strong sense of ancestral identity, genetic studies have revealed the presence of genetic admixture from various populations (Alshamali et al., 2009; Fernandes et al., 2015; Kousathanas et al., 2017; Al-Meer et al., 2019).

The study by Elliott et al. (2022) provides a crucial analysis of the genetic landscape of the UAE population, highlighting the influence of endogamy and consanguinity on genetic homogeneity and diversity. Despite global influences from historical migrations and trade, genetic studies in the Arabian Peninsula have been sparse. This research, analyzing 1,198 individuals with 1.7 million genetic markers, identified significant genetic homogeneity and long homozygosity tracts, aligning with the region's high consanguinity rates. The

study employed haplotype-based algorithms and admixture analyses to delineate gene flows predominantly from the Middle East, with some influences from Africa and South Asia. It also noted a strong intra-Emirati kinship, reflecting deep-rooted tribal connections, contrasting with more diverse inter-Emirati relationships.

1.3.The formation of the United Arab Emirates

The United Arab Emirates (UAE), also referred to as the Emirates, is a nation situated in the Eastern region of the Arabian Peninsula, positioned along the South-Eastern coastline of the Arabian Gulf and the North-Western coastline of the Gulf of Oman. Six Emirates (Abu Dhabi, Dubai, Sharjah, Ajman, Umm Al Quwain, and Fujairah) united on December 2, 1971, establishing the UAE as a federation (TEN Guide., 1972), while the seventh emirate, Ras al Khaimah, joined the federation on February 10, 1972 (Smith, 2004). Formerly known as the Trucial States, these seven sheikdoms entered into treaty relations with the British in the 19th century. The British announced their intent to withdraw from the region in 1968, leading to discussions among the Emirates for a united entity (Heard-Bey, Frauke,2005).

Archaeological discoveries in the UAE provide evidence of human habitation, transmigration, and trade spanning over 125,000 years (EarthSky, 2011). The region was previously inhabited by the Magan people (around 1300 BCE), who were known to the Sumerians and engaged in trade with coastal towns, as well as bronze miners and smelters from the interior (Bhacker and Bhacker, 1997). While there is no exact end date provided in the typical historical sources, the period of the Magan people's significance roughly

concludes by the end of the Bronze Age. The historical significance of trade with the Harappan culture of the Indus Valley is apparent from the discovery of jewelry and other artifacts, and there is substantial early evidence of trade with Afghanistan, Bactria, and the Levant (Abed and Hellyer, 2001).

Throughout the three distinct Iron Ages and the subsequent Hellenistic period, the UAE maintained its role as a crucial coastal trading hub (Abed and Hellyer, 2001). Following the spread of Islam in the 7th century, triggered by the response of the Al Azd tribe to the message of Muhammad, the region became Islamized (Hoyland, 2015). This transformation was solidified through the Ridda Wars and the decisive Battle of Dibba (632-633). The Islamic era witnessed the resurgence of region that now includes the UAE as a vital trade center. Through the ports of Julfar, Dibba, and Khor Fakkan, a connection to the extensive Eastern Arab trading network centered around the Kingdom of Hormuz, which served as a significant role in Arab trade dominance between the East and Europe (Hoyland, 2001). In the late Islamic era, small trading ports emerged alongside the development of agricultural oases like Liwa, Al Ain, and Dhaid, while tribal Bedouin societies coexisted with settled populations in the coastal areas (Heard-Bey and Frauke, 2004).

The Portuguese incursions and battles along the coast, led by Afonso de Albuquerque, disrupted Arab trade networks, and resulted in a decline in trade, as well as increased regional conflicts after the fragmentation of Hormuzi authority. Subsequent conflicts between the maritime communities of the Trucial Coast and the British culminated in the

sacking of Ras Al Khaimah by British forces in 1809 and 1819, leading to the signing of the first of several British treaties with the Trucial Rulers in 1820 (Lorimer and John, 1915). These treaties, starting with the General Maritime Treaty of 1820, fostered peace and prosperity along the coast, supporting a thriving trade in high-quality natural pearls and reviving regional commerce. In 1892, another treaty granted the British control over external relations in exchange for protectorate status (Heard-Bey and Frauke, 2005).

In early 1968, the establishment of a Federation was attributable to British decision to withdraw from their involvement in the Trucial States (Heard-Bey and Frauke, 2005). Upon receiving news of British withdrawal from Labour MP Goronwy Roberts, Sheikh Zayed and the nine Arabian Gulf sheikhdoms; including Bahrain and Qatar; embarked on an endeavor to establish a federation of Arab emirates (UAE Ministry of Presidential Affairs. (n.d.)). This agreement was initiated between two influential Trucial Rulers: Sheikh Zayed bin Sultan Al Nahyan of Abu Dhabi and Sheikh Rashid bin Saeed Al Maktoum of Dubai, and the concept of union was proposed in February 1968. An agreement in principle was made when they both met in the desert location of Argoub El Sedirah. Ultimately, inviting other Trucial Rulers to join the Federation, where a public announcement expressing their intent to form a coalition and extended an invitation to other Arabian Gulf states to join (Maktoum and Mohammed, 2012). Subsequently, a summit meeting took place later that month, attended by the rulers of Bahrain, Qatar, and the Trucial Coast. During this meeting, the government of Qatar suggested the formation of a Federation of Arab Emirates to be governed by a higher council composed of the nine

rulers. The declaration of the union proposal was accepted, and later announced. Despite that, various disagreements among the rulers rose regarding issues such as the capital's location, the drafting of the constitution, and the allocation of ministries (Zahlan, 1979). Overall, there were only four meetings held among the nine rulers.

The final meeting, which took place in Abu Dhabi, resulted in the election of Zayed bin Sultan Al Nahyan as the first President of the Federation. The gathering encountered several impasses, including the appointment of a vice-president, the defense strategy of the Federation, and the necessity of a constitution (Zahlan, 1979). The session's outcome revealed British Government's vested interests, urging Qatar's withdrawal due to perceived foreign interference in internal affairs. Consequently, the nine-emirate federation was terminated, and in succession, Bahrain achieved independence in August 1971. Consecutively, Qatar declared its independence in September 1971 (Kelly, 2016).

By December 1, 1971, the British-Trucial Sheikhdoms treaty expired, and the Trucial States became independent sheikhdoms. Four additional Trucial States, namely Ajman, Sharjah, Umm Al Quwain, and Fujairah, joined Abu Dhabi and Dubai in signing the UAE's founding treaty, which had a draft constituted in record time to meet the deadline of December 2, 1971. On that date, at the Dubai Guesthouse (now known as Union House), the Emirates agreed to form a union called the United Arab Emirates, which is the date the UAE celebrates the national day. Ras al-Khaimah joined the Federation in February 1972 (Smith and Simon, 2004).

Today, the UAE is a relatively recently developed country with a total population of 12.50 million as of 2024. Emiratis constitute around 10% of the population, approximately 1.25 million, according to the latest statistics from the Dubai Government (Dubai Statistics Center, 2024.). The expatriate population stands at 11.06 million, with Indians comprising 4.75 million (38%) and Pakistanis 1.80 million (14.4%). The other

The geographical location of the UAE falls on the Arabian Gulf sharing major boundaries with both Oman and Saudi Arabia (Hooglund and Toth, 1994). Figure 1.1 shows a map of UAE along with the borders of the seven cities. Figure 1.2 shows the geographical location of the UAE on the world map.

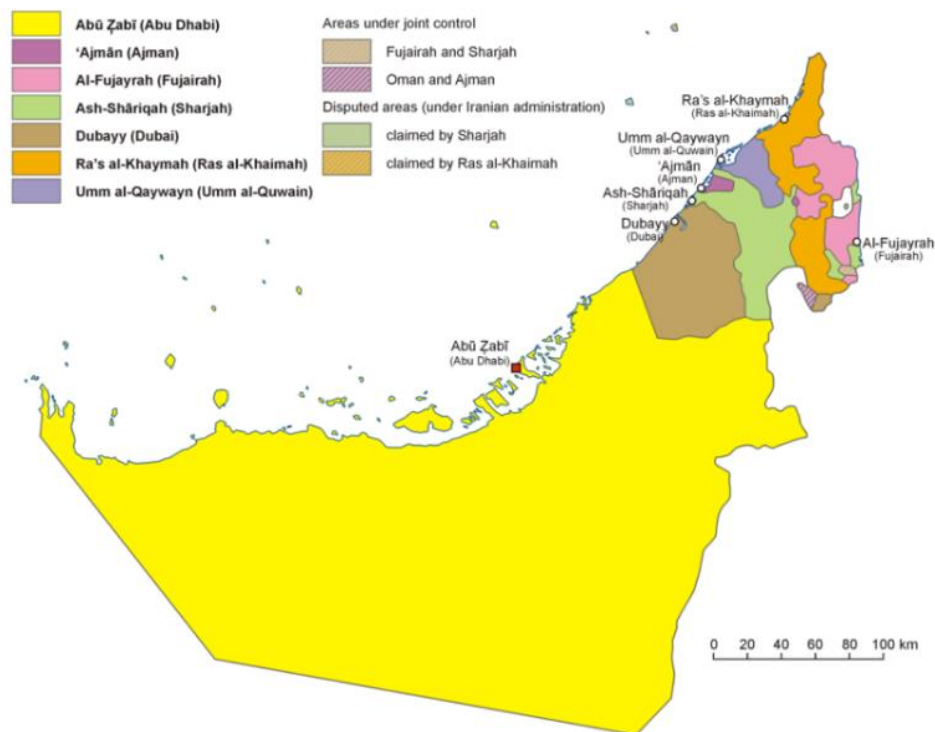


Figure 1.1.A map showing the borders of each city in the UAE (Taken from Abed and Hellyer, 2001).

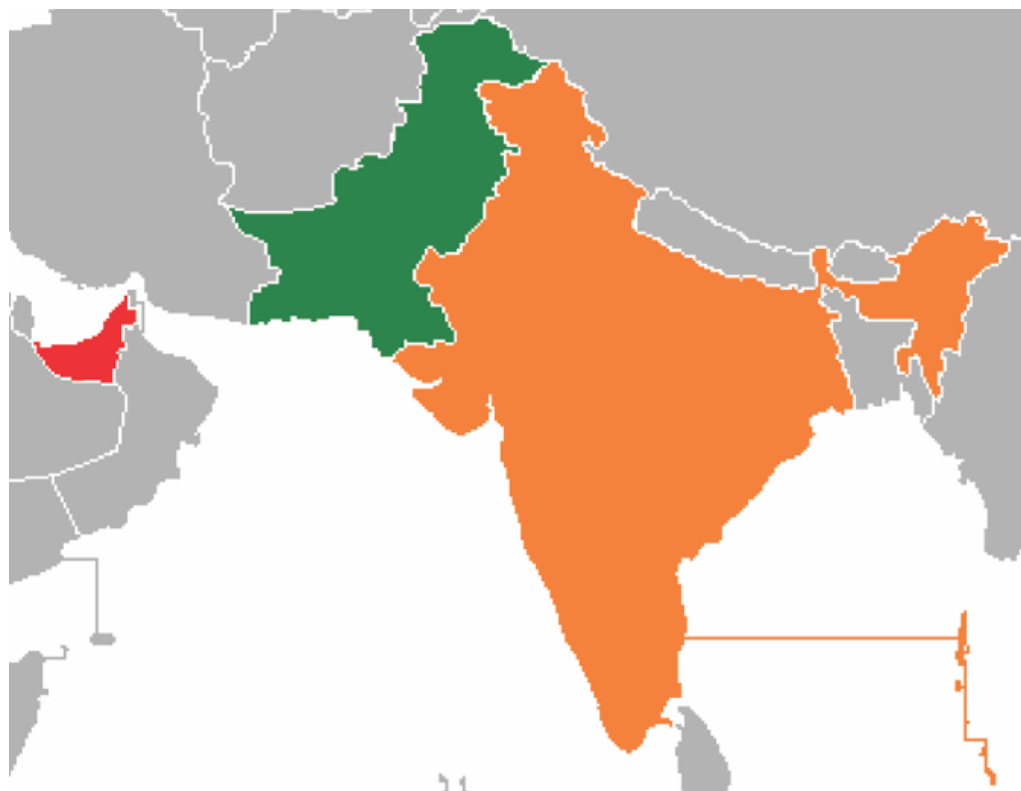


Figure 1.2. Geographical location of the UAE (red) on the world map relative to Middle East and East Asia, highlighting Pakistan and India (Taken from Grubb, 2007).

1.4.A brief history of Forensic Genetics

Forensic genetics, a discipline within the field of genetics, employs DNA analysis and comparison techniques to resolve legal issues (Butler, 2005). Its origins can be traced back to the early 20th century when Karl Landsteiner discovered the human ABO blood group polymorphism (Landsteiner, 1900). In the 1920s scientists began utilizing serological and protein electrophoretic methods to study the diversity of blood groups and polymorphic proteins. However, in a forensic context these methods were limited by the requirement for substantial amounts of biological material that were susceptible to bacterial contamination, rapid degradation, and were only relevant with certain body fluids. The

evidential value was generally low providing match probability ranging from 0.01-0.001 when applied to bloodstain analysis using eight different systems (Jobling and Gill 2005).

A pivotal moment in the field of forensic genetics occurred in 1984 when Sir Alec Jeffreys discovered hypervariable loci, also known as minisatellites or Variable Number Tandem Repeats (VNTRs). Using MultiLocus Probes (MLP), Jeffreys generated multi-band patterns that were initially known as DNA fingerprints (Jeffreys et al. 1985). This technique involved the digestion of DNA by restriction enzymes, followed by Southern blot analysis. However, the DNA fingerprinting method had limitations, including the need for large quantities of good-quality DNA, which posed challenges for practical implementation (Roewer, 2013).

The field of molecular biology, including forensic genetics, underwent a revolutionary transformation with the advent of the polymerase chain reaction (PCR) in 1983, developed by Kary Mullis (Mullis et al. 1986a). PCR enabled the amplification of minuscule quantities of DNA, even if they were degraded, which frequently occurs in forensic casework.

Over time, the older methods were gradually replaced with PCR-based approaches that amplified microsatellites, also known as Short Tandem Repeats (STRs). These repetitive and highly polymorphic DNA sequences offered both high sensitivity and discriminative power, making them well-suited for forensic applications (Roewer, 2013).

1.4.1. DNA Profiling

DNA profiling, is a forensic method used to identify individuals based on their unique genetic features. This technique has significantly transformed forensic science, law enforcement, and ancestry verification.

The methodology of DNA profiling enabled the identification of individuals with high degree of accuracy. Over the years, DNA profiling has experienced substantial technological evolution, shifting from methods like restriction fragment length polymorphism (RFLP) to more advanced techniques including short tandem repeat (STR) analysis using polymerase chain reaction (PCR) amplification.

1.4.2. STRs

STRs are characterized by DNA sequences with short repeating segments, typically comprising 2 to 6 bp. They are further categorized into different classes based on the number of nucleotides in the major repeat unit, spanning from mononucleotide to hexanucleotide repeats, with dinucleotide repeats being the most prevalent (Fan and Chu, 2007). Moreover, STRs are classified based on their structure into simple repeats, compound repeats, complex and hypervariable repeats.

In 1994, the United Kingdom Forensic Science Service (FSS) initially analysed four STR markers for profiling, constituting a first-generation quadruplex (STR loci); by 1995, they had incorporated six STR markers into routine analysis, and the sex-typing marker amelogenin.

In 1997, the FBI introduced The National DNA Index System (NDIS) and selected 13 autosomal STR loci for inclusion in CODIS, a database containing profiles obtained from forensic labs at various levels of government. This database allows forensic laboratories to compare DNA profiles with relevant cases or previously convicted offenders. As of February 2024, the National DNA Index (NDIS) contains records for over 17,026,097 offender profiles, 5,391,074 arrestee profiles, and 1,322,543 forensic profiles. CODIS has produced over 698,183 hits, assisting in more than 680,122 investigations (FBI, 2024). The 13 core STR markers were: TH01, vWA, FGA, D8S1179, D18S51, D21S11, CSF1PO, TPOX, D3S1358, D5S818, D7S820, D13S317, and D16S539. The nomenclature of a typical STR locus, like D21S11, follows a structured format: 'D' represents DNA, '21' signifies the chromosome number, 'S' indicates the presence of a single copy sequence, and '11' serves as a unique identifier (Parson et al., 2016).

STR loci fall into different categories: some are simple repeats, such as TPOX, CSF1PO, D5S818, D13S316, and D16S539, characterized by repeating the same nucleotide sequence, while others, like TH01, D18S51, and D7S820, are simple repeats with non-consensus alleles, featuring variations in the nucleotide sequence. Compound repeats with non-consensus alleles, such as vWA, FGA, D3S1358, and D8S1179, contain different repeat units. Additionally, D21S11 is an example of a complex repeat. Among the original dataset of 13 CODIS STR loci, markers FGA, D18S51, and D21S11 exhibit the greatest variability (polymorphism) across individuals, but variation is also observed in TPOX, CSF1PO, and TH01. When comparing all thirteen core loci in CODIS, the likelihood of a

random match averages below one in a trillion for unrelated individuals (Butler, 2012). In January 2017, CODIS profiles were expanded to include a total of 20 loci along with sex markers (FBI, n.d.).

The utilization of STR markers in forensic DNA profiling serves multiple purposes, including the identification of missing persons, verification of familial relationships, and matching of suspects to crime scenes. Y-chromosome STR markers are commonly employed in sexual assault and paternal kinship cases, given their male-specific nature. The identification of numerous STR markers on the Y chromosome has led to the development of kits that amplify these male-specific loci (Hill et al., 2007).

When using STR in genetic profiling, one significant challenge arises when the profile itself is complex, such as in the case of mixed, partial, or low template DNA profiles (Cerri et al., 2003).

1.5.Mitochondria

Mitochondria (single = mitochondrion) are double-membraned cytoplasmic organelles present in the cytosol of all nucleated mammalian cells (Figure 1.3) (Butler, 2012). The freely permeable outer membrane forms the outer surface of the mitochondrion, which is separated by an intermediate space from the largely impermeable, highly invaginated and structurally distinct cristae of the inner membrane. The outer membrane and the intermembrane space are considered as the mitochondrial outer compartment, while the inner membrane and the matrix space constitute the mitochondrial inner compartment. The primary

function of the mitochondria is the production of energy, which is carried out through the oxidative phosphorylation (OXPHOS) and the electron transport chain (Pfeiffer et al. 2001; Dykens and Will 2008). It also plays part in the intracellular signaling and apoptosis (programmed cell death) (Joza et al. 2001).

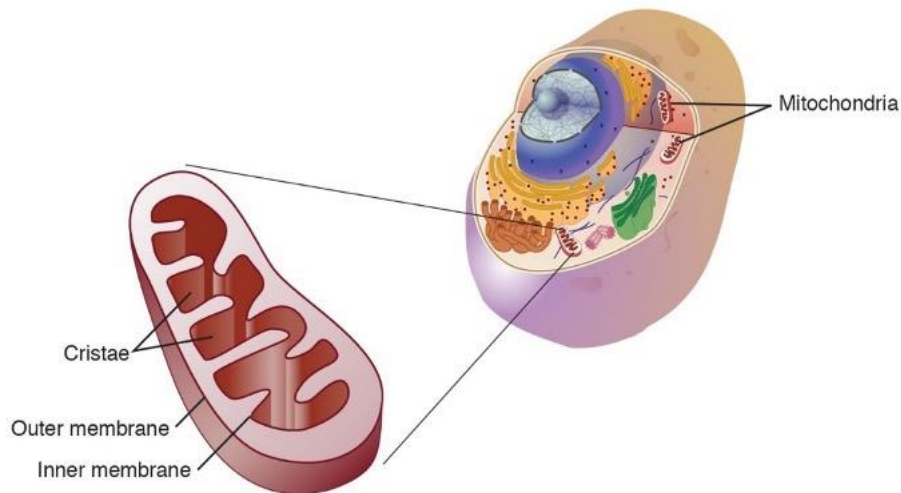


Figure 1.3. Illustrative image showing the position of mitochondria inside the biological cell and their general structure (National Human Genome Research Institute, 2023).

1.5.1. Mitochondrial DNA

The endosymbiotic theory was first articulated in the early 20th century but gained significant traction in the 1960s through the pioneering work of biologist Lynn Margulis. The theory challenges the idea that complex cells (eukaryotes) evolved solely through mutations and the gradual accumulation of complexity within ancient single-celled organisms. The origin of mitochondria, dating back to around 1.45 billion years ago (Martin and Mentel, 2010), is intricately linked to the emergence of eukaryotes during a critical phase of Earth's history (Embley and Martin, 2006). This period, extending from approximately 1.8 billion years ago to about 580 million years ago, was characterized by

largely anoxic oceans due to the prevalence of H₂S-producing bacteria. These environmental conditions significantly influenced the evolutionary trajectory of early eukaryotes, prompting them to develop anaerobic energy-producing pathways within their mitochondria to adapt to the low-oxygen conditions (Boxma et al., 2005). This evolutionary adaptation highlights the deep interconnection between mitochondrial development and Earth's environmental changes, underscoring the complex history of cellular life's evolution (Dyall et al., 2004).

Over evolutionary time, these endosymbiotic α -proteobacteria lost their ability to survive independently and retained a portion of their associated DNA (Andersson et al., 1998; Dolezal et al., 2006). Consequently, the endosymbiont transformed into mitochondria, and its genome became mtDNA, playing a crucial role in cellular energy production through oxidative phosphorylation and ATP synthesis (Timmis et al., 2004).

Supporting the endosymbiotic theory are several lines of evidence, including the presence of a membrane structure resembling gram-negative bacteria, circular double-stranded DNA lacking intron spacers, autonomous replication, polycistronic genes, and the existence of specific enzymes (Andersson et al., 1998).

The mtDNA is a small circular DNA molecule with approximately 16,569 base pairs (bp) but with variable lengths due to small insertions or deletions (Anderson et al., 1981). The coding region encompasses 37 genes, including 13 proteins involved in respiratory enzyme activity for oxidative phosphorylation and ATP production. For instance, at positions 514 to 524, there are varying numbers of dinucleotide repeats, commonly observed as

ACACACACAC or (AC)₅, but ranging from (AC)₃ to (AC)₇ in certain individuals (Szibor et al., 1997). The two strands of mtDNA are known as the “heavy strand” (H-strand) and the “light strand” (L-strand). The H-strand is rich in purine nucleosides (adenosine and guanosine) with a higher molecular weight, while the L-strand is rich in pyrimidine nucleosides (thymidine and cytidine) (Borst & Grivell, 1981).

mtDNA constitutes approximately 0.25% of a cell's total DNA and is divided into a small control region (CR) accounting for 6.8% and a large coding region comprising 93.2%. The CR, also known as the displacement loop or D-loop, serves as the initiation site for mtDNA replication and exhibits a three-stranded DNA loop structure visible through electron microscopy and atomic force microscopy during mitochondrial transcription and replication (Brown, 2010). Additionally, it contains 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNAs), specifically the 12S and 16S subunits, facilitating the decoding and translation of mtRNA into mitochondrial proteins (Anderson et al., 1981).

Figure 1.4 presents the mitochondrial genome, while the control region (~1,150 bp) of human mtDNA is highly polymorphic and extensively used in forensic and population genetics studies, depicted in Figure 1.5. It comprises three hypervariable regions known as HV1 (position 16,024–16,383), HV2 (position 57–372), and HV3 (position 438–574) (Lutz et al., 2000). Some population genetic studies encompass broader hypervariable regions, particularly HV1 (16,024–16,400) and HV2 (44–340), to capture phylogenetically significant positions (Brandstatter et al., 2004b).

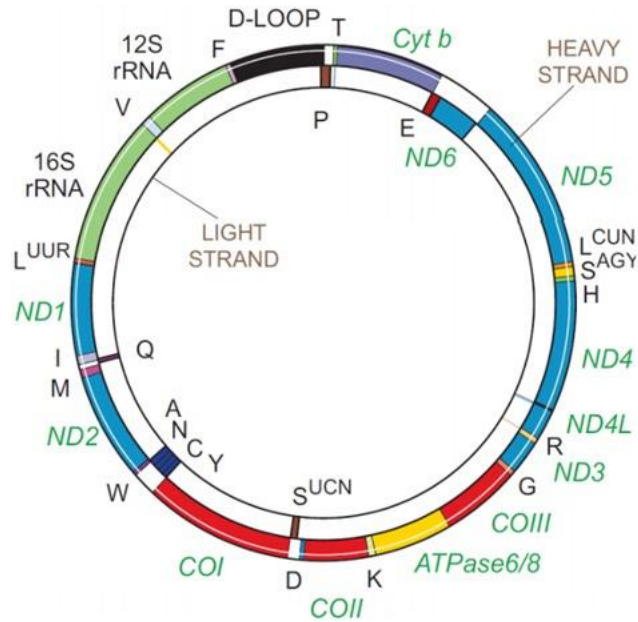


Figure 1.4. Diagram showing mitochondrial circular DNA structure (Zeviani et al. 2007).

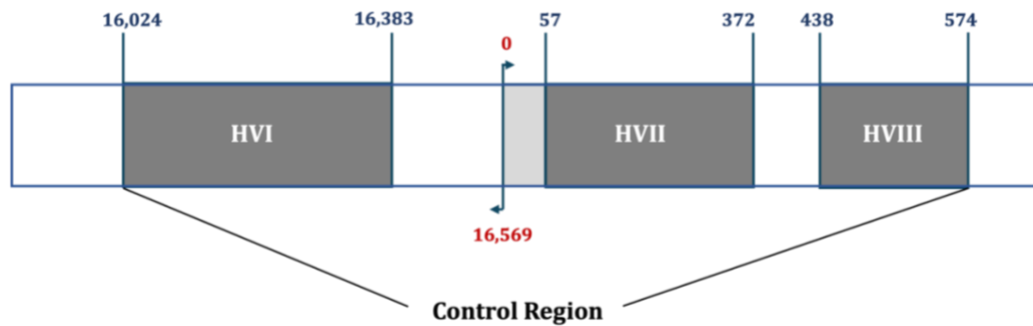


Figure 1.5. Expanded control region with hypervariable region indicated (Lutz et al., 2000).

1.5.2. Inheritance pattern of mtDNA

The inheritance pattern of mitochondrial DNA (mtDNA) is maternal, wherein mothers pass on their mtDNA to their offspring, and only female individuals can transmit it further (Sato and Sato, 2012). This can be attributed to two factors. Firstly, during fertilization, if paternal mitochondria present in the sperm tail enter the oocyte, they become diluted by the high number of maternal mitochondria (approximately 100,000) within the oocyte. As a result, the contribution of paternal mtDNA to the offspring is negligible (Manfredi

et al., 1997). Secondly, ubiquitination plays a role in the elimination of sperm mitochondria. Proteins specific to sperm mitochondria are tagged with ubiquitin, marking them for subsequent disposal through an elimination pathway. The precise timing of this elimination pathway, whether it occurs at the time of fertilization or thereafter, remains uncertain (Manfredi et al., 1997).

The haploid nature of the mitochondrial genome, determined by this maternal inheritance pattern, holds significant forensic relevance. It allows for the shared mtDNA sequence among relatives within the same maternal lineage. For instance, all siblings, regardless of sex, will typically possess the same mtDNA sequence as their mother. This characteristic proves valuable in confirming the identity of missing individuals, investigating cases of mass disasters, and studying population variations (Butler, 2012; Holland et al., 2013).

Parental leakage in mitochondrial DNA (mtDNA) inheritance refers to the rare phenomenon where mtDNA from both parents, rather than just the mother, is passed on to offspring (Luo et al., 2018). This anomaly occurs when paternal mtDNA somehow bypasses the usual mechanisms that prevent it from entering the oocyte during fertilization. The precise mechanisms behind paternal leakage are not fully understood and are the subject of ongoing research. Such instances are considered exceptionally rare (Ruis et al., 2019) but have been documented in humans and more frequently in some other animal species.

1.5.3. Copy Number

A key characteristic of mitochondrial DNA (mtDNA) that enhances its utility in forensic analysis is its abundance in cells. Unlike nuclear DNA, which is present in just two copies per nucleated cell, mtDNA can range from hundreds to thousands (Butler, 2012). On average there are 4 to 5 copies of mtDNA molecules per mitochondrion with a measured range of 1 to 15 (Sato & Kuroiwa 1991). This abundance offers significant benefits for forensic investigations, particularly when samples contain only minute quantities of DNA that are insufficient for traditional nuclear DNA analysis techniques. Moreover, the high copy number of mtDNA in cells enhances its detectability, especially in degraded samples, which is often the case in ancient DNA studies (Bogehagen, 2012; Cole, 2016; Robin and Wong, 1988; Naue et al., 2024).

1.5.4. Heteroplasmy

Heteroplasmy refers to the coexistence of multiple mitochondrial DNA (mtDNA) sequences either within a single cell or among cells within an individual. In the field of forensics, heteroplasmy serves as an additional discriminatory factor in the identification of human remains (Ivanov et al., 1996). It is widely recognized and routinely considered in forensic casework analysis (Melton, 2004; Melton et al., 2005). Although heteroplasmy is not an uncommon biological state, often it falls below the limits of detection in standard DNA sequence analysis (Comas et al., 1995; Bendall et al., 1997; Tully et al., 2001). Heteroplasmy can manifest in various ways, including the presence of multiple mtDNA types within a single tissue, different mtDNA types in different tissues, or a combination

of homoplasmy (presence of a single mtDNA type) and heteroplasmy in different tissue samples from the same individual (Carracedo, 2005).

Several factors contribute to the occurrence of heteroplasmy. These include the high copy number of mtDNA in eukaryotic cells, the relatively high mutation rate of the mtDNA control region which is about 10 times higher than that of the nuclear DNA genome. Heteroplasmy is observed in two forms: Single Nucleotide Heteroplasmy (SNH) or also known as Point Heteroplasmy (PH) and Length Heteroplasmy (LH). LH is the more prevalent form and is typically found in the control region of the mtDNA genome, particularly within long stretches of GC base pairs (Forster et al., 2010). LH is often caused by slippage of the replicative polymerase during DNA replication. Two main hotspots for LH are located at positions 16,184–16,193 in HV1 and 303–315 in HV2. These length variations can lead to difficulties in interpreting data obtained through conventional Sanger-based DNA sequencing methods, resulting in out-of-reading-frame or uninterpretable data (Melton, 2004).

Although the high mutation rate has been attributed to the low fidelity of the γ -polymerase involved in mtDNA replication, studies suggest that this replicative DNA polymerase is actually quite accurate (Longley et al., 2001). Mutations in mtDNA, resulting in SNH, can also occur due to DNA damage induced by Reactive Oxygen Species (ROS) generated during oxidative phosphorylation (Connell et al., 2022) and suboptimal repair processes (Yakes and VanHouten, 1997). The primary repair pathway for ROS-induced damage is the base-excision repair pathway (Driggers et al., 1993). As

mitochondrial DNA lacks the nucleotide excision repair mechanism and has attenuated double-strand break repair systems such as homologous recombination and non-homologous end joining pathways, mutations in mtDNA, resulting in SNH, can occur (Sykora et al., 2012).

1.5.5. Nuclear Mitochondrial DNA Segments (NUMT)

NUMTs, are intriguing genetic sequences that originate from mitochondria but are found within the nuclear genomes of eukaryotic organisms. These segments are essentially mitochondrial DNA that has been transferred to the nuclear genome over the course of evolutionary history. This transfer occurs through cellular processes where fragments of mitochondrial DNA are mistakenly incorporated into nuclear DNA, typically during repair of nuclear DNA damage. The presence of NUMTs in nuclear genomes is a widespread phenomenon observed across various eukaryotic species, illustrating a common genetic occurrence that bridges mitochondrial and nuclear genetic material (Hazkani-Covo et al., 2010).

Distinguishing NUMTs from genuine mitochondrial DNA (mtDNA) is crucial for accurate data interpretation, particularly in forensic contexts where precision is paramount. Recent advancements in sequencing technologies and bioinformatic methods have improved our ability to identify and filter out NUMTs, thereby enhancing the reliability of mtDNA analysis in degraded or mixed samples (Marshall et al., 2021). These developments ensure that forensic scientists can more accurately interpret mtDNA data, reducing the risk of misidentification and improving the overall quality of forensic genetic investigations.

NUMTs can impact genetic studies and diagnostics due to their similarity to mitochondrial DNA. They can lead to misidentification in genetic assays and have implications for understanding genetic diseases and evolutionary biology. The study of NUMTs offers insights into the mechanisms of DNA repair and transfer within cells and across generations, illuminating the evolutionary dynamics of genomes.

Woerner et al. (2020) introduced a novel tool, called Remove the NUMTs! (RtN!). This was designed to improve the accuracy of mitochondrial genome sequencing by identifying and removing NUMTs, which complicate variant calling and data interpretation. The study involved sequencing the whole mitochondrial genomes of 270 individuals using the Precision ID mtDNA Whole Genome Panel on the Ion S5 system, followed by read mapping to the revised Cambridge Reference Sequence and variant calling with Converge software. RtN! categorizes and filters reads based on sequence similarity to a large database of mitochondrial genomes, effectively removing NUMTs and low-frequency noise without impacting true mitochondrial variations. The tool demonstrated high efficiency in retaining correct mitochondrial haplotypes and reducing false positives in both single-source and mixed-source samples, thereby enhancing the reliability of mitochondrial genomics analyses. RtN! is available as an open-source application, offering significant improvements for population genetics, forensic analysis, and biomedical research applications. (Taken from Woerner, 2020. RtN. GitHub. <https://github.com/Ahhgust/RtN>).

1.6. Analysis methods development

The study of mtDNA variation has made significant progress since the pioneering work of Anderson, who first sequenced the complete human mitochondrial genome over 45 years ago. This landmark achievement led to the establishment of a reference sequence known as the Cambridge Reference Sequence (CRS) (Anderson et al., 1981). Subsequently, in 1999, a re-sequencing of the original placenta sample revealed errors and rare polymorphisms that were specific to that sample. To address these discrepancies, a revised version of the Cambridge Reference Sequence (rCRS) was introduced (Andrews et al., 1999). The rCRS now serves as the updated reference sequence for human mtDNA analysis.

1.6.1. Restriction Fragment Length Polymorphism (RFLP)

During the early 1980s mitochondrial DNA (mtDNA) analysis predominantly relied on restriction enzyme assays targeting the entire mtDNA genome (Cann et al., 1987). This technique involved the enzymatic digestion of mtDNA using specific restriction enzymes that recognize and cleave DNA at specific sites. The resulting fragments were then separated through gel electrophoresis, and the patterns of fragment lengths were examined to identify genetic variations. Known as low-resolution Restriction-Fragment Length Polymorphism (RFLP) analysis, this method typically involved a limited number of samples and a small set of enzymes (Brown, 1980). A notable study during this period investigated 21 individuals from diverse ethnic and geographic backgrounds, demonstrating the utility of mtDNA RFLP patterns in tracing human genetic history

(Cann et al., 1987). In this study, 18 restriction endonucleases were employed, while seven enzymes yielded identical fragment sizes across all 21 samples, the remaining enzymes revealed differences attributed to single base substitutions, enabling the individual characterization of each sample (Brown, 1980). Subsequently, in the mid-1980s, a higher-resolution restriction RFLP analysis was introduced, utilizing 12 or 14 enzymes, to further explore the origins of modern humans (Cann et al., 1987).

1.6.2. Polymerase Chain Reaction (PCR)

PCR allows the amplification of specific regions of mtDNA using DNA primers that flank the target sequence. This technique enables the detection and analysis of mtDNA variants even in limited DNA samples (Wilson, 1997).

1.6.3. RFLP-PCR Assays

Combining RFLP and PCR techniques allows the detection of specific mtDNA mutations. The PCR amplification step is followed by restriction enzyme digestion, and the resulting fragment patterns are analysed to identify the presence or absence of particular mutations. One of the early works of this assay was carried out by Douglas C. Wallace where he, and his team reported the use of RFLP-PCR assays to identify a specific mitochondrial DNA mutation linked to Leber's Hereditary Optic Neuropathy (LHON), an inherited form of vision loss. (Wallace et al., 1988).

1.6.4. Sanger Sequencing

The Sanger sequencing method is widely utilized for mitochondrial DNA (mtDNA) analysis due to its ability to provide detailed information about the position and base changes in each polymorphic position within the analysed region, compared to the revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999).

The application of mtDNA testing in forensic science coincided with the introduction of the Polymerase Chain Reaction (PCR) in the mid-1980s. PCR enables the amplification of specific DNA sequences, allowing for the generation of millions of copies within a short period. This breakthrough revolutionized the field of DNA analysis (Mullis et al., 1986b).

Sequencing the hypervariable segments of the control region, known as the D-loop, has proven valuable in revealing individual and population-specific variations. The high mutation rate in these non-coding regions results in significant variability within a relatively short and easily sequenced region of the mtDNA, contributing to the popularity of mtDNA in forensic studies (Mullis et al., 1986b; Salas et al., 2007).

In addition to detecting variability in the hypervariable regions (HV1, HV2 and HV3), sequencing methods can also be employed for Single Nucleotide Polymorphism (SNP) analysis in the coding region. This approach enhances the discriminatory power of mtDNA typing and facilitates the assignment of mtDNA profiles to specific haplogroups, supporting human population studies and medical research (Bogenhagen and Clayton, 1977; Alvarez-Iglesias et al., 2007).

During the early 1990s, mtDNA analysis expanded to investigate human origins in different geographic areas. Researchers conducted large-scale studies on specific continents, employing techniques such as RFLP surveys of the entire mtDNA molecule, with a focus on the coding region, or sequencing of the first hypervariable segment (HV1) in the control region, known for its rapid evolution (Macaulay et al., 1999). Forensic geneticists recognized the significance of mtDNA testing for identification purposes, especially in cases involving highly degraded biological materials and telogen hair shafts with minimal nuclear DNA content (Salas et al., 2007).

Mitochondrial DNA typing often emphasizes polymorphism in HV1 and HV2, although a fraction of mtDNA genomes sharing identical HV1 and HV2 sequences may exhibit non-identical sequences in HV3 (Lutz et al., 1997). Recent studies have increasingly included hypervariable regions HV1, HV2, and HV3, with reports of mutations in the intervening sequences of the control and coding regions of human mtDNA (Penta et al., 2001; Chomyn and Attardi, 2003). As a result, a comprehensive mtDNA sequence database covering the entire control region has become valuable.

Several guidelines issued by the ISFG and SWGDAM for mtDNA typing, recommending the sequencing of the entire mtDNA control region instead of one or two hypervariable regions in population genetic studies for forensic databases. The guidelines also stressed the importance of conducting negative and positive controls, as well as extraction reagent blanks, throughout the laboratory process (Parson et al., 2014).

1.7.Comparison of Nuclear DNA (nDNA) and mtDNA

While both contribute to an organism's overall genetic makeup, nDNA and mtDNA possess several significant differences. These differences highlight the complementary roles of mtDNA and nDNA in genetics and their distinct contributions to understanding evolutionary history and individual phenotypes. Differences in inheritance pattern, size, number of copies and furthermore. The differences are summarized in Table 1.1.

Table 1.1 Characteristics of nDNA and mtDNA (Butler, 2012).

Characteristics	Nuclear DNA (nDNA)	Mitochondrial DNA (mtDNA)
Genome Size	3.2 billion bp	16,569 bp
Copies Per Cell	1 copy per cell	Hundreds to thousands per cell
Percentage of total DNA	99.75%	0.25%
Structure	Linear	Circular
Inheritance Pattern	Maternal and Paternal	Maternal
Chromosomal Pairing	Diploid	Haploid
Generational Recombination	Yes	No
Replication Repair	Yes	No
Uniqueness	Unique	Identical within the Maternal Lineage
Mutation Rate	Low	At least 5-10 times more than the nDNA
Reference Sequence	Described in 2001 by the Human Genome Project, now GRCh38 (hg38).	rCRS (Anderson et al., 1981)

1.8.mtDNA Population Genetics

Mitochondrial DNA offers a robust framework for conducting population genetic studies, enabling researchers to delve into the maternal lineage and historical migrations of human populations. Mitochondrial DNA is particularly useful for studying population genetics

because it mutates at a relatively consistent rate, allowing researchers to estimate the timing of genetic divergences. This "molecular clock" provides insights into evolutionary events and can help track the migration patterns of ancient human populations.

The application of mtDNA in population genetics extends to understanding human evolutionary history. By analyzing variations in mtDNA sequences among different human populations, scientists can reconstruct a genetic tree that illustrates the relationships and temporal divergence among these groups. This approach has been instrumental in confirming the out-of-Africa theory of human evolution and in identifying distinct genetic populations across the globe.

1.8.1. Haplotypes

Haplotypes are specific combinations of alleles or DNA sequences that are closely linked on a chromosome and inherited together from a single parent. This term, derived from the phrase "haploid genotype", is crucial in the study of genetics because it refers to a specific combination of alleles or sequence of DNA bases that are closely linked on one chromosome and tend to be inherited as a unit. This linkage occurs because the genes or DNA sequences are located close to each other on the chromosome, reducing the likelihood of recombination during meiosis, the process that generates gametes (sperm or egg cells). As a result, these combinations are passed down across generations in blocks. Due to this linkage, the haplotype can provide a clear picture of the genetic diversity and evolutionary history within populations. Techniques such as SNP genotyping and sequencing methods are used to analyse genetic samples. By comparing these

sequences across different individuals within a population, scientists can determine which alleles tend to be inherited together, forming a haplotype. In genetic linkage studies, haplotypes are crucial for mapping genes associated with diseases. They help researchers to identify genetic markers that are co-inherited with a disease trait. This is particularly useful for tracking the inheritance patterns of complex traits that are influenced by multiple genes, as it allows for the identification of genetic loci that contribute to disease phenotypes.

Haplotypes are particularly important because they help to identify which gene variations often occur together in an individual's genome. Each person has two sets of chromosomes—one inherited from each parent—and thus, two sets of haplotypes. The phylogenetic trees is a conceptualized method for grouping haplotypes into haplogroups. Phylogenetic trees are constructed based on genetic similarities and differences observed in the haplotype data, illustrating the evolutionary relationships among different haplogroups.

1.8.2. mtDNA Haplogroups and Phylogenetic Tree

Individual haplotypes can be grouped into larger sets called haplogroups. Moving from haplotypes to haplogroups, the concept expands as haplogroups consist of a group of similar haplotypes that share a common ancestor. The concept of haplogroups as clusters of similar haplotypes that share a common ancestor, showing how genetic drift and selection can lead to the formation of distinct haplogroups. This is particularly useful in studying human evolutionary history, as haplogroups can trace the geographical and

temporal origins of different populations. The analysis of mtDNA haplogroups reveals the diversity and migration patterns of human populations. Research conducted by Cann et al., (1987) utilizing mitochondrial DNA restriction site polymorphisms concluded that modern human evolution began in East Africa approximately 150,000 – 200,000 years ago (KYA) and subsequently, radiated out of Africa around 70,000 years ago. Haplogroups are classifications of mtDNA variants characterized by specific mutations (Torroni et al., 1996). These variants trace back to ancestral lineages that migrated out of Africa, with African haplogroups falling under the macrohaplogroup L. In East Africa, the African haplogroup L3 gave rise to two additional macrohaplogroups, M and N, during or after the migration out of Africa (Watson et al., 1997; Torroni et al., 2006).

The dispersion of M and N mtDNA clusters outside Africa can be explained by two routes. The Northern route traversed the Sinai Peninsula and the Levant, while the Southern route followed the Asian coastline and eventually reached Australia (Nei and Roychoudhury, 1993; Foley and Lahr, 1997). The Bab Al Mandab strait served as an alternative to the northern route during glacial stages due to aridity in the Levant acting as a barrier. It is believed that the southern route was favored by modern humans, as the Bab Al Mandab strait was narrow and shallow at that time (Valladas et al., 1988; Mercier et al., 1993; Maca-Meyer et al., 2001).

The descendants of the M macrohaplogroup gave rise to haplogroups C, D, G, and M1–M20 during migration to Southeast Asia and Australia. The N macrohaplogroup followed two directions. One branch moved through Southeast Asia to Australia, while the other

branch migrated Northward into central Asia, generating haplogroups A and Z. Additionally, the N macrohaplogroup gave rise to European haplogroups I, X, and W. The submacrohaplogroup R in Western Eurasia originated from N and further diversified into remaining European haplogroups (H, J, U, K, T, U, and V) (van Oven and Kayser, 2009). In Eastern regions, sub-macrohaplogroup R led to Asian mtDNA haplogroups B and F (Wallace and Chalkia, 2013) (see Figure 1.6).

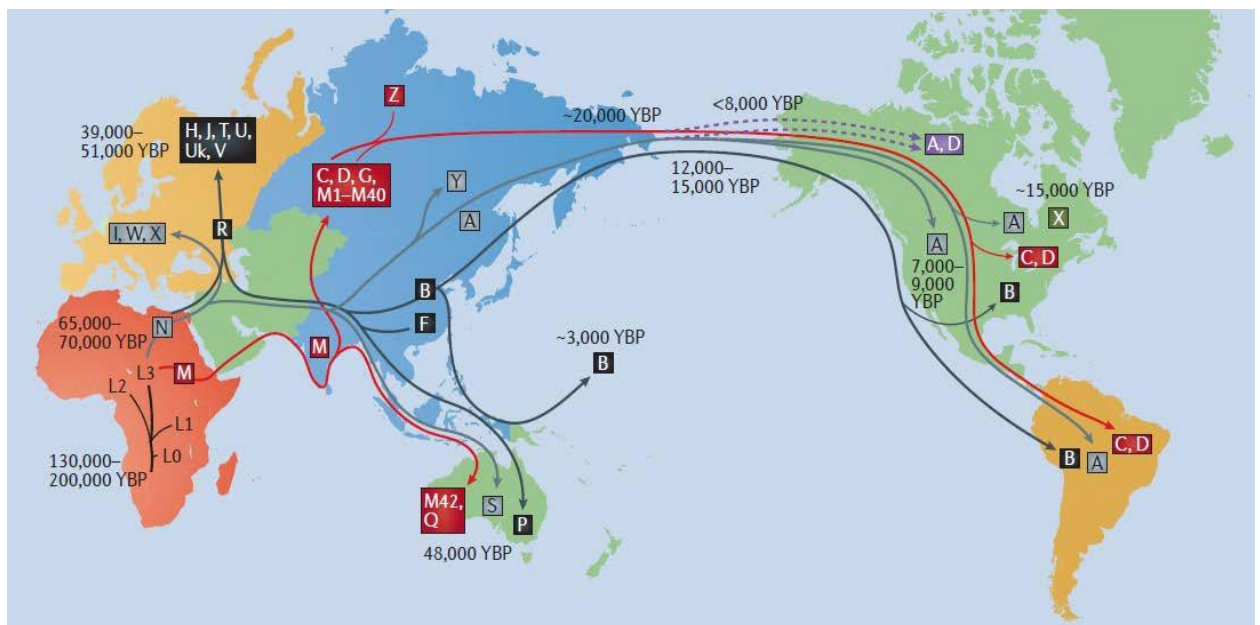


Figure 1.6. Worldwide spread of human population migrations pattern and major mtDNA haplogroups (Taken from Nature Reviews Genetics 2015, 16 (9): 530-542).

Haplogroup classification can be accomplished through analysis of control region (CR) polymorphisms or restriction fragment length polymorphism (RFLP) analysis of mtDNA. Haplogroups were initially named alphabetically in order of discovery (Figure 1.7) and later grouped to some degree by continent (Cann et al., 1987; Vigilant et al., 1991; Chen et al., 2000; Denaro et al., 1981; Jobling et al., 2014). Recent discoveries, such as the Jebel Irhoud fossils in Morocco, challenge the notion that the origins of modern humans are

solely restricted to Eastern, Central, and Southern Africa, indicating the early presence of the *Homo sapiens* clade in Morocco over 300 thousand years ago (Hublin et al., 2017).

The study of mtDNA haplogroups provides valuable insights into the evolutionary history and migration patterns of human populations (Cann et al., 1987; Torroni et al., 1996; Watson et al., 1997; Torroni et al., 2006). The role of phylogenetic trees in representing haplogroups is crucial in understanding human evolution and migration patterns. These trees help trace the historical migrations of populations and the evolutionary divergence of haplogroups over time.

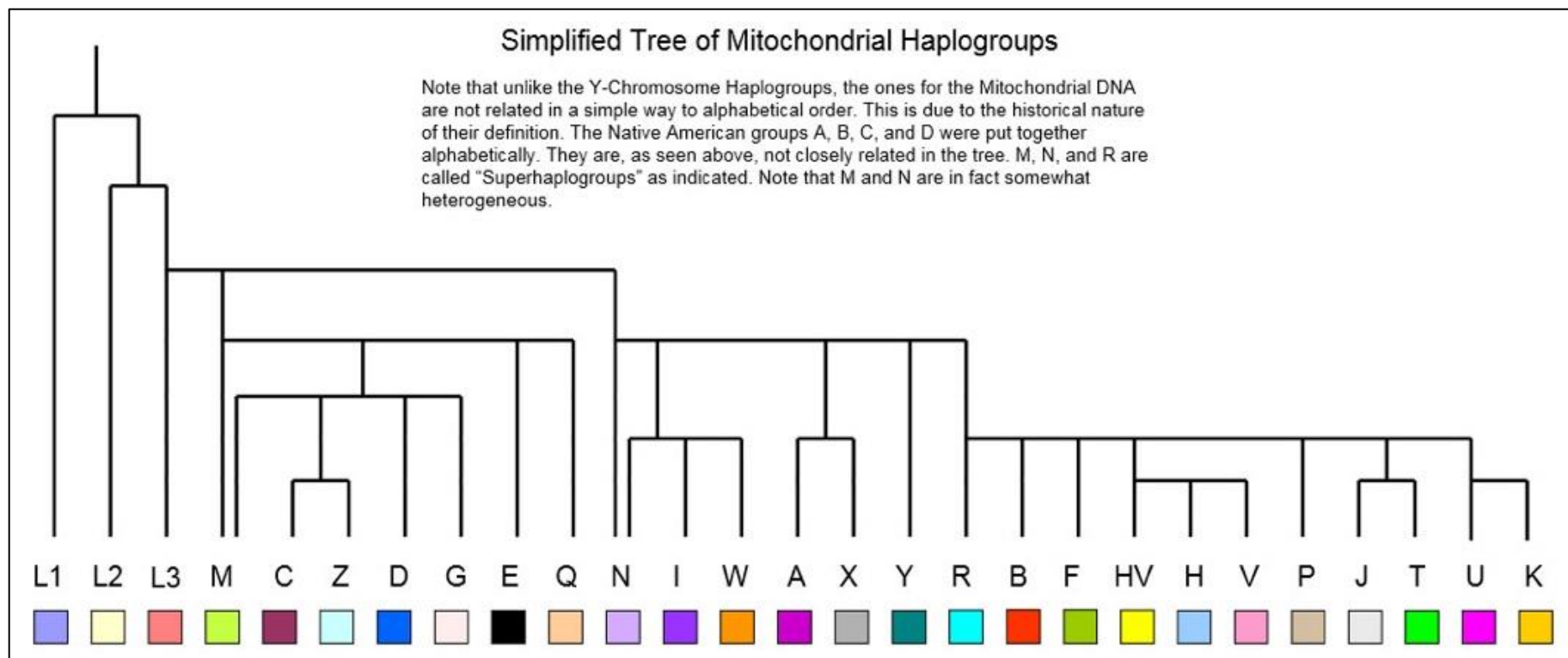


Figure 1.7. Human mitochondrial DNA phylogenetic tree. (Wikitree n.d.)

Over time, these haplogroups diversified into more specific sub-haplogroups due to genetic mutations and geographical separation. For instance, the T haplogroup, which is thought to have originated approximately 45,000 years ago in the Near East, has branched out into several sub-haplogroups as populations migrated and adapted to new environments. A notable example is the T2 sub-haplogroup, which further diversified into T2b, and even more specifically into T2b7a. Each successive sub-haplogroup represents a lineage that carries unique genetic markers, illustrating the dynamic and complex nature of human ancestry. As these sub-haplogroups become more specific, their geographical distribution narrows significantly. For example, while the T haplogroup is widespread, T2 is less common and found more in Europe and the Near East. Moving further down the lineage, T2b is even more specific to certain parts of Europe. Finally, T2b7a1 is highly localized, being found in specific regions of the world (Figure 1.8).

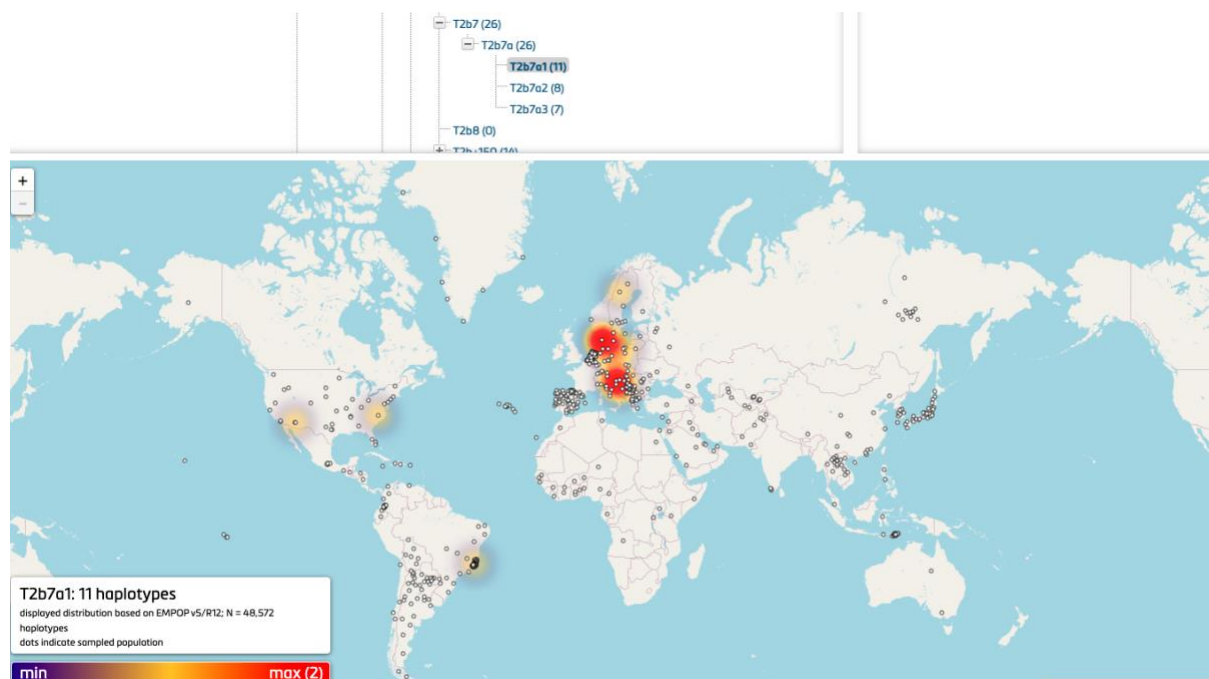


Figure 1.8. Haplogroup T2b7a1 distribution on a world map obtained from EMPOP.

1.8.3. Forensic Genetics

The use of mitochondrial DNA analysis is particularly appropriate when there are degraded skeletal remains or when hairs without roots are encountered in forensic casework or human identification cases. Two primary advantages offered by mitochondrial DNA over nuclear DNA analysis are, the higher sensitivity due to the presence of thousands of mtDNA copies in a cell compared to two copies of nuclear DNA, and the maternal inheritance of the mtDNA, which enables distant maternal relatives to be compared to the analysed samples for relationship hypothesis testing or when the original depositor of the sample is not available (Holland et al., 2013). However, mtDNA, for individual identification, has a limited discriminatory power compared to that of STR analysis (Holland et al. 2013). Mitochondrial DNA can also be used exonerate or tie a person to a crime scene just as autosomal DNA analysis and it can be used either as a sole circumstantial evidence or as a complement to nuclear DNA evidence in court (Leland, 1998).

The quantity and quality of nuclear and mitochondrial DNA in human hair shafts, revealing that nuclear DNA, despite being highly fragmented, is more abundant than previously thought (Brandhagen et al., 2018). The findings in their work highlighted the challenges of obtaining complete nuclear DNA profiles from shed hairs due to DNA degradation and fragmentation, particularly in aged samples, which affects the success rate of traditional STR profiling.

The mtDNA copy number variation relative to nDNA quantity was assessed across nine different tissues from 32 deceased individuals to understand the absolute and relative amounts of mtDNA and their impact on downstream genotyping (Naue et al., 2024). The study found that nDNA quantity is a weak surrogate for estimating mtDNA quantities among different tissues, with significant tissue-specific and inter-individual variations observed. Hair showed extreme variations in mtDNA content, highlighting the need for parallel determination of mtDNA quantity and quality before genotyping. The findings suggest that mtDNA is generally more stable than nDNA, but degradation is tissue-dependent, and accurate mtDNA quantification is critical for reliable genotyping results. Mitochondrial DNA analysis on 691 casework hairs showed high success rates in obtaining profiles, influenced by hair age, colour, and diameter, with 8.7% of samples showing mixtures likely due to surface contamination, and 11.4% exhibiting sequence heteroplasmy (Melton et al., 2005).

1.9. Whole mtDNA genome

The analysis of whole genome mtDNA involves a combination of traditional and modern techniques that have significantly advanced our understanding of this vital genetic component. Historically, Sanger sequencing, a widely used method, played a fundamental role in deciphering the mitochondrial genome. By utilizing polymerase chain reaction (PCR) amplification and capillary electrophoresis, Sanger sequencing allowed for the determination of nucleotide sequences across specific mtDNA regions. This method served as the gold standard for identifying mitochondrial haplotypes and investigating

maternal lineage relationships. However, the advent of Massive Parallel Sequencing (MPS), also known as Next-Generation Sequencing, revolutionized mtDNA analysis. MPS techniques provide high-throughput sequencing with improved coverage and sensitivity, enabling the analysis of the entire mitochondrial genome in a single experiment. This expanded capability has not only increased the accuracy of mtDNA analysis but has also facilitated the discovery of novel genetic variants and the identification of rare mitochondrial haplotypes. The combination of Sanger sequencing and MPS has greatly enhanced our understanding of mtDNA diversity, population structure, and evolutionary history, opening up new avenues for research in fields such as forensic genetics, population genetics, and human migration studies. For mtDNA analysis, various kits for extraction and amplification are available commercially, depending on the research needs. As for amplification, these are some of the available technologies: Qiagen REPLI-g Mitochondrial DNA Kit, Illumina NextSeq or MiSeq Kits, NEB LongAmp Taq 2X Master Mix, Sigma-Aldrich REPLI-g Mitochondrial DNA Kit, and Thermo Fisher Scientific two Precision ID kits for mtDNA; control region and whole genome.

1.10. Mitochondrial DNA Analysis in UAE Populations

Numerous publications have extensively explored the control region of mtDNA within diverse populations worldwide. These studies are pivotal for unraveling population histories and facilitating description of mtDNA haplotype distribution. However, research on mtDNA control region analysis in Arab populations, particularly those in the Arabian Peninsula, remains relatively scarce. Despite significant urbanization, the region has

managed to preserve its traditions and cultural heritage. As a result, the Arab population exhibits a high degree of consanguinity and limited migration to other geographical areas with the preference of settlement in the Arabian peninsula along the Arabian Gulf (Dreaming in Arabic 2016; Tadmouri et al., 2009). Consequently, the Arabian peninsula population potentially possesses a distinct genetic structure, necessitating the establishment of a dedicated database for accurately estimating the rarity of mtDNA, particularly in the context of forensic investigations (Butler, 2011).

In 2007, Alshamali et al. conducted a study reporting the first mtDNA control region data for 249 native Arabs residing in Dubai. The study revealed a Random Match Probability (RMP) of 0.007. However, this study had certain limitations. Primarily, it did not encompass the entire Arab population in the UAE, as the samples were solely collected from native individuals in Dubai and was based only on the control region. The findings of this study is summaized in Figure 1.9.

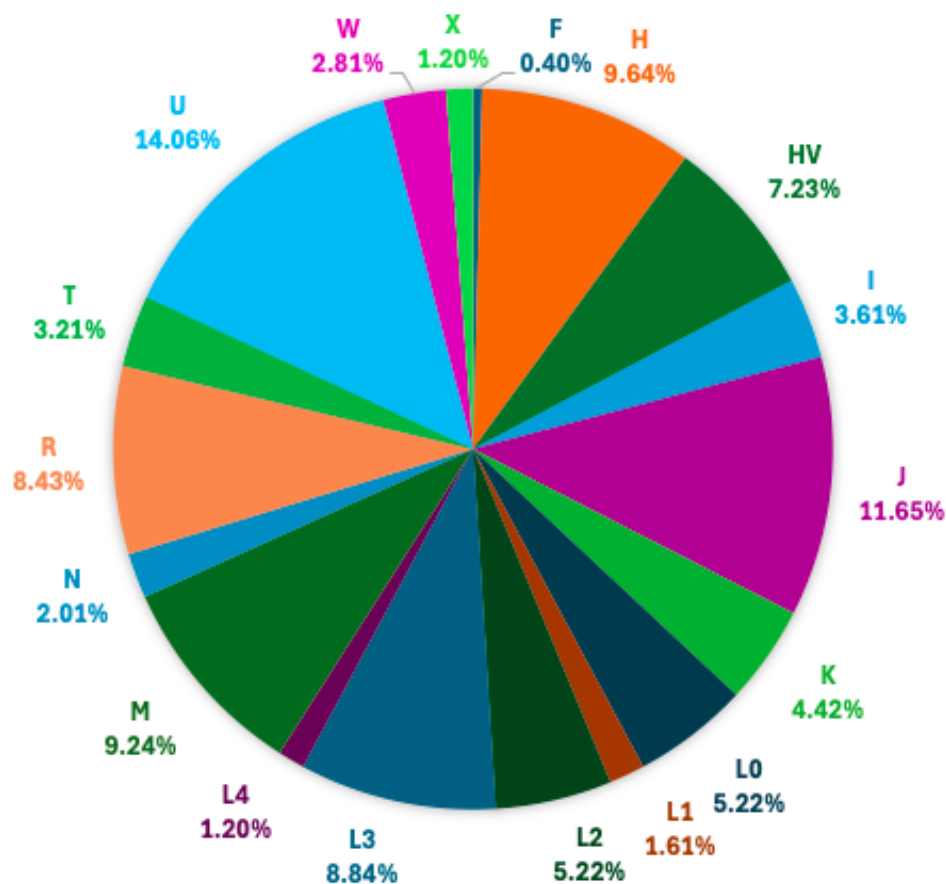


Figure 1.9. Representation of haplogroups in the Dubai dataset derived from control region sequencing (Alshamali et al., 2007).

In 2019, a study aimed to generate and analyse the first whole genome sequences of two UAE nationals to contribute to the understanding of genetic diversity and disease associations within the region (AlSafar et al., 2019). Whole Genome Sequencing was performed on two Emirati individuals, UAE S001 and UAE S002, mtDNA haplogroup reported R2 + 13500 for UAE S001 and G2a1 for UAE S002. The study provides a foundational reference for genetic research and precision medicine in the UAE. It highlighted the genetic diversity and potential health implications for the UAE population, paving the way for targeted healthcare strategies and personalized medicine.

Another study was published in 2020, that aimed to explore the mitochondrial DNA (mtDNA) landscape of the UAE native population, assessing genetic diversity, haplogroups, heteroplasmy, and demographic history, using whole mitochondrial genome sequencing, variant annotation, haplogroup analysis, and demographic history reconstruction (Aljasmi et al., 2020). Long-range PCR was conducted using two sets of overlapping primers, each targeting approximately 8500 base pairs, adopted from Fendt et al. (2009). Two PCR amplicons were used for the NGS library preparation. Subsequently, a 200-base pair sequencing kit (Ion Xpress Plus Fragment Library Kit; Life Technologies, Carlsbad, CA, USA) was employed to create a short mitochondrial genome fragment library. Ion Torrent™ Personal Genome Machine™ (PGM) system platform was utilized for whole mtDNA genome sequencing. 232 female UAE nationals were sequenced. 15 haplogroups were identified, including L, U6, M1, R, J, and K with a total of 968 mtDNA variants were found. Haplogroup U was predominant (16.81%), followed by R (15.08%), and M (12.06%). Sub-Saharan haplogroups L accounted for 8.62%, indicating African migration influences. Figure 1.10 summarizes the distribution of mitochondrial haplogroup frequencies within this study.

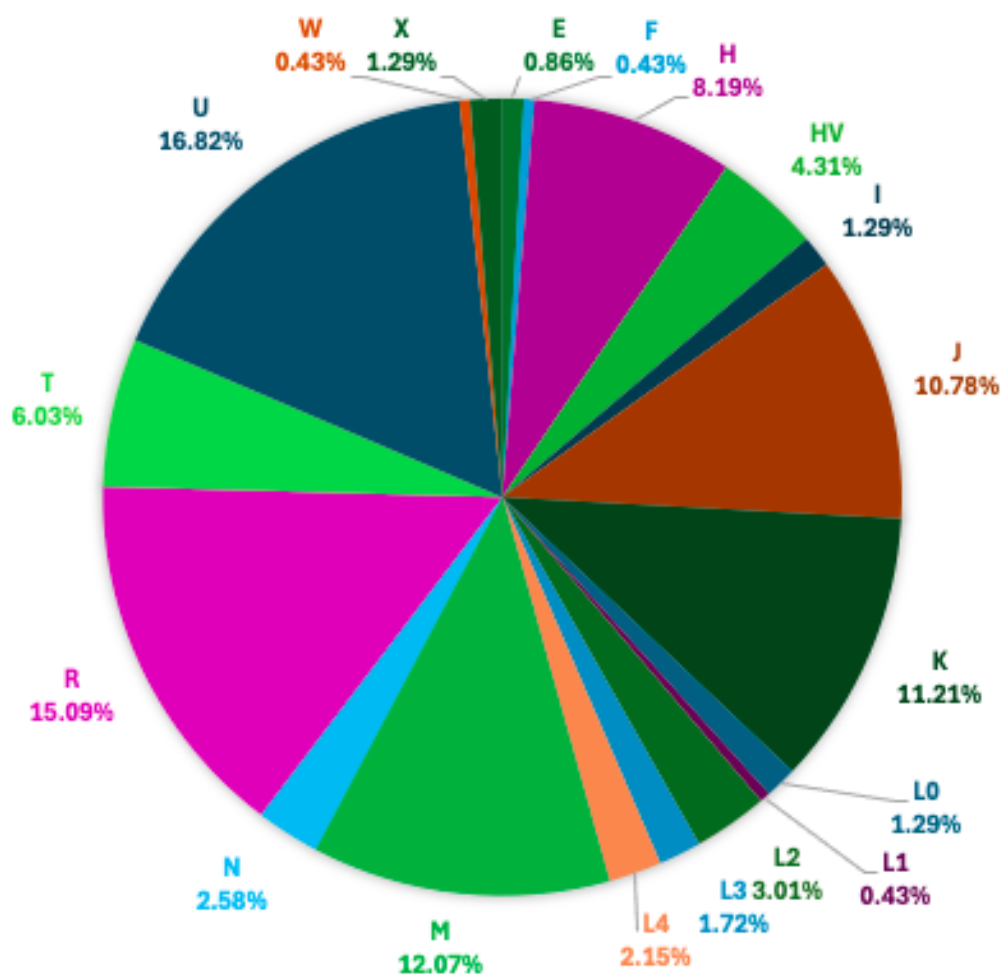


Figure 1.10. Representation of haplogroups frequencies in a whole mtDNA pie chart from the study by Aljasmi et al., 2020 (N=232).

1.11. Mitochondrial DNA Analysis in Middle East Populations

In 2009, a study of mtDNA CR sequences Egyptian population with 277 samples, resulted in 238 different haplotypes (Saunier et al., 2009). Egypt's geographical location connects three continents: Africa, Asia, and Europe. Throughout its history, Egypt has been ruled by various cultures, including Greeks, Romans, Arabs, Turks, French, and British. These historical interactions have contributed to the diverse genetic makeup of the population. Most of the haplogroups were R (22%), followed by L3 haplogroups L3 (12.3%). A representation of all the haplogroups summarized in Figure 1.11.

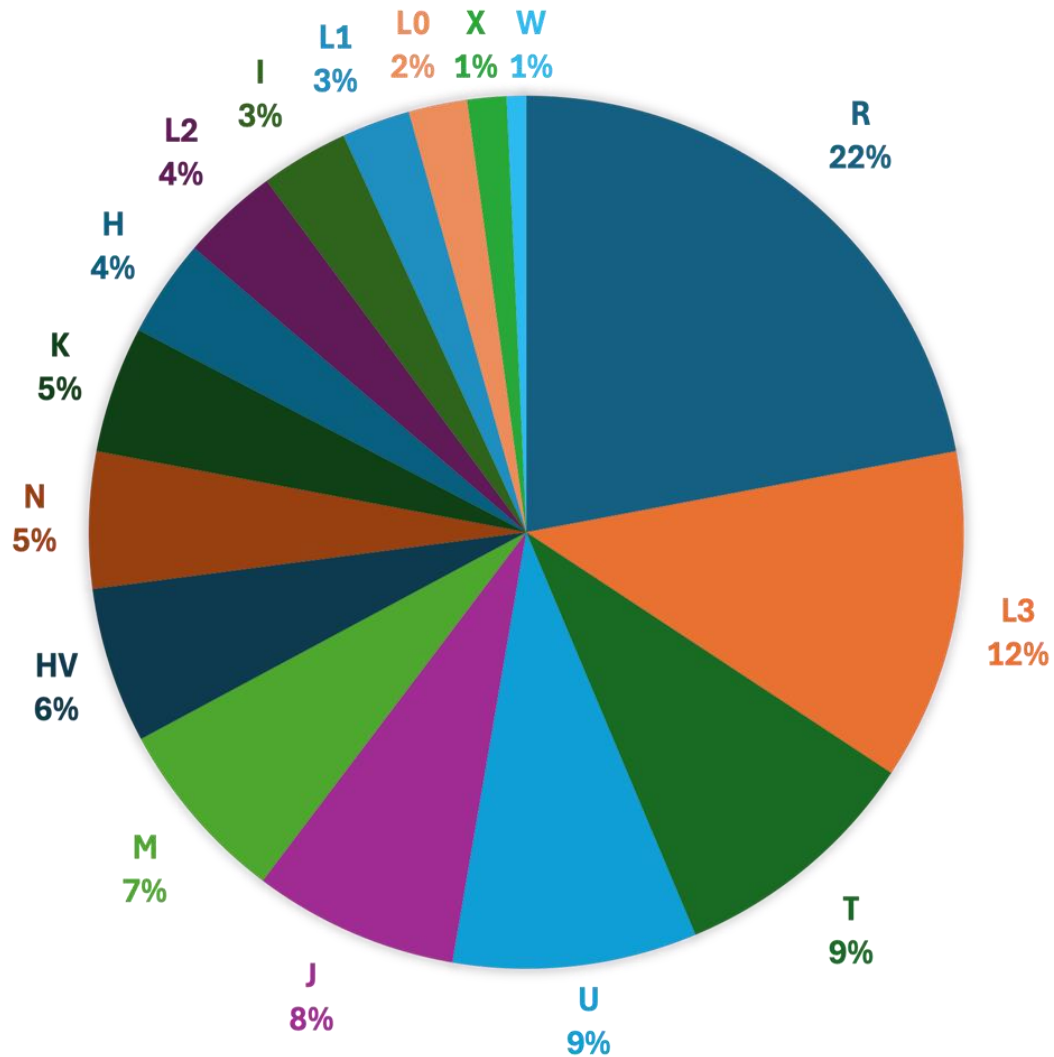


Figure 1.11. Representation of haplogroups in the Egyptian population mtDNA control region (Saunier et al., 2009).

In a study conducted in 2011, Scheible et al. examined the Kuwaiti population using mtDNA CR sequences, analyzing a total of 381 samples. The study reported a high level of genetic diversity with a calculated value of 0.9979. They identified 297 distinct haplotypes and observed three major geographical regions: 235 samples had Western Eurasian (74%), 38 African (10%), and 108 Asian (16%) (Scheible et al., 2011). Figure 1.12 provides a comprehensive summary of the frequencies of various mitochondrial haplogroups identified in the study.

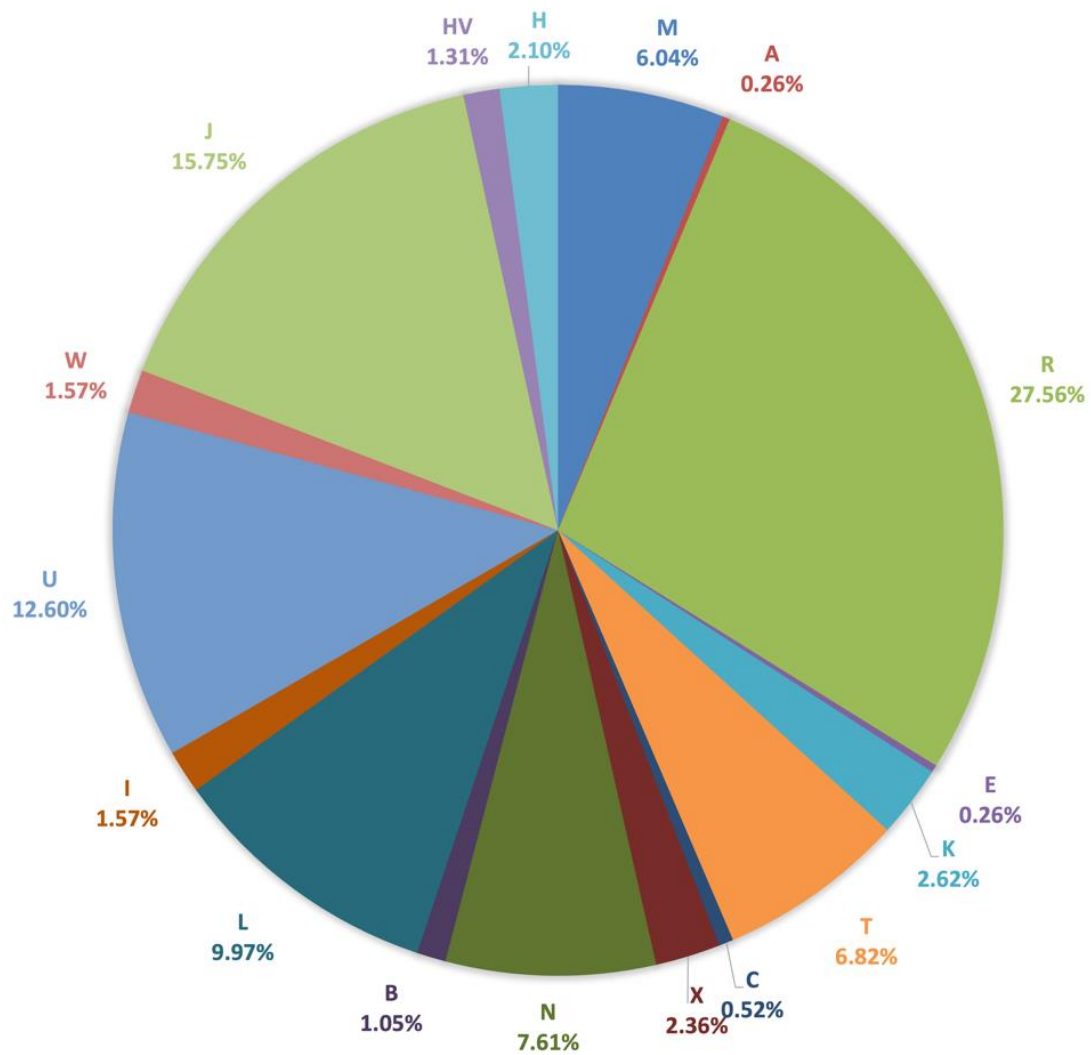


Figure 1.12. Pie chart representation of Kuwaiti population haplogroups of mtDNA CR (Scheible et al., 2011)

A study in which 85 samples were sequenced for HV1 and HV2 from mtDNA genome sequences. Specific haplogroup categories N(xR) paralog, which includes all lineages within haplogroup N except for the R subclade, prevalent in eastern Africa, Arabia, and the Near East) were analyzed (Fernandes et al., 2012). The mtDNA was amplified using 32 overlapping fragments (Maca-Meyer et al., 2001), and sequencing was performed on a 3100 DNA Analyzer (AB, Applied Biosystems, Foster City, CA). Haplogroups N1 (including I), N2 (including W), and X are described as rare and having patchy distributions

across Eurasia. Haplogroup N1b is rare, but one of its subclades, N1b2, is relatively common in Jewish populations, has a Near Eastern origin and is associated with a founder effect in Ashkenazi Jewish ancestors. Haplogroup N1 likely originated in Southwest Asia. N1c is primarily found in Southwest Asia, especially in Arabia. N1a, is most frequent in Arabia, with some deeply rooted lineages in Ethiopia, Somalia, and Yemen. N1e, which split from haplogroup I is found in the Arabian Peninsula and Russia. Haplogroup I is the most frequent clade within N1 is predominantly found in Europe, but it exhibits a frequency peak in the Gulf region, with high diversity values in the Gulf, Anatolia, and southeast Europe. A sub-haplogroup of I5a shows a recent founder on the island of Socotra in the Gulf of Aden. Haplogroup N1b is primarily found in Southwest Asia. Two Somali samples share similarities with N1b, are labelled as part of N1f. N1f likely resulted from ancient gene flow between Arabia and the Horn of Africa. In haplogroup N2, there is a rare subclade known as N2a44, which is mainly found in eastern Europe and the Caucasus, has a minor presence in regions such as Iran, Arabia, and Ethiopia. The major subclade of N2 is haplogroup W, which is more frequent and widespread than N2a44, with some populations in Eastern Europe showing over 10% frequency. Haplogroup W is less common in the Near East and Arabia. The Arabian Peninsula is suggested as the most likely place for the earliest branching of haplogroup N, which includes N1, N2, and X, along with the major Eurasian haplogroup R. Southern Asia saw the emergence of a new N(xR) branch, N5, which is rare and has been identified in Iran. These haplogroups are described as branching directly from the first non-African founder node, the root of haplogroup N.

The research identifies these haplogroups as ancient relics of modern human dispersal out of Africa, approximately 60 thousand years ago. These haplogroups are considered to have a relict distribution that suggests an ancient ancestry within the Arabian Peninsula. They likely spread from the Gulf Oasis region toward the Near East and Europe during a pluvial period between 55–24 thousand years ago. The findings support the hypothesis that Arabia was indeed the first staging post in the spread of modern humans around the world, contributing significantly to the genetic diversity seen across various populations today.

The publication of Zimmermann et al. 2019 studied mtDNA CR variation in three Middle Eastern countries, Jordan, Lebanon, and Bahrain. It consisted of 195 Lebanese, 202 Jordanians, and 213 Bahrainis. The Bahraini population samples had a significant portion (25%) of haplogroup L lineages, L2, L3, L1, and a single occurrence of L5b were observed multiple times in the dataset. Exclusive occurrences of haplogroup L0 lineages, which were found six times. Haplogroup U was the second most common cluster, with a frequency of 16%, and a relatively high proportion of haplogroup M (11%). The frequency of haplogroup R0 in Bahrain was consistent with that in the neighboring country, the United Arab Emirates Alshamali et al., where both were observed at a rate of 8%.

Haplogroup H was dominant haplogroup in Lebanese set (16.41%), while haplogroup U was dominant in Jordanian set with a frequency of 16.8%. It was noteworthy that haplogroup U3 was observed from 8 samples in both Lebanese and Jordanian samples sets, consistent with previous study from the Dead Sea population set (Rowold and

Herrera, 2010). Furthermore, four lineages of Haplogroup U6 were observed in the Jordanian set, but completely absent in the Lebanese set. Haplogroup G was observed once in the Lebanese sets; G2a+152; and was not observed in the other sets. Haplogroup X was present in both Lebanese and Jordanian sets while absent in the Bahraini set. Figures 1.13, 1.14, and 1.15 summarizes the findings of the study.

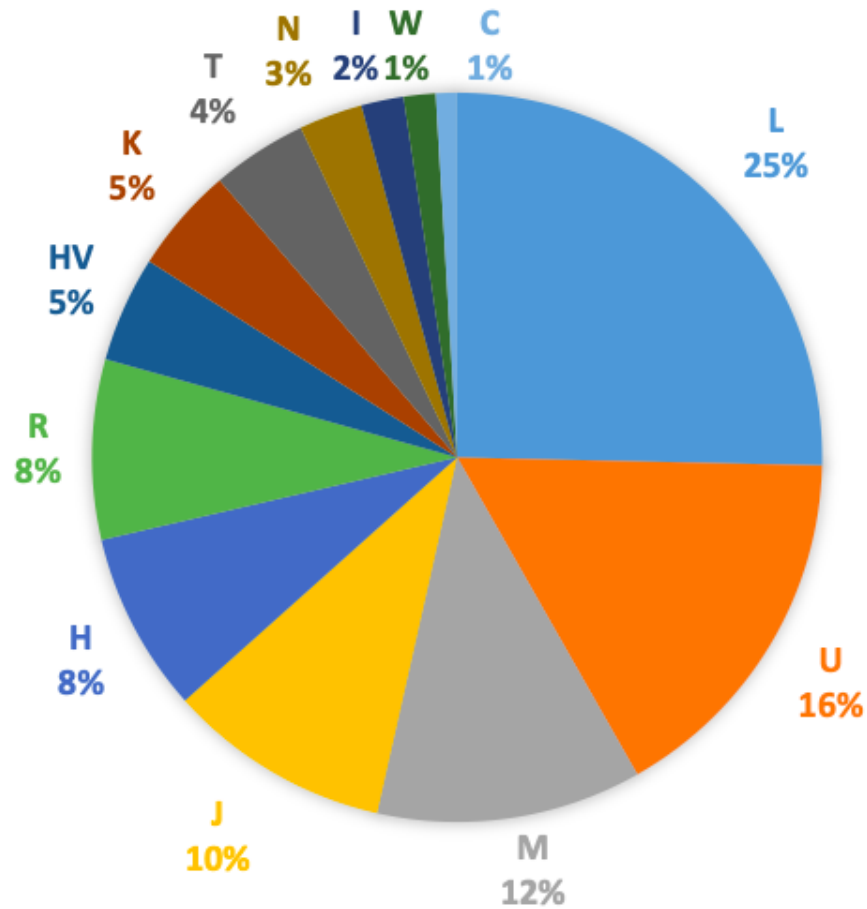


Figure 1.13. Bahraini population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)

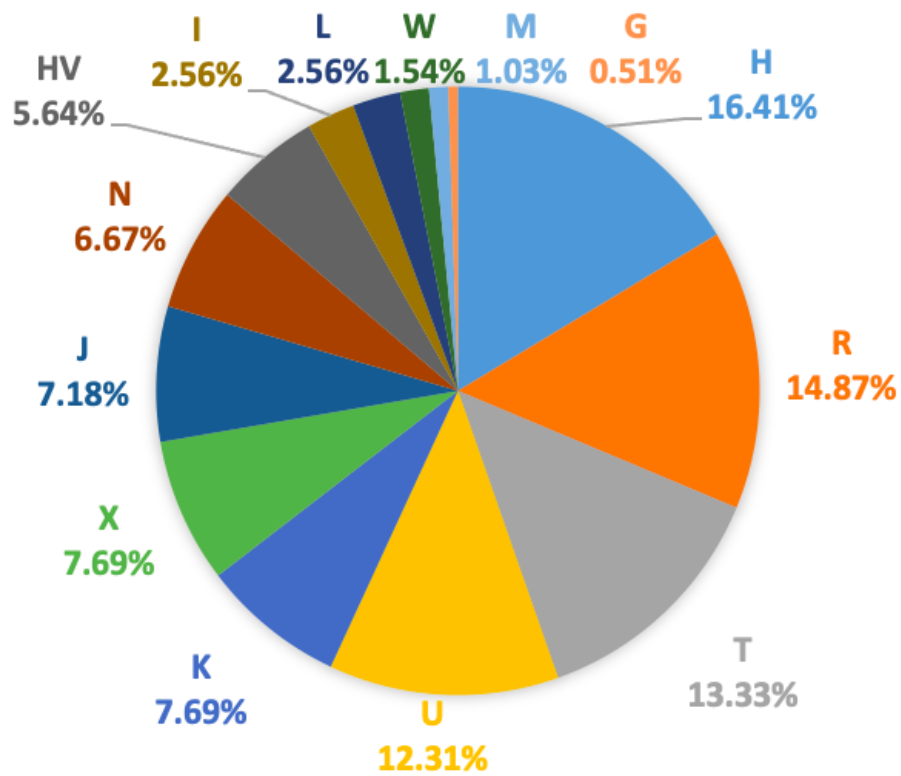


Figure 1.14. Lebanese population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)

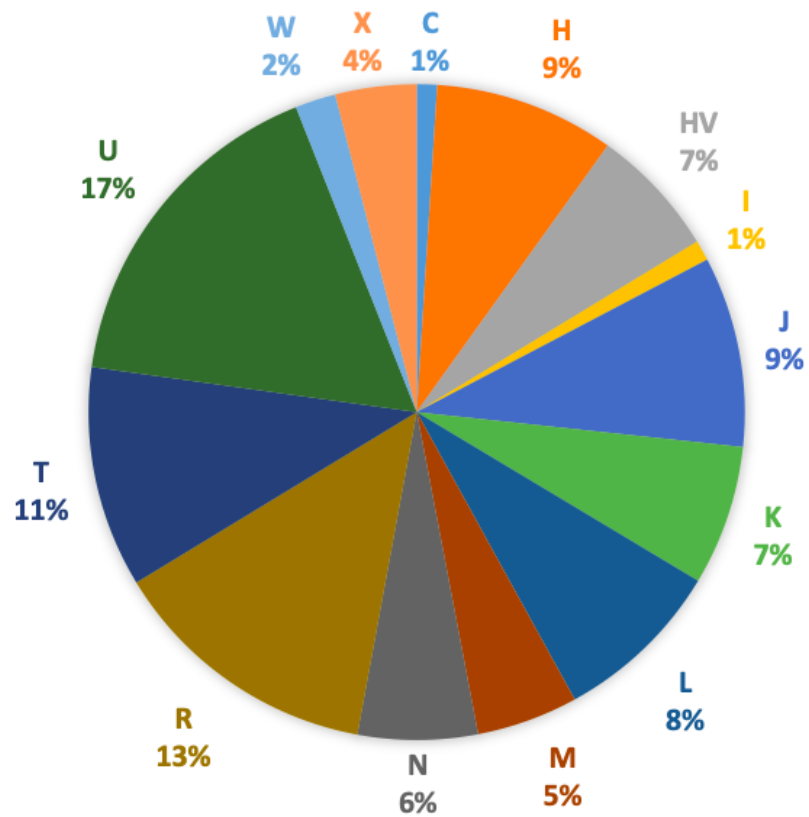


Figure 1.15. Jordanian population set mtDNA CR haplogroup distribution (Zimmermann et al., 2019)

1.12. Mitochondrial DNA Analysis Sub-Asian Population

Large-scale population studies have been conducted on various Sub-Asian ethnic groups to provide a comprehensive understanding of mtDNA diversity in this region. Previous research has highlighted the distinct haplogroup distributions and migration patterns within these populations. For example, a study by Kivisild et al. (1999) analysed mtDNA variations among different Indian subpopulations, revealing significant genetic diversity and ancient maternal lineages. Similarly, Thangaraj et al. (2006) focused on the Andaman Islanders, identifying unique haplogroups that provide insights into early human migrations out of Africa. These studies have helped to construct a framework for current findings, align present readings, and build confidence in this study by providing comparative data and validating our methodologies.

1.12.1. Indians

In 2004, a study investigated significant haplogroups pan India (Metspalu et al., 2004). It was found that Haplogroup M is prevalent in India, with variations across regions. Haplogroup W is the most common mitochondrial DNA haplogroup in India but follows a distinct spatial pattern. Haplogroup U is the most frequent sub-clade of haplogroup R in India, with U2 sub-groups evenly distributed. Haplogroup U7 follows a pattern similar to Haplogroup W and is also found in neighboring regions. The study also extended its findings to include Iran, with West Eurasian haplogroups predominating. Haplogroup U surpasses Haplogroup H in Iran, leading to the presence of sub-groups like U7 rarely found in Europe. Haplogroup M in Iran appears to have a different regional origin. The

study reveals important insights into the genetic diversity and historical migrations in India. It finds that certain haplogroups, such as M2, M2a, M2b, and R5, are prevalent in India and have deep coalescence ages, indicating their ancient presence in the region. The frequency of haplogroup M2 is high among Austro-Asiatic tribal populations, but a re-evaluation by Kumar et al. (2008) suggests it may have been overestimated due to incomplete coding region data in earlier studies. West Eurasian-specific mtDNA haplogroups are prevalent in western states of India and Pakistan but decline towards the South and East, more common in caste populations than tribal groups. East Eurasian-specific mtDNA haplogroups are less common in India but exhibit geographic variation, with elevated frequencies in regions near Tibet, Myanmar, and Southeast Asia. Surprisingly, some southern Indian populations, particularly in Tamil Nadu, show relatively high frequencies of East Eurasian-specific mtDNA haplogroups. Central Asian-specific haplogroups like D4c and G2a are rare in India but present in Iran and Southern Indian states.

Another study was carried out that focused on mtDNA variations in 24 different tribal populations across India (Kumar et al., 2008). The study investigates the distribution of macrohaplogroup M within 24 Indian tribal populations by analyzing 2,768 mitochondrial DNA (mtDNA) samples. It reveals that macrohaplogroup M is present in 69.39% of the samples. Interestingly, tribes from the Western region of India exhibit significantly lower frequencies of macrohaplogroup M, typically around 50% or less. The study also identifies the presence of M2 within the samples, with a notable absence of M2 among the eight

tribes in Northeast India, except for a single instance in the Sonowal Kachari tribe. Excluding the Northeast tribes, the frequency of the M2 haplogroup among the studied tribes is approximately 13.86%, with variations across tribes. The Betta Kuruba tribe in the Southern region has the highest frequency of M2 (39.13%), while the adjacent Jenu Kuruba tribe has a much lower frequency (7.02%). The distribution of subclade M2b within the tribes was also explored. M2b is absent among Indo-European speakers in Western and Central India but is prevalent among Dravidian-speaking tribes and some other populations. Its frequency varies significantly, with the Betta Kuruba tribe having the highest frequency (35.65%). This distribution pattern of M2b suggests population-specific differences. A phylogenetic tree was constructed based on the sequencing of 72 mtDNAs of the M2 haplogroup and 4 additional M2 complete sequences from existing literature. This tree reveals two main sister clades, M2a and M2b. M2a further divides into three basal branches: M2a1, M2a2, and M2a3.

Another study focused on the distribution and genetic diversity of mtDNA haplogroup R7 within different linguistic groups in India (Chaubey et al., 2008). The findings of the study suggest that the presence of haplogroup R7 is ununiformly distributed in India. The study incorporated 35 new genetic sequences, revealing the existence of eight new subclades within six branches of haplogroup R in the Indian subcontinent, suggesting their ancient origins. Haplogroup R7, specifically subclade R7a1, is more prevalent among Munda speakers with a geographic concentration in the Austroasiatic "heartland" in states like Bihar, Jharkhand, and Chhattisgarh. Haplogroup R6, also more frequent among

Austroasiatic speakers, is part of the main cluster that includes populations from all linguistic groups. Further analysis of R7 mtDNAs reveals two deep-rooted subclades, R7a and R7b. R7a dates back approximately 3,000 to 7,000 years ago, depending on the mutation rate used, and is a converging point for all Austroasiatic individuals. In contrast, the coalescent times for R7 variation among Dravidian and Indo-European speakers are older.

Another recent study focused on two populations in India; including Gujjars from the Jammu region of Jammu and Kashmir (J&K) and Ladakhis from the Ladakh region (Singh et al., 2020). Gujjars are a nomadic/semi-nomadic group inhabiting North Western regions of the Indian subcontinent, while Ladakhis have diverse genetic makeup due to historical trade routes in the Ladakh region. The study found that Gujjars exhibited the lowest mitochondrial DNA (mtDNA) nucleotide diversity among all reference Indian populations, indicating relatively lower genetic diversity within this group. In contrast, Ladakhis displayed higher genetic diversity. Haplogroup analysis based on mtDNA revealed that Gujjar samples could be assigned to 19 unique haplogroups, with M30f being the most abundant, followed by R5a, M30, and U2a. Ladakhis, on the other hand, exhibited 42 unique haplogroups, with M9 being the most abundant, followed by haplogroups A, C4, and D4.

1.12.2. Pakistanis

In 2014, a study focused on the Makrani people of Pakistan was carried out. The study analyzed 100 samples and reported a genetic diversity value of 0.9688, a random match

probability of 0.0408, and a power of discrimination of 0.9592. The researchers observed multiple haplogroups, including African haplogroups (28%), West Eurasian haplogroups (26%), South Asian haplogroups (24%), and one East Asian haplogroup, while the remaining 20% mtDNA of the sampled individuals could not be confidently assigned a continental origin (Figure 1.16). Furthermore, they identified a total of 70 haplotypes, with 54 unique haplotypes and 16 haplotypes shared among multiple individuals (Siddiqi et al., 2015).

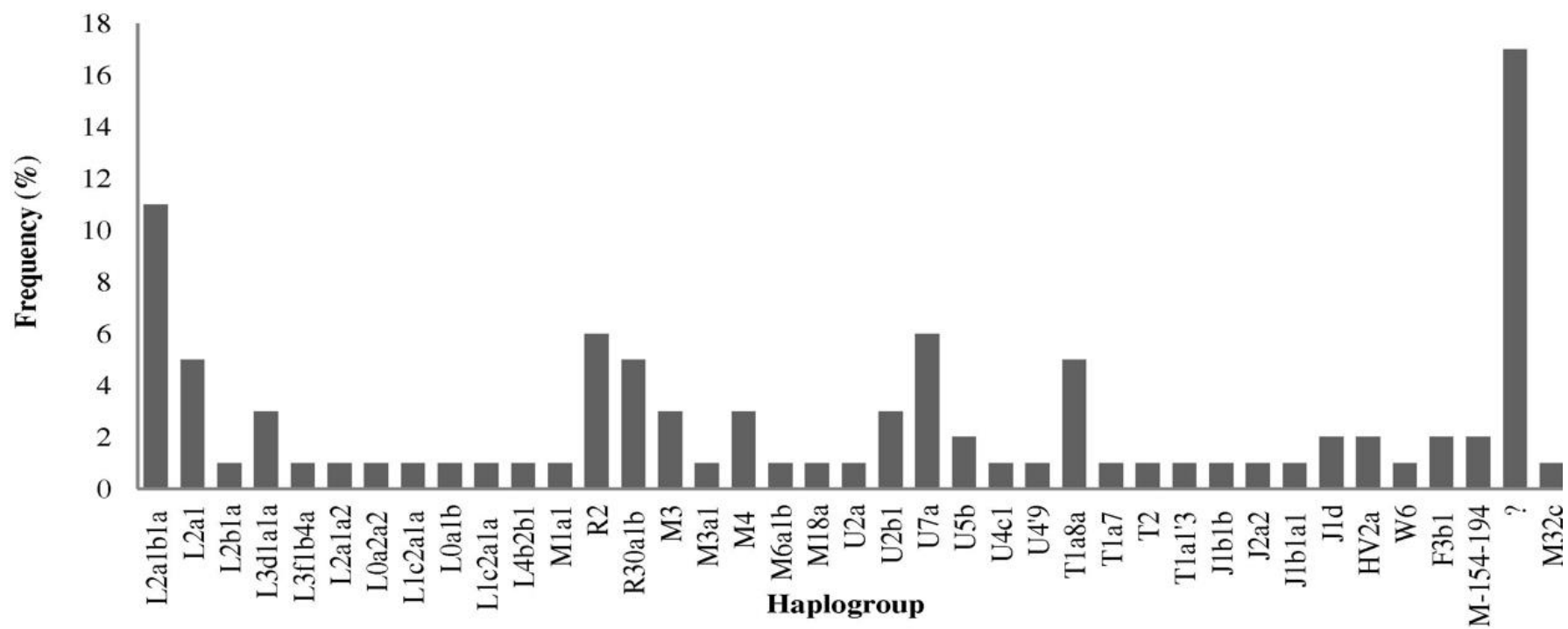


Figure 1.16. Graphical representation of frequencies of the mtDNA haplogroup composition of the 100 sampled Makrani from Pakistan (Taken from Siddiqi et al., 2015)

In a 2017 study by Yasmin et al., the genetic composition of the Sindhi ethnic group in Pakistan was examined. The study analyzed 88 samples from the Sindhi population and identified several haplogroups, with the most common ones being M (50%), U (18.18%), H (13.63%), R (10.22%), W (3.41%), J (2.27%), L (1.13%) and T (1.13%) (Figure 1.17). These haplogroups represented South Asian, West Eurasian, and African genetic influences. A comparison was made between the Sindhi population and other ethnic groups in Pakistan, as well as various population datasets from different regions. The analysis showed a high level of genetic variation within the Sindhi population, while the variation between populations was relatively low.

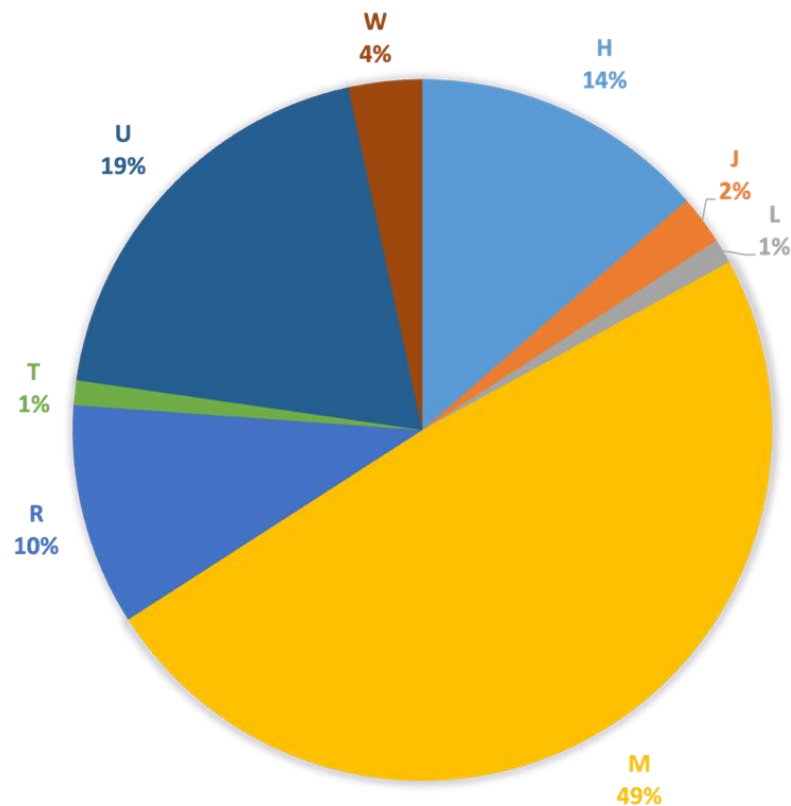


Figure 1.17. Demonstration of the relative proportions of the observed haplogroups in Sindhi population (Yasmin et al., 2017)

Working hypothesis:

An evaluation of mtDNA of the 610 samples of the UAE population who represent the whole geographical region, which includes Indians and Pakistanis residing in the UAE.

1.13. Aim

This PhD thesis was designed to re-evaluate mtDNA of the UAE by representing the whole geographical region. This study constructed mtDNA database of the UAE population by investigating haplotypes and their corresponding haplogroups. The project also evaluated the potential of mtDNA in forensic application where genomic DNA seems to be weak or degraded, by applying the MPS technology using whole mtDNA genome of casework samples.

Previous studies on the mtDNA of the United Arab Emirates (UAE) population have not fully captured the genetic diversity across all seven Emirates: Abu Dhabi, Dubai, Sharjah, Ajman, Umm Al Quwain, Ras al Khaimah, and Fujairah. This study was able to obtain samples from individuals from each of these Emirates, as well as critical casework samples such as bones and touch samples with weak STR profiles. The study also examined the Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific) through a concordance study to evaluate its accuracy and reliability. The research addressed the lack of comprehensive whole mtDNA population genetics data for Indians and Pakistanis residing in various parts of the world, particularly those who have migrated to the Middle East and the UAE. Although some existing studies cover these populations, they often overlooked the effects of migration movements and genetic flow. The aim of the research

was to sequence the whole mitochondrial genomes of individuals from these communities living in the UAE, thereby creating high-quality, comprehensive mitochondrial genome haplotypes that accurately represent the UAE population. The objective was to establish a robust database that would provide forensic analysts with precise tools for estimating haplotype frequencies. This would enable analysts to determine the rarity of a match during forensic comparisons. The uniqueness and originality of the study lie in its potential to significantly enhance the understanding of mtDNA analysis and forensic genetics in the UAE. By conducting this research, the study would create a foundational resource for forensic applications, ultimately improving the accuracy and effectiveness of forensic investigations in the region.

1.14. Objectives

- To obtain Ethical Approval and Sample Collection: Upon obtaining necessary ethical approval, the collection of biological samples that represent the Emirati population set.
- To undertake a Diverse Population Sampling: In addition to the native population, sample collection from two major populations residing in the UAE, Indians and Pakistanis.
- To optimization all Laboratory Procedures: Refine DNA extraction protocol and optimization of PCR to ensure the highest quality and efficiency in Sanger sequencing for mtDNA control regions.

- To undertake mtDNA Sanger sequencing by performing Sanger sequencing on 93 samples for mtDNA CR from the three population groups to establish a genetic baseline for MPS.
- To undertake a Massive Parallel Sequencing (MPS) Implementation: Implement MPS using the Precision ID Whole mtDNA Genome panel to sequence the complete mitochondrial genomes of samples from the three populations, ensuring a comprehensive analysis of mitochondrial DNA.
- To carry out a Concordance Evaluation: Conduct a concordance study to compare the mitochondrial DNA profiles obtained from Sanger sequencing and MPS. This comparison will assess the consistency and reliability of genetic data across different sequencing technologies.
- To evaluate the MPS data to identify and characterize the mitochondrial haplogroups present within the populations. This analysis will include assessing the quality of sequencing data, the depth of coverage, and the detection of any novel variants.
- To undertake Casework Data Evaluation: Analyse casework data to explore the practical applications of mtDNA sequencing in forensic workflow, which assesses the utility of mtDNA in real-world scenarios and its implications for population genetics studies.

Chapter 2

2. Materials and Methods

2.1. Materials

- **General Materials:**

Auto-Mate Express™ Instrument

Bleach

Benchguard® sheets

Centrifuge machine

Cutting mat

Dithiothreitol (DTT)

Ethanol

Eppendorf tubes

FTA® cards (Whatman® Bioscience, UK).

FTA® purification reagent (Whatman® Bioscience, UK)

Micro-Amp® Optical 96-well reaction plate

Nuclease-free tubes

Optical Adhesive Covers (Applied Biosystems, USA)

Personal Protective Equipment

PCR tubes

Pipettes

PrepFiler Express BTA™ kit (Thermofisher)

PrepFiler Express™ kit (Thermofisher)

QIAGEN disposable Harris Uni-Core™ punch with 1.2 mm I.D. tip

Racks

Thermomixer incubator

Tips

TE buffer

UV-box benchtop decontamination chambers

Virkon®

Vortexer

Veriti® 96-Well Thermal Cycler (Thermo Fisher Scientific, USA)

- **Samples:**

510 blood samples from the UAE

50 blood samples from Indians

50 blood samples from Pakistanis.

56 bone samples

56 casework samples (blood, saliva, touch DNA, body fluids)

GEDNAP blood samples

High primates blood samples (Gibbon, male chimp, female chimp)

Origene 9947A control DNA sample

Promega 9947A control DNA sample

Standard reference materials (SRM-2392)

- **Sanger sequencing:**

AutoLys® 96-deep well plate

AutoLys Tubes

STARLet® instrument

EZ1 DNA Investigator Kit

Forward and reverse primers by Eurofins MWG Operon (L15879, H727, F15975, R635, R240)

Platinum® master mixes

Quantifler™ Human DNA Quantification kit

Quantifiler® Trio kit

Reddymix™ master mixes

- **Gel electrophoresis and sequencing**

ABI 3500 Genetic Analyser (Applied Biosystems, USA).

BigDye XTerminator® Purification Kit

Casting trays

Connecting cables

Electrodes

GeneScan POP-7 (Applied Biosystems, USA),
Hi-Di™ formamide (Applied Biosystems, USA)
HyperLadder™ 1 kb DNA size marker (Bioline ®)
MicroCLEAN™ reagent
Power supply
SafeView™ (NBS Biologicals Ltd)
Submarine chamber tank
Well comb
2.5% Agarose gel
1X Genetic Analyser Buffer with EDTA

- **Massive parallel sequencing:**

CleanNGS Reagent (Clean NA, ZH, Netherlands)
Ion Library TaqMan™ Quantification Kit
Ion Chef™ instrument
Ion S5 XL instrument
Ion S5™ Precision ID Sequencing Reagents cartridge
Ion S5™ Precision ID Chef Solutions
Ion S5™ Precision ID Chef Supplies
Ion 530™ Chip.
Ion 520™ Chip.
Ion S5™ Precision ID Wash Solution bottle
Ion S5™ Precision ID Cleaning Solution bottle
Precision ID mtDNA Whole Genome Panel
Qubit® dsDNA HS Assay Kit
Qubit® 3.0 fluorometer
QuantStudio™ 5 Real-Time PCR system

- **Softwares:**

BioEdit software
Converge™ software (Thermo Fisher Scientific, USA)
EMPOP mtDNA database

Haplogrep 3.2.1 database
Hamilton® AutoLys Robotic system.
Image Lab software
mtDNAprofiler software
SeqScape™ software 4.0
Sequencing Analysis Software v5.4 (Applied Biosystems, USA).
Torrent Suite™ Software
4Peaks software

2.2. Samples from United Arab Emirates Population

The target study population came mainly from local UAE Emiratis, and to lesser extent Indians and Pakistanis. Sample collection comprise a total of 610 samples of female and male individuals from the three ethnicities, encompassing 510 samples from the UAE, 50 samples from Indians, and 50 samples from Pakistanis. These samples were obtained with informed consent from blood donors in Dubai Health Authority and were provided to the Department of Forensic Sciences and Criminology in Dubai (Ethical Approval Grant Ref. DSREC-SR-03/2018_03). All paperwork on informed consent was approved by the University of Central Lancashire STEMH Committee and the Government of Dubai. Samples were stored and transferred on FTA® cards (Whatman® Bioscience, UK). For Emirati participants, maternity was confirmed to be Emirati through the verification of official documents, such as birth certificates and family registration books (Khulasat Al-Qaid), Emirati women typically belong to the Arab ethnic group, specifically the Arab ethnic subgroup known as the "Emirati Arabs." An Emirati woman is a citizen of the United Arab Emirates, that legally holds the Emirati nationality and recognized as a national of

the UAE. Her nationality is passed from an Emirati father (*jus sanguinis*), and not acquired legally from an Emirati husband through citizenship.

2.3. Amplification of mtDNA Control Region

The amplification of the mtDNA control region is normally a critical step in obtaining high-quality genetic data for further analysis. The methods used to amplify this region were examined for efficiency and effectiveness. To minimize experimental steps and ensure suitability for Sanger sequencing, a thorough investigation of existing literature was conducted. The goal was to identify an approach that uses the least number of primers without compromising the quality of the results. The method described by Cardena et al. (2013) was found to be optimal, allowing for the amplification of the control region using a single pair of primers. The primers selected based on their methods are detailed and were ordered from Eurofins MWG Operon® based on the published sequences.

2.3.1. Primers Selection

In order to minimise the steps involved in the experiment the literature was investigated to identify an approach to amplify mtDNA Control Region (CR) with the least number of primers taking into consideration its suitability for Sanger sequencing without compromising the quality of the results. The methods described by Cardena et al., 2013 allowed the amplification of the control region using one pair of primers, as shown in Figure 2.1. The primers detailed in their methods are used in sequencing. Table 2.1 shows

all primers which were ordered from Eurofins MWG Operon® based on the published sequences.

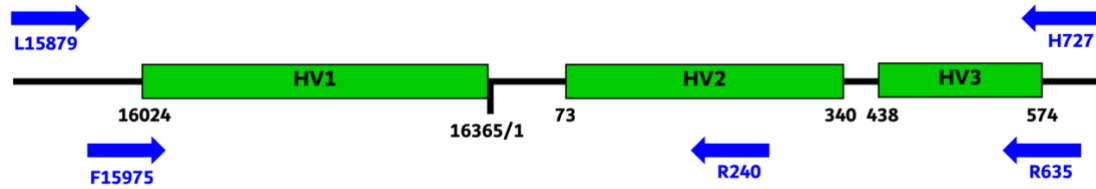


Figure 2.1. A Schematic diagram showing the primer chosen to amplify mtDNA control region.

Table 2.1. Forward and reverse primers obtained from the online protocol (Yonsei University, n.d.) for the amplification of the mtDNA CR.

Names	Primer 5' - 3'
L15879	AAT GGG CCT GTC CTT GTA GT
H727	AGG GTG AAC TCA CTG GAA CG
F15975	CTC CAC CAT TAG CAC CCA AA
R635	GAT GTG AGC CCG TCT AAA CA
R240	TAT TAT TAT GTC CTA CAA GCA

2.3.2. Primers Quality Control

All primers used in this project were ordered at HPLC grade by Eurofins MWG Operon (Schoske et al. 2003).

2.4. Quality Control

Quality control practices in sample extraction were pivotal for accuracy and reliability in the laboratory. Ensuring the safety of personnel and sample integrity involved systematic use of Personal Protective Equipment (PPE), including lab coats, gloves changed between steps, face masks, closed-toe shoes, and additional hair coverings. Hand hygiene and

adherence to safety regulations completed the comprehensive approach to a secure laboratory environment during extraction. Standard Operating Procedures (SOPs) from kits and instruments provided a consistent approach. Calibration, regular equipment maintenance, validation of extraction methods, and the use of blanks and controls (positive and negative) were aspects of quality assurance. Proper record-keeping, from samples collection to extraction, including deviations documentation, was crucial for traceability. Environmental controls, emphasizing cleanliness and temperature regulation, minimized contamination risks. Constantly decontaminating the pipettes, tips, racks, and Eppendorf tubes using the UV-box benchtop decontamination chambers. Also, working surfaces were disinfected using bleach and Virkon®. Benchguard® sheets were used and discarded after every lab session. These practices collectively enhanced the overall quality and reliability of laboratory outcomes.

2.5. DNA Extraction and Purification

Automated extraction and purification for conventional extraction techniques were facilitated employing the EZ1 DNA Investigator Kit (Qiagen) in conjunction with the EZ1 Advanced XL system. Additionally, the PrepFiler Express™ and PrepFiler Express BTA™ kits (Thermofisher) were utilized in tandem with the Automate Express Nucleic Acid Extraction System as well as Hamilton Robotics, encompassing the STARlet™ and Microlab AutoLys™ STAR platforms. Direct PCR and DNA extraction methods were both put into practice for product optimization and validation. Detailed information on the

extraction methodologies and the corresponding sample quantities processed are outlined in Table 2.2.

Table 2.2. Different extractions kits used in the studied samples.

Extraction methods	Number of Samples
PrepFiler Express™	619
PrepFiler Express BTA™ (Bone Samples)	56
Hamilton Robotics	56

2.5.1. Population Samples Extraction

Six discs (2.0 mm) punched out of each FTA® card stained with blood and placed into barcoded spin-basket/filter-equipped AutoLys Tubes and introduced into the Hamilton® AutoLys Robotic system. Each sample was then treated by the addition of 500 µL of PrepFiler® lysis Buffer and 10 µL of freshly prepared dithiothreitol (DTT). Samples were incubated at 70 °C for 40 min followed by centrifuging for 2 min at 13000 xg. The filtrate was then transferred to 96-deep wells plate and made ready for purification on STARLet®. Extracted DNA samples were dispensed in 96-deep well plate by AutoLys® and made ready for purification. The plates, and extracts, were transferred to the STARLet® instrument for purification using Magnetic Beads, Wash Buffer and Elution Buffer. All samples were eluted in 40 µL of TE buffer (10 mM Tris-HCL, 0.1 mM EDTA, pH 8.0).

Simultaneously, an equivalent quantity of the punched-out discs was transferred to appropriately labelled PrepFiler LySep™ Columns. To this, 500 µL PrepFiler™ Lysis Buffer and 5 µL freshly prepared DTT was added, and the tubes were then closed firmly and

transferred to the Thermomixer incubator. Samples were subjected to an incubation of 40 min at 70 °C and 750 rpm. Following incubation, a centrifugation step for 2 min at 10,000 xg was carried out, facilitating the transfer of the lysate to the sample tubes. Activation of the AutoMate Express™ Instrument followed onscreen directives step by step. Finally, elution tubes were obtained for each samples containing 50 µL extracted DNA, which was stored at -20°C.

2.5.2. Casework Samples

The samples chosen; with a total of 56; for inclusion in this study, were directly obtained from Forensic Biology section – Dubai Police. Samples were received and were already extracted using the PrepFiler Express™ Kit and Hamilton® AutoLys Robotic system. Table 6.1 in Chapter 6.1 summarized the casework samples.

2.5.3. Bone Samples

Bone samples were obtained in powdered form from Forensic Biology section – Dubai Police. PrepFiler Express BTA™ Forensic DNA Extraction Kit was used. Following the user guide, the protocol of bone was used. Table 6.2 from Chapter 6.1 summarized 56 solid tissue samples (bones) obtained from the forensic biology. Approximately, 50 mg of bone powder were weighed and subsequently placed into a PrepFiler™ bone Lysate Tube. For each sample, a lysis mixture was added, with 220 µL of PrepFiler BTA™ Lysis Buffer, 5 µL of freshly prepared 1 M DTT, 7 µL of Proteinase K, and 230 µL BTA solution. Following that, the tubes were vortexed and centrifuged briefly. The tubes were then positioned

within a rotating incubator at 56 °C, and a rotational speed of 1100 rpm was maintained throughout the overnight incubation. Upon completion of the incubation, the tubes were subjected to a centrifugation step at 10,000 xg for 90 s. The resulting clear supernatant was then transferred to a newly labelled PrepFiler™ Sample tube. With the AutoMate Express™ Instrument in operation the PF Express BTA option was chosen to initiate the sample processing with adherence to onscreen guidelines. Elution tubes containing 50 µL of the extracted DNA were carefully preserved at -20 °C.

2.5.4. Direct PCR Extraction

For each PCR reaction, a disc sized 1.2 mm of FTA® card stained with blood was punched and transferred to a clean PCR tube. A volume of 200 µL of FTA® purification reagent (Whatman® Bioscience, UK) was added to the disk in the PCR tube and incubated at room temperature for 15 min, then the reagent was discarded, and the disc was washed twice with 200 µL of TE buffer. Finally, the disc was dried at room temperature before adding the PCR amplification mix.

2.5.5. DNA quantitation

The nuclear DNA concentrations were determined using Quantifler™ Human DNA Quantification kit (Applied Biosystems, USA) run on an AB 7500 real-time PCR system (Applied Biosystems, USA). Quantification standards were prepared by serial dilution in TE buffer of the 200 ng/µL Human DNA Standard included in the kit. Eight serial dilutions were prepared (50, 16.7, 5.56, 1.85, 0.620, 0.210, 0.068 and 0.023) ng/µL. The procedure

was carried out according to the manufacturer's protocol, but instead of using full-scale reactions, the total volume of this assay was halved (Westring et al. 2007). The half-scale reaction consisted of a 5.25 μ L of Quantifiler™ PCR Reaction Mix, 6.25 μ L of Quantifiler™ Human Primer Mix and 1 μ L of the DNA sample for a final reaction volume of 12.5 μ L. Samples, including the Non-Template Control (NTC) and the DNA standards, were loaded into MicroAmp® Optical 96-well reaction plate and sealed using Optical Adhesive Covers (Applied Biosystems, USA). Each standard dilution was run in duplicate, with replicates placed in wells near each other.

The sample sheets were set up and the samples were run using the standard thermal amplification profile of 95 °C for 10 min; 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Standard curves were generated, and amplification plots were used to compare the results. Internal Positive Control (IPC) results were monitored for the presence of PCR inhibitors.

From the results obtained, the samples with low DNA concentrations (< 0.01 ng) were re-quantified to confirm no error had happened during the sample preparation. The samples which still showed very low DNA amounts were re-extracted from the original FTA card using 12 punches of FTA® cards stained with blood sized 1.2 mm and then quantified with Quantifiler™ kit.

Also, Quantifiler® Trio was performed for the forensic samples which included body fluids, touch DNA, bones, and teeth. The kit included Quantifiler™ THP DNA Standard (100 ng/ μ L), Quantifiler™ THP DNA Dilution Buffer. Starting with standards preparation, five

microcentrifuge tubes labelled S1, S2, S3, S4, and S5. 10 μL of Quantifiler™ THP DNA Dilution Buffer was pipetted into the tube labelled S1, and 90 μL of the same buffer was pipetted into each of the tubes labelled S2, S3, S4, and S5.

To prepare Standard 1 (S1), 10 μL of Quantifiler™ THP DNA Standard (100 ng/ μL) was added to the tube labelled S1. The tube was then vortexed thoroughly to mix the contents, resulted in 50 ng/ μL DNA standard solution. To prepare Standard 2 (S2), 10 μL of the S1 solution was transferred to the tube labelled S2. The tube was then vortexed thoroughly to mix the contents, resulted in 5 ng/ μL DNA standard solution. For Standard 3 (S3), 5 μL of the S2 solution was transferred to the tube labelled S3. The tube was vortexed thoroughly to mix the contents, resulted in 0.5 ng/ μL DNA standard solution. To prepare Standard 4 (S4), 10 μL of the S3 solution was transferred to the tube labelled S4. The tube was then vortexed thoroughly to mix the contents, resulting in 0.050 ng/ μL DNA standard solution. For Standard 5, 10 μL of the S4 solution was transferred to the tube labelled S5. The tube was vortexed thoroughly to mix the contents, resulting in 0.005 ng/ μL DNA standard solution. The standards concentrations were assigned manually in the real-time PCR setup.

The reaction mix was prepared by multiplying the total number of samples in the run by 10 μL of Quantifiler Trio Reaction Mix and 8 μL of Quantifiler Trio Primer Mix, and then combining these in a clean, nuclease-free tube. The mixture was thoroughly mixed by gentle vortexing and briefly centrifuged to collect the droplets at the bottom of the tube.

The reaction mix was dispensed into each well of a 96-well reaction plate. Each well received 18 μ L of the reaction mix.

Following with the addition of 2 μ L of each DNA sample, DNA standard, and NTC to the appropriate wells containing the reaction mix. The plate was then sealed with an optical adhesive cover, vortexed and briefly centrifuged to ensure all droplets were at the bottom of the wells.

The reaction plate was placed into the real-time PCR instrument. The PCR cycling program was set up according to the manufacturer's instructions, Quantifiler® Trio typically involving an initial hold at 95°C followed by 40 cycles of denaturation at 95°C and annealing/extension at 60°C.

The data were analyzed using the instrument's software, which generated a standard curve from the known DNA standards and calculated the concentration of DNA in the unknown samples based on this curve. The calculated DNA concentrations were reviewed. The results were recorded and used to determine the suitability of the DNA samples for downstream applications.

2.6.The PCR Amplification conditions

All PCR experiments were done using four samples consist of:

- Positive control, pGEM® -3Zf (+) (Applied Biosystem, USA)
- Negative controls, PCR Graded water and extraction blank (Applied Biosystem, USA)
- R1 and R2: samples of two unrelated females' individuals (Internal known samples).

2.6.1. Primers Selections

Different set of primers were utilized for amplification PCR and sequencing the mtDNA CR in this study, the selected primers were summarized in the Table 2.3 and explained further in Chapter 3.

Table 2.3. Primers for PCR and Sequencing

Primer	Sequence (5'→3')
L15879	AAT GGG CCT GTC CTT GTA GT
H727	AGG GTG AAC TCA CTG GAA CG
F15975	CTC CAC CAT TAG CAC CCA AA
R635	GAT GTG AGC CCG TCT AAA CA
R240	TAT TAT TAT GTC CTA CAA GCA

2.6.2. Control Region Amplification

The Veriti® 96-Well Thermal Cycler (Thermo Fisher Scientific, USA) was used for the PCR. The primers L15879 and H727 mentioned in Table 2.3 were used to amplify the mtDNA control region following the PCR conditions in Table 2.4. PCR reaction comprised of 0.5 µL of each primer (forward and reverse), 5 µL of either Platinum® or Reddymix™ master mixes and DNA template (3 µL of extracted DNA or 1.2 mm disk of purified FTA® with an addition of 3 µL PCR grade water).

Table 2.4. PCR conditions for amplification of mtDNA control region

Temperature	Time Points	No. of Cycles
95 °C	10 min	1 cycle
95 °C	30 s	
56 °C	45 s	
72 °C	1 min	35 cycles
72 °C	7min	
15 °C	∞	Hold

2.6.3. Gel Electrophoresis

To check the performance of the PCR, a volume of 5 µL of the amplified products was electrophoresed in a 2.5% agarose gel stain with 5 µL of SafeView™ (NBS Biologicals Ltd). HyperLadder™ 1 kb DNA size marker – Bioline® was used and loaded together with the samples onto the gel. The gel was photographed using Image Lab software.

2.7. Sanger Sequencing

To sequence the mtDNA control region, Sanger sequencing was employed. This process involved two critical steps: the purification of PCR products and the subsequent sequencing reaction using the BigDye® Terminator chemistry.

2.7.1. PCR Product Purification

Following the PCR amplification process of mtDNA control Region, PCR products were purified using microCLEAN™ reagent. A volume of 10 µL of microCLEAN™ was added to the PCR product, mixed by pipetting and left at room temperature for 5 min. Afterward, PCR tubes were centrifuged at 13000 xg for 7 min, the supernatant was removed twice

and then the pellets were resuspended in 10 μ L of 10mM Tris-Cl buffer pH 8.0 (TE) and left to rehydrate for 5 min at room temperature.

2.7.2. BigDye® Sequencing Reaction

Sequencing reactions were performed in reaction volume of 10 μ L with 0.75 μ L of BigDye® Terminator Reaction Mix V3.1 (Applied Biosystems), 1.7 μ L BigDye® Sequencing Buffer, 0.32 μ L (3.2 pm) of forward or reverse primer, 2.23 μ L of PCR grade water and 5 μ L of purified PCR product from the previous step. Thermal cycling was performed using a Veriti® (Applied Biosystems, USA) using the following cycles conditions shown in Table 2.5.

Table 2.5. Sequencing PCR cycles conditions.

Temperature		Time Points
96 °C		1 min
96 °C	10 s	
50 °C	5 s	25 Cycles
60 °C	4 min	
12 °C		∞

2.7.3. BigDye® Sequencing Product Purification

Purification of PCR product before sequencing to remove dNTPs, primers and other nonspecific PCR products was essential to prevent the results from displaying noisy data. In this research, two approaches were tested; these were DNA precipitation method and BigDye XTerminator® Purification Kit. In DNA precipitation method, following sequencing reaction, a volume of 1 μ L of -4 °C 3M NaOAc, pH 4.6, 1 μ L of glycogen (20 μ g/ μ L),

1.0 μ L of EDTA (100 mM) and 30 μ L -20 °C ethanol (96%) were added and left to incubate for a minimum of overnight and a maximum for 24 h at room temperature. Then samples were centrifuged for 30 min at 4 °C, and then the supernatants were carefully removed, pellets were washed twice with freshly prepared 70% ethanol and left to dry in PCR machine at 50 °C for 10 min. The BigDye XTerminator® Purification Kit, was much simpler, and it was performed by adding Xterminator reagent, vortex and finally centrifuge for 30 min at maximum speed before loading to CE.

2.7.4. Sequence Detection and Analysis

The detection of mtDNA control region sequencing reaction products were done with the ABI 3500 Genetic Analyser (Applied Biosystems, USA). The samples were prepared by adding 11 μ L Hi-Di™ formamide (Applied Biosystems, USA) to each sample in the PCR tube. The samples were then left for 10 min at room temperature to rehydrate and finally transferred to the corresponding well in the 96-well plate. A total of 8 samples was injected at a time for 10 s at 3,000 v and separated at 15,000 v for 120 min with temperature run of 60 °C. The separation was achieved using GeneScan POP-7 (Applied Biosystems, USA), 1X Genetic Analyser Buffer with EDTA and 50 cm capillary arrays. After the data collection samples were analysed by Sequencing Analysis Software v5.4 (Applied Biosystems, USA).

2.8. Whole mtDNA Sequencing

Massive Parallel Sequencing (MPS) enabled the simultaneous analysis of multiple small DNA fragments. In high-quality samples, MPS was previously shown to efficiently cover the entire genome with just two extensive overlapping amplicons. However, forensic samples often showed signs of degradation and fragmentation, making this approach impractical. To achieve full genomic coverage in such cases, it was essential to use a larger number of smaller amplicons.

This method posed significant technical demands and consumes more time, primarily because of the extensive set of primers needed for the small-amplicon assay. Additional challenges included the inadvertent co-amplification of nuclear mitochondrial DNA segments (NUMTs) (Chaitanya et al., 2015), and the detection of various phylogenetic motifs or mutation patterns that were linked to specific haplogroups and haplotypes (Strobl et al., 2019). The complexity of these challenges underscored the value of commercially available panels.

The application of MPS technologies to whole mtDNA sequencing normally marked a substantial progression in fields like forensic genetics, medical research, and evolutionary biology. MPS offers an exhaustive analysis of the mitochondrial genome, providing detailed insights into genetic variations with higher resolution than traditional methods (King et al., 2014). Unlike Sanger sequencing, which was restricted to smaller genomic regions and often necessitated multiple assays to encompass the entire mitochondrial genome, MPS could sequence thousands to millions of DNA fragments concurrently. This

capability not only ensured complete coverage in a single assay but also enhanced the detection of low-frequency variants and heteroplasmy (Peck et al., 2016). These features were vital for forensic investigations and the study of mitochondrial diseases. By sequencing entire mitochondrial genomes, MPS enhances the precision of haplogroup classification and supported the simultaneous processing of numerous samples, thereby saving both time and costs compared to older sequencing methods.

The evolution of sequencing technologies over the past decade has reinforced the concordance between Sanger and MPS data. Breitingner et al. (2024) highlight this technological progression, which has streamlined forensic workflows and improved analytical consistency.

2.9. Commercially Available Whole mtGenome MPS workflow

Two commercial kits for sequencing the entire mitochondrial genomes were recently introduced to the market. The Precision ID mtDNA Whole Genome Panel from Thermo Fisher Scientific was one such kit. It comprised 162 amplicons divided into two multiplexes, each amplicon averaging about 163 base pairs in size (Parson et al., 2013). Since its introduction, this panel had undergone numerous assessments and a comprehensive developmental validation (Cihlar et al., 2020), specifically for use on the Ion Torrent™ sequencing platform by Thermo Fisher Scientific.

The panel, however, was not originally designed or optimized for use with the MiSeq FGx™ sequencing platform by Verogen, Inc., which was the MPS platform utilized at ESR Ltd

(McElhoe et al., 2014). For this research, a workflow was crafted enabling the sequencing of the Precision ID mtDNA Whole Genome Panel on the MiSeq FGx™. This approach had effectively produced complete mitochondrial genome haplotypes from both high-quality and degraded samples, including buccal swabs and hair shafts, based on previous studies conducted in this laboratory.

The ForenSeq™ mtDNA Whole Genome Solution by Verogen, Inc., San Diego, CA, USA, represents the second commercially available panel designed for the analysis of whole mitochondrial genomes from forensic-type samples, specifically tailored for the MiSeq FGx™ platform. This solution comprised 245 amplicons distributed across two multiplexes, each with an average size of 131 base pairs. Recent developmental validations for the ForenSeq™ mtDNA Whole Genome and Control Region Solutions had been documented in the literature (Holt et al., 2021), though not as extensively as the Precision ID panel. The ForenSeq™ kit was a comprehensive system that included automatic sequencing library analysis through the ForenSeq™ Universal Analysis Software (UAS). However, the UAS did not currently support the masking of personally identifiable information (PHI) sites, which is a critical requirement for forensic mitochondrial genome analysis in New Zealand. This limitation had necessitated the development of a custom analysis pipeline using GM-HTS to mask PHI sites during data alignment, thereby adapting the kit for local forensic requirements. This adaptation ensured compliance with New Zealand's privacy regulations and allowed for accurate forensic analysis without compromising sensitive information.

Furthermore, Promega™, based in Madison, WI, USA, was at the time of starting this project preparing to launch the PowerSeq™ Whole Mito System. This upcoming panel would produce 161 amplicons, averaging 167 base pairs each, organized into two multiplexes.

The primary objective of this doctoral research was to evaluate the Precision ID mtDNA Whole Genome Panel by Thermo Fisher Scientific for its utility in MPS of the mitochondrial genomes on the Ion Torrent™ platform, specifically for potential forensic applications. This assessment was intended to determine the panel's efficacy and suitability within a forensic context, considering the unique challenges associated with the conditions of forensic samples. The evaluation focused on the panel's performance metrics, its ability to generate reliable and accurate mtDNA profiles, and its compatibility with the intricacies and requirements of forensic DNA analysis.

2.10. Ion Torrent workflows

Ion Torrent was developed as DNA sequencing technology platform developed by Ion Torrent Systems Inc., a subsidiary of Thermo Fisher Scientific. It leverages semiconductor sequencing, an innovative technique that directly converted chemical signals into digital data. Ion Torrent technology could be distinguished by its speed, cost-efficiency, and scalability, making it a robust tool in genomics research. In contrast to traditional sequencing methods that relied on optical detection, Ion Torrent used semiconductor chips to detect hydrogen ions released during DNA synthesis. This approach could facilitate direct and real-time monitoring of nucleotide incorporation.

The Ion Chef™ instrument, was developed by Thermo Fisher Scientific, as an automated sample preparation system specifically designed to enhance the workflow of MPS applications. This instrument was intended to be used in conjunction with Ion Torrent™ sequencing platforms, providing a streamlined and simplified approach to sample preparation.

The Ion Chef™ instrument performed several key functions, including template preparation, clonal amplification, and chip loading. It could automate the process of template preparation, which involved the clonal amplification of DNA fragments on ion sphere particles (ISPs) or Ion Chef™ Chip. The Ion Chef™ system used automated liquid handling and robotic movements to carry out these steps efficiently and reproducibly. Central to this technology was the Ion Chip, which contains millions of microscopic wells. Each well could harbor a single DNA template along with a polymerase enzyme. The incorporation of a nucleotide into the DNA strand could release a hydrogen ion, resulting in a pH change that the chip detects.

2.10.1. Qubit® 3 Fluorometer Quantification

Qubit® dsDNA HS Assay Kit was used for DNA quantification. The method was validated and recommended by Thermo Fisher Scientific for mtDNA quantification prior library preparation. (Per kit of 100 assays, Qubit® dsDNA HS Reagent (250 µL), Qubit® dsDNA HS Buffer (50mL), Qubit® dsDNA HS Standard #1, and Qubit® dsDNA HS Standard #2). The concentration of the kit ranged for the detection of samples between 1 ng/mL to 500 ng/mL in a volume of 200 µL after dilution. Following the user's guide recommendation

for Precision ID mtDNA panels employment, 0.1 ng of genomic DNA (gDNA) for each targeted amplification reaction. All samples were then normalized by manually diluting them using Nuclease Free water (Invitrogen™). All samples were prepared using Qubit® Assay tubes that are 500 µL tubes specially designed to fit the fluorometer. Primarily, Qubit® working solution was prepared in 1.5 µL tube by mixing an adequate amount of the Qubit® dsDNA HS Reagent and Buffer. Table 2.6 summarized the user's guide Qubit® working solution preparation. Using the volumes provided, a total of 200 µL from the Qubit® working solution was eluted in each 0.5 µL tube. Table 2.7 summarized the user's guide for sample preparation.

Table 2.6. Qubit® working solution preparation per sample (n)

Reagent	Volume
Qubit® dsDNA HS Reagent	1 µL x <i>n</i>
Qubit® dsDNA HS Buffer	199 µL x <i>n</i>

Table 2.7. Qubit® dsDNA HS Assay samples' preparation

	Volume	Qubit® working solution
Standards	10 µL	190 µL
Samples	3 µL	197 µL

Initially, the Qubit® 3 Fluorometer was calibrated by creating a calibration curve to generate quantification results. This was accomplished by preparing Standard #1 and Standard #2. First, two clean tubes were used, and only the tube lids (this was done to avoid any interference with the readings) were labelled as S1 for Standard #1 and S2 for Standard #2. Then, 190 µL of Qubit® working solution was eluted into both tubes. Next,

10 μL of Qubit® dsDNA HS Standard #1 was added to the tube labelled S1, and 10 μL of Qubit® dsDNA HS Standard #2 was added to the tube labelled S2. Both tubes were vortexed for 2–3 s and incubated for 2 min at room temperature.

While the incubation was in progress, the fluorometer was set up. On the fluorometer's home screen, the DNA option was selected, and dsDNA High Sensitivity was chosen as the assay type, followed by following the on-screen prompts. The "Read standards" screen was displayed. The tube containing Standard #1 was inserted into the sample chamber, the lid was closed, and then the Read standard button was pressed. After the reading was completed, Standard #1 was removed, and the same steps were repeated for the tube containing Standard #2. The results were displayed graphically on the screen, with the plot showing a positive slope, as illustrated in Figure 2.2. It was crucial to insert the standards in the correct order (Standard #1 followed by Standard #2) during the calibration step. It was also noted that after the incubation period, the fluorescence signal remained stable for 3 hours at room temperature.

For sample quantification, the tubes were labelled accordingly, and 197 μL of the working solution was dispensed into each tube and, 3 μL from each sample was added to each tube. Both tubes were vortexed for 2–3 s and incubated for 2 min at room temperature. Meanwhile, on the fluorometer screen, the sample volume and unit for the Qubit® quantification assays were selected. In this case, the volume was set to 3 μL , and the output sample units were set to ng/ μL . The sample tube was then inserted into the sample

chamber, the lid was closed, and the Read tube button was pressed. After reading, the tube was removed, and the process was repeated for all samples.

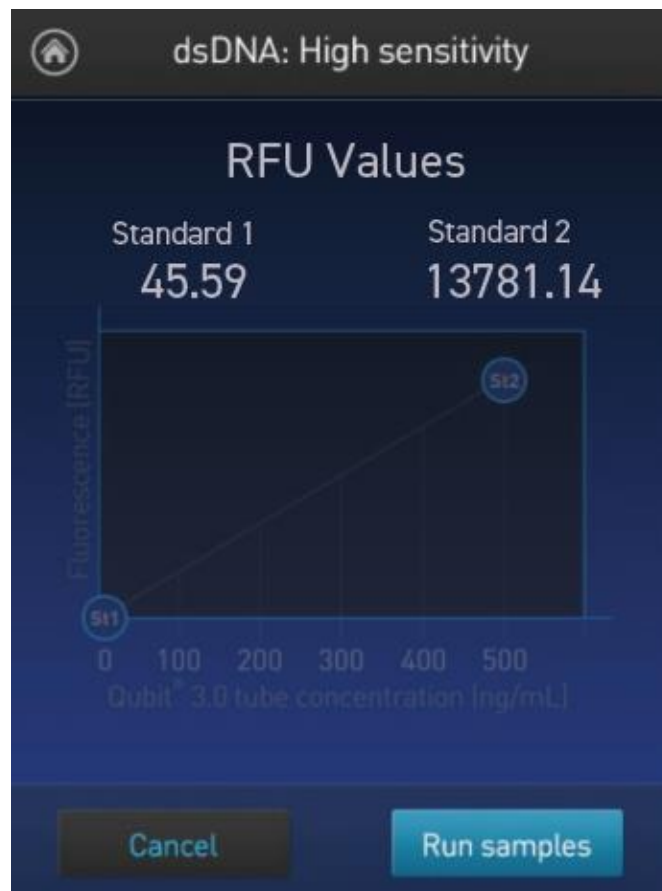


Figure 2.2. Calibration graph of Qubit® quantification assays

During the readings of the results, if a reading fell within the assay's range, it was displayed on the screen. However, if it's outside the assay's range, an "Out of Range" message was displayed instead. The results screen typically presented the results in two key values: The top value, displayed in large font, represents the concentration of the original sample. The bottom value represented the dilution concentration. This format provided users with both the concentration of the undiluted sample and the concentration after any necessary dilution, ensuring accurate and informative quantification results.

2.10.2. Library Preparation

The library preparation process was conducted both manually and automatically using the Ion Chef instrument with the Precision ID Library Kit, specifically the Precision ID mtDNA Whole Genome Panel. This library kit employed a plate-based procedure, which offered advantages in terms of sample tracing and compatibility with both automated and manual methods. The DNA material for this process was obtained from bodily fluids and bone samples. The Precision ID mtDNA Whole Genome Panel included an assortment of unlabelled PCR primers that were meticulously designed using the Ion AmpliSeq™ methodology. This design was aimed to facilitate the creation of libraries from mitochondrial DNA (mtDNA). Notably, the kit featured dual-panel pool systems that were strategically structured to ensure comprehensive coverage of the mtDNA genome. These pools were characterized by a well-planned arrangement of small amplicon overlaps, which aimed to achieve an average 11-bp overlap between the two pools. The typical amplicon size was 163 bp, the resulting libraries incorporated an additional ~80 bp due to the presence of barcode adapters. Each of these pool systems consisted of 81 primer pairs per tube, resulting in a total of 162 primer pairs for each kit of the Precision ID mtDNA Whole Genome Panel. In a complete kit, three tubes were provided: 5X Ion AmpliSeq™ HiFi Mix (red cap), Precision ID mtDNA panel Pool 1, and Precision ID mtDNA panel Pool 2 (see Table 2.8). The Precision ID Library Kit complemented this process and included essential components such as FuPa Reagent, Switch Solution, DNA Ligase, and Low TE.

In adherence to the protocol, all samples were prepared following a conservative method, especially designed for non-degraded samples.

2.10.2.1. Ion Chef Library Preparation

During the first step, template preparation, the Ion Chef™ instrument facilitated the amplification and preparation of sequencing templates by combining DNA samples with Ion AmpliSeq™ libraries or target-specific primer sets. It carried out essential steps like emulsion PCR or template fragmentation, ensuring the templates are properly amplified and prepared for subsequent sequencing processes. Then clonal amplification was performed as a second step, during which the Ion Chef™ system employed a proprietary emulsion PCR technology to amplify the prepared DNA fragments onto the Ion Chef™ Chip. This process resulted in the generation of clonal amplicons, where each well on the chip contains a single DNA fragment. The emulsion PCR technique enables efficient amplification and isolation of individual DNA fragments, facilitating downstream sequencing analyses. Finally, the chip loading, the Ion Chef™ instrument produced a chip that contained the clonally amplified DNA fragments that were ready for sequencing. This chip was then loaded manually into an Ion Torrent™ S5 XL System sequencing device.

2.10.2.2. Manual Library Preparation

After fluorometer quantification, all samples were normalized to the required concentration of 0.1 ng of genomic DNA (gDNA) for each targeted amplification reaction. Following that, the samples were prepared for amplification using the pools. Two master mixes were prepared for Pool 1 and Pool 2 in two separate Eppendorf tubes. Table 2.8

listed the components of the mix with the volumes needed. Additionally, two separate MicroAmp 96-Well plates were used for each pool, where feasible. In each well of the MicroAmp 96-Well plates, the mixture for each pool was dispensed, followed by the addition of the adequate amount of the sample. The total volume of each well was 10 μ L. The plate was sealed with a MicroAmp™ Clear Adhesive Film, ensuring a tight seal by applying pressure using the applicator. To prevent evaporation, based on personal preference, an adhesive PCR Plate Foil was applied on top of the clear adhesive layer. Subsequently, each plate was vortexed, and then they were transferred to the centrifuge to collect droplets. Amplification was achieved using the Veriti® 96-Well Thermal Cycler (Applied Biosystems, CA, USA) following the conditions outlined in Table 2.9. Once the amplification reaction was completed, 10 μ L from Pool 2 was transferred into the well containing Pool 1 of the same sample, resulting in a total volume of 20 μ L per well. If the work was to be suspended at this stage, the target amplification reactions could be stored at 10 °C overnight on the thermal cycler or at -20 °C for longer-term storage, for up to one month. During library preparation, each well containing both Pool 1 and Pool 2 of the same sample was be assigned one barcode adapter.

Table 2.8. Target amplification PCR mix preparation.

Pool 1		Pool 2	
Component	Volume	Component	Volume
5X Ion AmpliSeq™ HiFi Mix (red cap)	2 µL	5X Ion AmpliSeq™ HiFi Mix (red cap)	2 µL
Precision ID mtDNA panel Pool 1	5 µL	Precision ID mtDNA panel Pool 2	5 µL
gDNA, 0.1 ng	X µL	gDNA, 0.1 ng	X µL
Nuclease-free water	3- X µL	Nuclease-free water	3- X µL
Total	10 µL	Total	10 µL

Table 2.9. PCR conditions for whole mtDNA target amplification.

Stages	Temperatures	Time Points
Hold	99 °C	2 min
21 Cycle	99 °C	15 s
	60 °C	4 min
Hold	10 °C	∞

The next step in the process involved the digestion of amplicons. To each well, 2 µL of FuPa reagent (brown cap) was added, bringing the total volume to 22 µL in each well. The plate containing the samples was sealed and vortexed to mix the contents thoroughly. Then the plate was centrifuged to collect any droplets and ensure proper mixing of the reagents. Reaching to the PCR step, the plate was loaded onto the thermal cycler, and the conditions specified in Table 2.10 was applied for the digestion step. Once the PCR digestion is completed, this stage provided another stopping point in the process. At this point, the plate could be stored at -20°C for future use or analysis.

Table 2.10. PCR conditions for digestion of target amplicons.

Temperature	Time Points
50 °C	10 min
55 °C	10 min
60 °C	20 min
10 °C	Hold (for up to 1 hour)

The last PCR reaction was the ligation of adapters to the amplicons. Following the PCR is the purification step. Amplicons are ligated to barcode adapters in this step, Ion Torrent™ Dual Barcode Kit 1-96 was used in all the work. Table 2.11 simplifies the addition order of each component with the adequate quantities.

Table 2.11. Perform the ligation reaction.

Order of Addition	Components	Volumes
1	Switch Solution (Yellow Cap)	4 µL
2	Ion Torrent™ Dual Barcode Kit 1-96	2 µL
3	DNA Ligase (Blue Cap)	2 µL
-	Total volume	~30 µL

It was particularly important that the DNA ligase was added last and not combined with other components. After the components' addition was done, the plate was sealed, vortex thoroughly, then centrifuge to collect droplets and loaded in the thermal cycler. Table 2.12 contain the thermal cycler parameters to perform the ligation reaction.

Table 2.12. Thermal cycler parameters of ligation reaction

Temperatures	Time Points
22 °C	30 min
68 °C	10 min
10 °C	Hold (for up to 1 hour)

Post PCR, Samples could be stored overnight at 10°C on the thermal cycler. For longer periods, store at –20°C, serving another accessible stopping point. Finally, the libraries were set for purifying step. Using the MicroAmp plate, the plate seal was carefully removed, then 45 µL of CleanNGS Reagent (Clean NA, ZH, Netherlands) was added to each well (library). This was pipetted up and down 5 times to mix the bead suspension with the DNA thoroughly, then incubated for 5 min at room temperature. The plate was later placed on a magnetic rack then incubated for 2 min or until the solution clears. The supernatant was carefully discarded without disturbing the pellet. A volume of 150 µL of freshly prepared 70% ethanol was added, and the plate was moved side-to-side in the two positions of the magnet to wash the beads. The supernatant was discarded without disturbing the pellet. The ethanol step was repeated for a second wash. Then, the plate was kept in the magnet rack to air-dry the beads at room temperature for 5 min to ensure no traces of ethanol droplets. The plate containing the library was removed from the magnet, and 50 µL of Low TE was added to the pellet to disperse the beads. The plate was sealed, vortexed, incubated for 5 min at room temperature, and then centrifuged to collect droplets. Based on personal preference, the samples could be stored with beads at 4°C for up to one month. Alternatively, for long-term storage at –20°C, the supernatants were transferred to a new plate without beads. Generally, it was avoided to store libraries at –20 °C in the presence of beads. In this study, the supernatants of all samples (libraries) were transferred immediately to a new plate and stored without beads at -20 °C.

The quantity of libraries was determined using the Ion Library TaqMan™ Quantification Kit (Kapa Biosystems, MA, USA) on the QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, CA, USA) (Bender et al. 2004). From the libraries plate, 1:100 dilutions were prepared for each sample. In a new MicroAmp plate, 198 µL of Nuclease-free Water was added to each well. Then, 2 µL of each library (well) from the libraries plate was removed and added to the adjacent well on the dilution plate. Subsequently, the standards were prepared using the *E. coli* DH10B Control Library with the dilution specified in Table 2.13.

Table 2.13. Three 10-fold serial dilutions of the *E. coli* DH10B Control Library

Standards	Control Volumes	NFW Volumes	Concentrations
1	5 µL (undiluted)	45 µL	6.8 pM
2	5 µL Std 1	45 µL	0.68 pM
3	5 µL Std 2	45 µL	0.068 pM

The PCR mixture was prepared by combining 10 µL of Ion Library TaqMan™ qPCR Mix with 1 µL of Ion Library TaqMan™ Quantitation Assay (20X). In a new PCR plate, 11 µL from the prepared mixture was aliquoted into each reaction well. Subsequently, 9 µL each of the diluted library plate, or control library dilution (duplicate of each standard), and negative control was added to the reaction wells. The total reaction volume per well was 20 µL. Following this, the QuantStudio™ 5 Real-Time PCR System was set up, with the parameters outlined in Table 2.14.

Table 2.14. QuantStudio™ 5 Real-Time PCR run parameters

Stages	Temperatures	Times
Hold	50 °C	2 min
Hold	95 °C	20 s
40 Cycles	95 °C	3 s
	60 °C	30 s

After the run was completed, the average concentration of each undiluted library was calculated by taking the mean quantity and multiplying it by the dilution factor of 100. Subsequently, all libraries were diluted to a final concentration of 30 pM, which was the recommended concentration for the Ion Chef™ System. The minimum volume used for each library was 25 µL. In the Planned Run setup, the templating size was set to 200 bp. Following the dilution of the sample libraries to their target concentration in pM, equal volumes of multiple diluted libraries were pooled together. For reference, Figure 2.3 showed an example of two library calculations. Libraries could also be calculated through a newly developed equation that was validated through the scientific team at Thermo Fisher directly from the Qubit results. The average library size in the equation is always set to 200.

$$L_{nM} = \frac{Qubit_{ng/uL}}{660_{g/mol} \times Avg.Lib.size} \times 10^6$$

<u>Well Position</u>	<u>Barcode Ion Dual 1 - 96</u>	<u>Barcode Position</u>	<u>Sample Number</u>	<u>Sample Name</u>	<u>Quantity</u>	<u>Multiply by 100</u>	<u>Volume of lib 100 pm</u>	<u>vol of water</u>	<u>Final Loading (Pm)</u>	<u>Prepared Diluted Library</u>	<u>Volume of total pooled lib</u>	<u>Vol of water</u>
G2	55	G7	Sample 15	Ref Sample: Promega	7.601	760.1	3.95	26.05	30	2	7.5	17.5
H2	56	H7	Sample 16	Ref Sample: ORIGene	7.642	764.2	3.93	26.07	30	2		
											25 ul	

Figure 2.3. Library dilutions to target concentration 30 pM.

Use the pooled libraries in template preparation reactions on the Ion Chef™ Instrument.

The last step in manual library preparation was libraries loading on Ion Chef™ for chipping.

For Precision ID mtDNA Whole Genome Panel, Ion 520™ Chip and Ion 530™ Chip were compatible. For the Ion 520™ Chip, it supported the loading of 25 samples maximum, while Ion 530™ Chip supports 32 samples. The screen-prompt step by step set up was followed on the instrument. The screen-prompt was followed to aid the load Ion S5™ Precision ID Chef Reagents, Ion S5™ Precision ID Chef Solutions, Ion S5™ Precision ID Chef Supplies and Ion 530™ Chip. Figure 2.4 is an illustration of the Ion Chef deck.

Torrent Suite™ Software was required for template preparation.

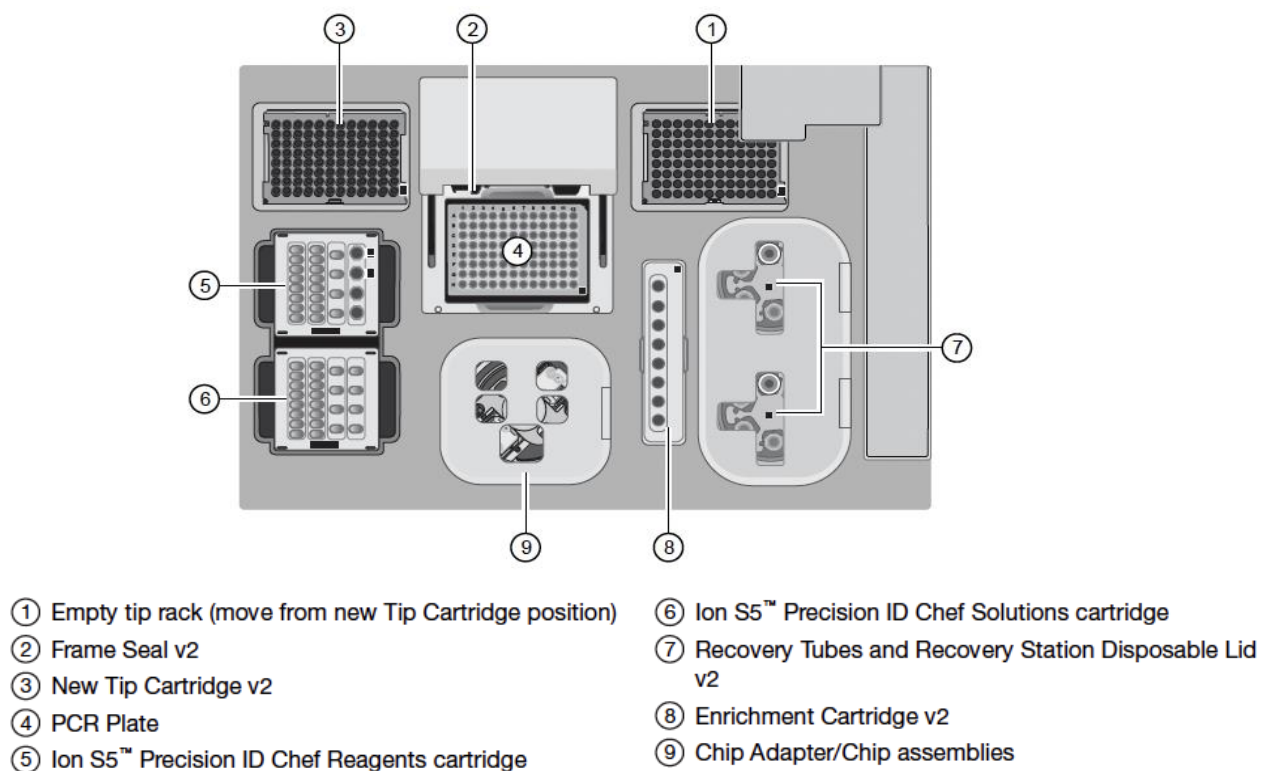


Figure 2.4. Ion Chef™ Instrument deck (Illustration obtained from Precision ID mtDNA Panels with the HID Ion S5™/HID Ion GeneStudio™ S5 System Application Guide)

Sign into Torrent Suite™ Software, on the dashboard “Plan” tab was selected, then “Templates” was chosen. Then Human Identification > Precision ID mtDNA Whole Genome Panel - S5, navigated to Plan screen. Table 2.15 summarized the different options with the appropriate action.

Table 2.15. Torrent Suite™ Software Plan screen setup

Options	Actions for Precision ID mtDNA Whole Genome Panel
Run Plan Name	Enter a name for the run plan.
Analysis Parameters	Select Default (Recommended).
Reference Library	Select PrecisionID_mtDNA_rCRS(Mito)
Target Regions	Precision_ID_mtDNA_Whole_Genome_Panel_Targets_vX.X.bed.
Hotspot Regions	Select None
Number of chips	1 (New plan will be created to each chip regardless of chips loaded together)
Sample	Enter a name for each sample
Sample ID	Optional
Sample Description	Optional
Reference	Leave blank
Sample Tube Label	Fill according to the tube label (Ex: C0940464)
Chip Barcode	Fill according to the chip label (Ex: DAFJ00872)
Bead Loading (%)	Default (30)
Key Signal	Default (30)
Usable Sequence (%)	Default (30)

After the setup on the Torrent Suite™ Software dashboard, the setup was proceeded to the Ion Chef™ Instrument. On the touch screen, “Set up run” was chosen to initiate the instrument, followed by Quick start option. Instructions were followed as per the prompted on the screen. When prompted, the instrument door was closed by first lifting it slightly to disengage the locking mechanism, and then it was pushed down until the locks engaged. After the door closed, the instrument vision system activated, and "Start check" was tapped to initiate Deck Scan. Upon completion of Deck Scan, "Next" was tapped to access the Data Destination screen. The information displayed on the screen

was carefully verified to ensure it included the correct kit type, chip type, chip barcodes, and planned run. If necessary, the correct options could also be chosen from the dropdown list, and "Next" was tapped. On the "Run Options" screen, the desired completion time for the run was entered, allowing adjustment of the run termination to coincide with the return to the lab, ensuring that the chips were not left in the instrument for an extended period. Once one chip was later loaded on the sequencer, the other chip was stored at 4 °C until the sequencing of the first chip was complete. Finally, "Start run" was tapped. After the run was concluded, the deck was cleaned by removing and disposing of the used consumables.

2.10.3. DNA Sequencing

The Ion Torrent sequencer employed in this study was the Ion S5 XL, a highly adaptable platform capable of supporting a diverse array of applications, ranging from small targeted panels to whole exome sequencing.

Following "Library Preparation" step, 32 mtDNA template libraries were loaded onto an Ion 530™ Chip. Subsequently, these libraries underwent sequencing on the Ion Torrent™ Ion S5™ XL System using the selected protocol for 500 nucleotide flows as specified during the "Plan" setup. The reporting during sequencing was based on pH changes resulting from the production of hydrogen ions during the incorporation of each nucleotide into the newly formed DNA strand. The reagents employed in the procedure included the Ion S5™ Precision ID Sequencing Reagents cartridge, Precision ID Wash Solution bottle, and Precision ID Cleaning Solution bottle. Within the instrument, there

was the chip clamp and waste reservoir. The system was equipped with built-in RFID technology, allowing identification through tags and readers. Reagents exceeding their expiration date or usage count triggered an error message, prompting the user to replace the reagent before initiating the run.

2.10.3.1. Initialization

The instrument underwent an initialization step initially, with touchscreen prompts guiding the positioning of components. On the Ion GeneStudio touchscreen, the "Initialize" button became active and ready for operation. The initialization process, lasting approximately 30–40 minutes, ensured the proper installation of all required consumables. The "Initialize" button was clicked and the door, chip, and Reagent cartridge clamps were unlocked. Following the prompted start, the Precision ID Wash Solution bottle was removed to access the waste reservoir. The waste reservoir was then removed, emptied, and reinstalled. Subsequently, new Precision ID Wash Solution and Precision ID Cleaning Solution bottles were installed, along with a new cartridge of Ion S5™ Precision ID Sequencing Reagents. For initialization purposes only, an old chip was placed in the chip clamp, and after closing the door, "Next" was clicked on the screen. The instrument confirmed the proper installation of consumables and the chip, as well as verified that the Precision ID Cleaning Solution bottle contained sufficient reagent for the post-run clean. All on-screen instructions were followed. Upon completion of initialization, "Home" was tapped.

2.10.3.2. Sequencing

Upon completion of initialization, "Run" was displayed on the touchscreen of the instrument. Clicking the "Run" button released the door, and the chip clamp was unlocked. The chip to be sequenced was secured in the clamp, which was then returned to the locked position. After closing the door, "Next" was clicked. The correct Planned Run automatically appeared upon chip detection. Typically, two sequencing runs per single initialization were performed. For run 1 out of 2, the "Enable post-run clean" checkbox was deselected before tapping "Review." All the displayed information was confirmed, and then "Start run" was tapped. After the completion of the run, the instrument was ready for the second chip load. At this stage, "Enable post-run clean" was selected, allowing the instrument to automatically perform the cleaning procedure. During the sequencing process, a sensor detected changes in pH, and the nucleotide was identified based on the pH signal. The Ion Torrent sequencing software interpreted the electrical signal, identifying the nucleotide incorporated at each position in the DNA strand. The software generated a readout of the sequence data, utilized for downstream analysis with Converge™ software (Thermo Fisher Scientific, USA).

2.10.4. Statistical data Analysis

Statistical data analysis was done using various metrics to characterize genetic variation across the Emirati, Indian, and Pakistani sample sets. Key measures included the Number of Haplotypes (Ht) and Number of Haplogroups (Hg) to evaluate the diversity of haplotypes and haplogroups within each population. Haplotype Diversity (Hd) and

Haplogroup Diversity (HgD) were calculated to further quantify the genetic variability among individuals and groups. Additionally, Probability of Discrimination (PD) was used to assess the ability to differentiate individuals based on their mtDNA profiles, while Probability of Identity (PI) provided insight into the likelihood of two randomly selected individuals sharing the same haplotype. Together, these methods offered a comprehensive framework for understanding genetic diversity to support the forensic applications across the studied populations.

Chapter 3

3. Sanger Sequencing Analysis of the Control Region

3.1.Introduction

The Sanger sequencing technique is widely recognized as the standard method for DNA sequencing and is widely applied in forensic laboratories for mtDNA examination. Numerous forensic databases are built on Sanger sequencing and it has been used to characterise mtDNA in questioned biological samples to aid the inclusion or exclusions of potential suspects, through the detection of SNPs and Indels, mainly in the control region of the mtDNA (Wilson and Allard, 2005).

There are numerous approaches published in the literature to analyse the control region and reliable methodologies which differ depending on the targeted control region, number of primers, limited kits, or region of interest. Primer sets can vary from single to multiple sets, to cover single region of the control region, or the whole control region encompassing HV1, HV2 and HV3 (Parson and Dür, 2007; Just et al., 2015; Bär et al., 2000).

In forensic genetic analysis dye terminator chemistry is the gold standard which is employed for sequencing, however primer design and number of primers used to analyse the sequence along with this chemistry must undergo optimization and standardized prior to application (Butler, 2012; Wilson et al., 2014).

3.2. Chapter Aim

The aim of this chapter of the study was process samples from three population groups using Sanger sequencing to generate concordance data, serving as a baseline for the subsequent implementation of MPS technology. This included optimization of extraction methods, validating the sequencing methodology, establishing haplotypes for comparison, and ensuring data consistency before transitioning to high-throughput sequencing methods.

3.3. Chapter Objectives

- To optimize samples extraction methods
- To validate a set of primers for mtDNA control region amplification.
- To optimize the sequencing methodology.
- To carry out sequence 30 Emiratis samples, 30 Indians samples and 30 Pakistanis samples to establish the data for a concordance study

3.4. Methods

See for details of methodology in chapter 2.

3.4.1. Sanger sequencing Optimization

Sanger sequencing begins with the amplification of a targeted mtDNA control region, followed by sequencing. This strategy of the study was to amplify the whole control region. The primer set chosen for this study was based on a review of previous published works (Wilson et al., 2002). Additional primer sets were incorporated to optimize sequencing

using the BigDye™ Terminator Cycle Sequencing Ready Reaction Kit v3.1 (Thermo Fisher Scientific, USA). The resulting products were then analysed using an ABI 3500 capillary electrophoresis (Thermo Fisher Scientific, USA). Product purification was tested with different commercial kits to compare the amount of DNA recovered. Different extraction methods were also tested to assess the best extraction method for the sample type being used.

3.4.2. DNA Extraction

In this study, optimization was firstly carried out to choose the best extraction method to carry out the downstream analysis. Extraction methods tested were direct amplification from Whatman® FTA® cards and PrepFiler®-based extraction of FTA® cards.

The use of Whatman® FTA® cards (Qiagen, Hilden, Germany) has become a standard practice for reference samples DNA extraction, purification, and storage in forensics workflow. These cards are reliable for preserving blood samples at room temperature in a dry, clean environment. In the process of optimizing the steps of Sanger sequencing two candidate methods were tested: firstly, direct amplification from 1.2 mm disk of purified FTA® cards stained with blood, and secondly, extraction of blood on FTA® using PrepFiler®-based extraction that was semi-automated using Hamilton® robotics. This selection was based on the methods available at the Police laboratory in Dubai. DNA samples extracted using PrepFiler® were eluted into 40 µL in order to have adequate amount of extracted DNA for multiple downstream analyses . All DNA extracts generated

using PrepFiler® were quantitated using Quantifiler™ and range between 26 - 38 ng/μL of nuclear DNA (nDNA).

3.4.3. Control Region Amplification

Initially in this study, a single primer set was utilized to amplify mtDNA control region (HV1, HV2, HV3) ranging from positions 16024 to 574: forward primer L15879 (5'- AAT GGG CCT GTC CTT GTA GT -3') and reverse primer H727 (5'- AGG GTG AAC TCA CTG GAA CG -3') (Cardena et al., 2013). This primer set was used to generate a fragment 1417 bp long. To ensure the specificity of the amplification, it is recommended to use primers with known sequences that have been extensively optimized by previous researchers (Wilson et al., 2002). These primers had been shown to specifically target mitochondrial DNA (mtDNA) while excluding most nuclear-mitochondrial DNA (NUMT) sequences. By minimizing the inclusion of NUMTs, the accuracy and reliability of mtDNA analysis are enhanced (Parson and Bandelt, 2007). However, given that full-length NUMT insertions can occur in the genome, it is challenging to completely avoid their amplification. The primer set mentioned was used to amplify the mtDNA control region following the PCR conditions outlined in Table 2.4. The PCR reaction comprised 0.5 μL of each primer (forward and reverse), 5 μL of either Platinum® or Reddymix™ master mixes, and DNA template (4 μL of extracted DNA or 1.2 mm disk of purified FTA® with an addition of 3 μL PCR grade water).

Four samples were chosen to run the comparison for extraction method and PCR optimization. Each sample underwent evaluation of the extraction methods, direct punch

against extraction. The extracted DNA of the same sample was later evaluated for master mixes Platinum® and Reddymix™.

Table 3.1. Four samples used for comparison for FTA® card and PrepFiler® extraction, and two master mixes Platinum® and Reddymix™

Sample Names	Extractions	Master Mix	Agarose Gel Lanes
Sample 1A	FTA® card	Reddymix™	2
Sample 1B	FTA® card	Platinum®	3
Sample 1C	PrepFiler®	Reddymix™	4
Sample 1D	PrepFiler®	Platinum®	5
Sample 2A	PrepFiler®	Reddymix™	6
Sample 2B	PrepFiler®	Platinum®	7
Sample 2C	FTA® card	Reddymix™	8
Sample 2D	FTA® card	Platinum®	9
Sample 3A	PrepFiler®	Reddymix™	10
Sample 3B	PrepFiler®	Platinum®	11
Sample 3C	FTA® card	Reddymix™	12
Sample 3D	FTA® card	Platinum®	13
Sample 4A	FTA® card	Reddymix™	14
Sample 4B	FTA® card	Platinum®	15
Sample 4C	PrepFiler®	Reddymix™	16
Sample 4D	PrepFiler®	Platinum®	17

Amplification of the mtDNA control region was initially evaluated using gel electrophoresis. The technique allowed visual assessment to the integrity and size of the DNA fragments.

Simultaneously, the master mix efficiency was evaluated on the agarose gel as well. Comparison of master mixes was assessed by the intensity of the amplicon and amount of primer dimers. The representative images of the agarose gel electrophoresis analysis

of the DNA samples are shown in Figure 3.1. Based on the results, both approaches of extraction (FTA® card and PrepFiler®) were successful. However, the PrepFiler® extraction method resulted in higher amounts of the PCR product. This is most likely due to the higher amount of template mtDNA initially presented in the PCR, when 4 µL of eluted DNA was used compared 1.2 mm washed disc of blood-spotted FTA card. The extracted DNA samples were found to exhibit very distinct PCR-amplified bands on the agarose gel, without any smearing effect. Also, there is a possibility that FTA® cards may have residues of some inhibitors following washing.

Interestingly, both Platinum® Taq Polymerase and 2X ReddyMix™ master mixes shown the same intensity on gel electrophoresis and same pattern between direct amplification of FTA® disk and extracted DNA and ultimately both yielded good quality results. A decision was made to use DNA extraction and proceed with the 2X ReddyMix™ to make the process more cost effective.

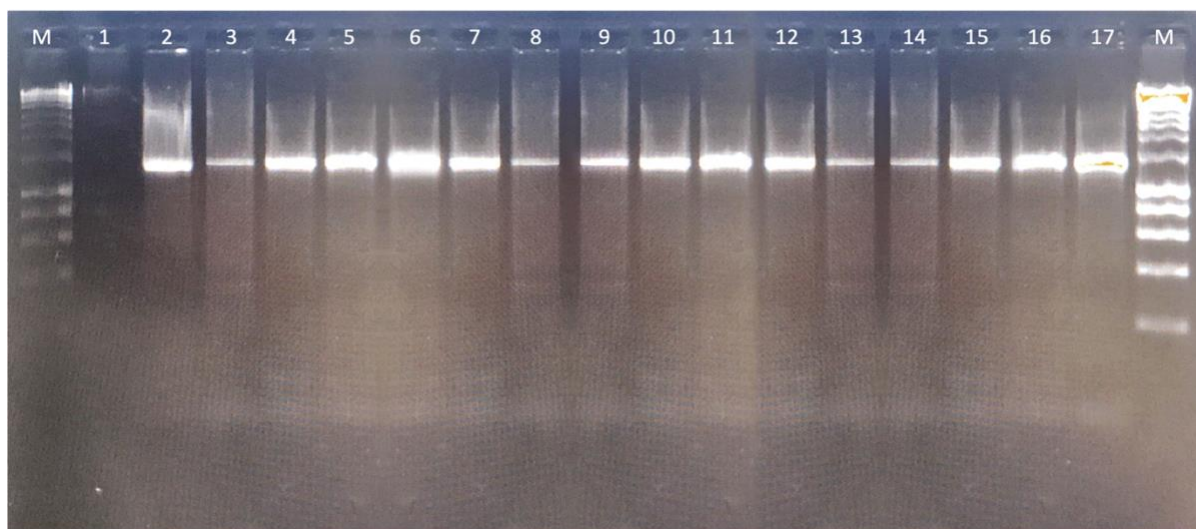


Figure 3.1. A captured run of samples on 2% agarose gel electrophoresis. HyperLadder™ 1 kb is in lane M (far right and left of the sample wells). Lane 1 is the negative control. Lanes 2-5, 6-9, 10-13, and 14-17 corresponds to female blood samples 1, 2, 3, and 4 respectively, with four replicates (A-D). See Table 3.1 for sample details.

3.4.4. BigDye® Terminator Sequencing

Based on the optimization results, both extraction method and control region amplification, PrepFiler® extraction was chosen to proceed downstream with amplifications undertaken in 10 µL Reddymix™ reaction volume. To evaluate the result of the amplified product by the set of primers, 6 samples that were extracted using PrepFiler® and processed with Reddymix™ (from the earlier steps) were chosen to be further analyzed using BigDye® terminator sequencing. Prior to the BigDye® terminator reaction, a purification step was employed. All amplified products were purified using MicroClean® reagent. Following this step, each amplified control region of mtDNA was sequenced twice, initially, using both forward primer (L15879) and reverse primer (H727) in duplicate. Following the BigDye® Terminator reaction, the sequenced products were purified using two approaches, 3 samples were purified using DNA precipitation and the

other 3 samples were purified using the BigDye XTerminator® Purification Kit. The amplicons after the cycle sequencing were then separated using capillary electrophoresis (3500 Genetic Analyzer, Thermo Fisher Scientific, USA) after adding 11 µL HiDi formamide to the cleaned product and loaded into a MicroAmp™ Optical 96-well reaction plate. The sequence data were analysed using Sequencing Analysis Software v5.4 (Thermo Fisher Scientific, USA). An example of the precipitation purification method is shown in Figure 3.2 and Figure 3.3 for comparison. The results shows an example of a sample cleaned using BigDye XTerminator® Purification Kit. The DNA precipitation method was chosen to proceed with the rest of the samples. Figure 3.4 illustrates the electropherogram obtained using the reverse primer (H727) and the precipitation purification method with ethanol, showing a smooth baseline. In contrast, Figure 3.5 displays the electropherogram of the mtDNA control region sequence with noticeable baseline noise using the reverse primer (H727) after purification with the BigDye XTerminator® Kit."

Although the quality of the recovered sequences was very high with up to 600 bp of sequence for each primer, forward and reverse, it was still not enough to span the whole control region of mtDNA, as signals started to fall in both directions. To overcome such issue, two additional primers were used for sequencing with priming positions approximately 100 bp ahead of the primers initially used for amplification of the control region. These primers were F15975 (5'- CTC CAC CAT TAG CAC CCA AA -3') and reverse R635 (5'- GAT GTG AGC CCG TCT AAA CA -3') (Lee et al. 2010).

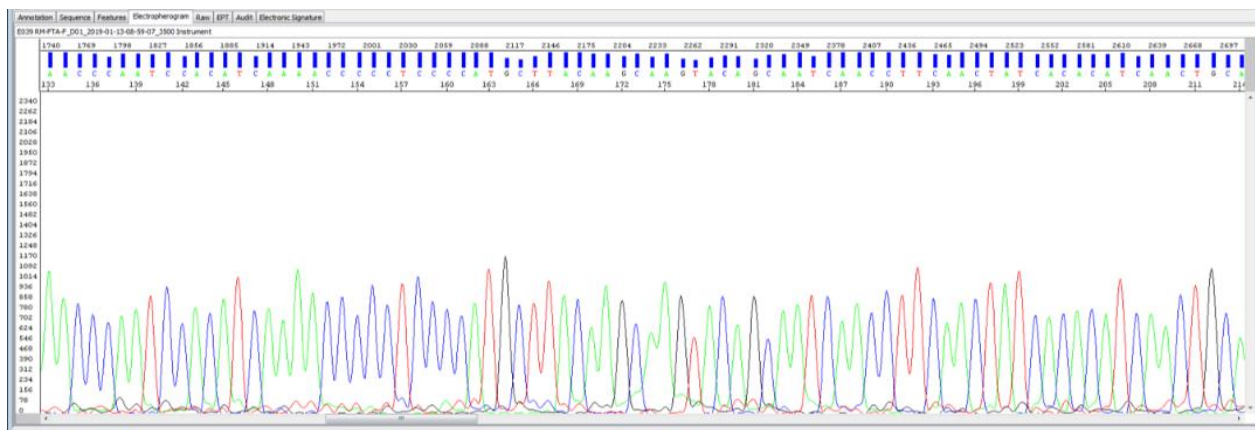


Figure 3.2. An electropherogram showing the mtDNA control region sequence with smooth baseline using forward primer (L15879) using precipitation purification method with ethanol. In this and subsequent figures, electropherogram is typical of four different experiments.

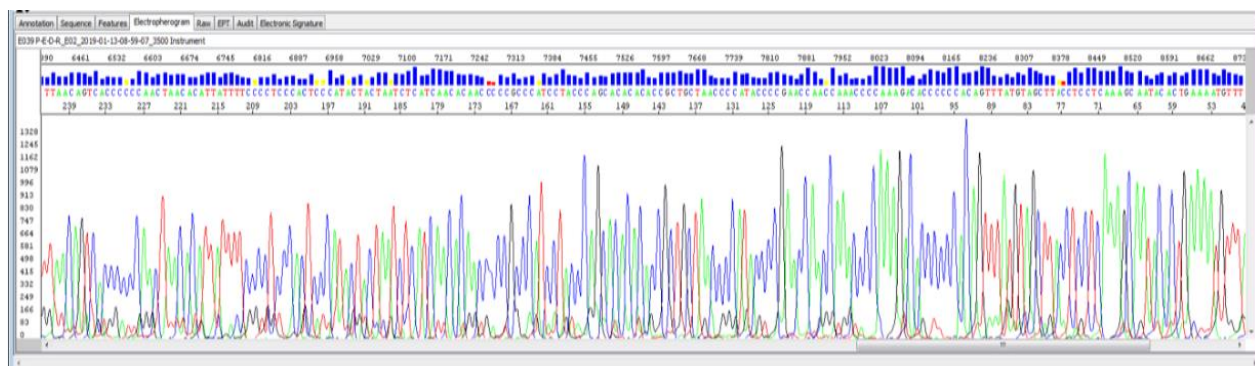


Figure 3.3. An electropherogram of the mtDNA control region forward primer (L15879) sequence with high baseline noise using BigDye XTerminator® Purification Kit.

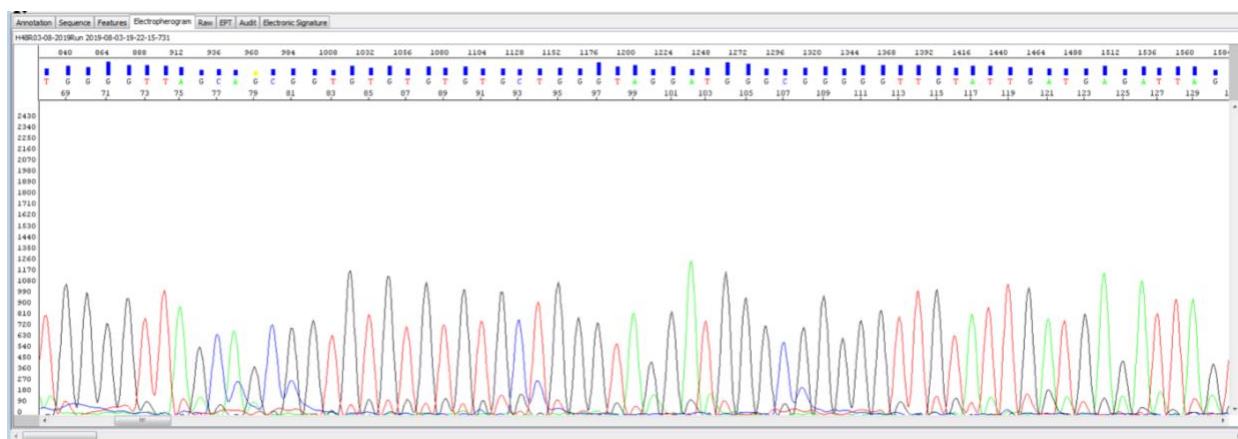


Figure 3.4. An electropherogram showing the mtDNA control region sequence with smooth baseline using reverse primer (H727) using the precipitation purification methods with ethanol.

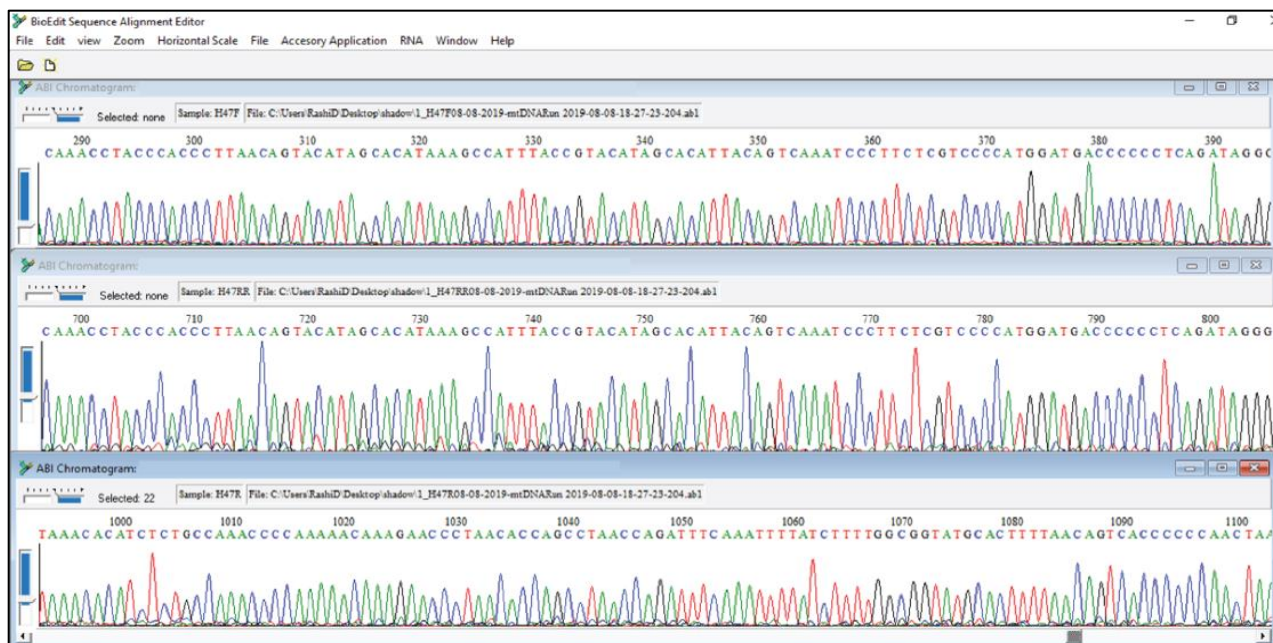


Figure 3.6. BioEdit™ software result showing an example of sequence quality generated for one sample using three primers F15975 (top), R240 (middle), and R635 (bottom).

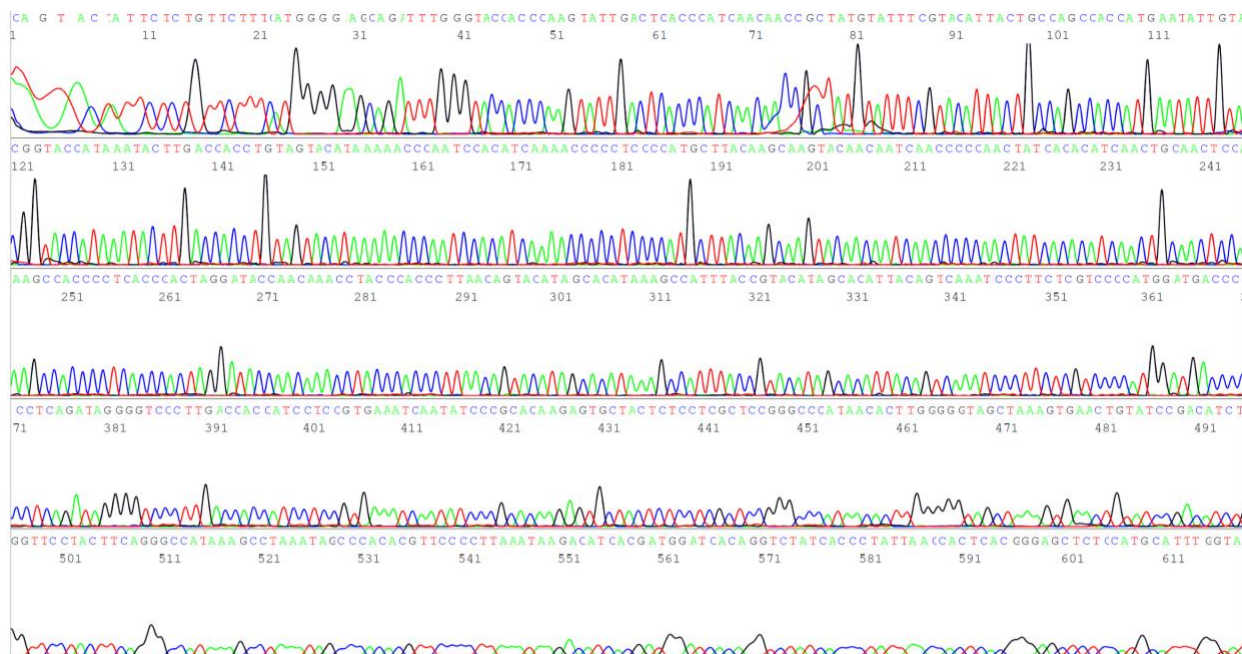


Figure 3.7. Electropherogram of a single random sample showing the dropping effect in sequencing the control region using a single forward primer (F15975).

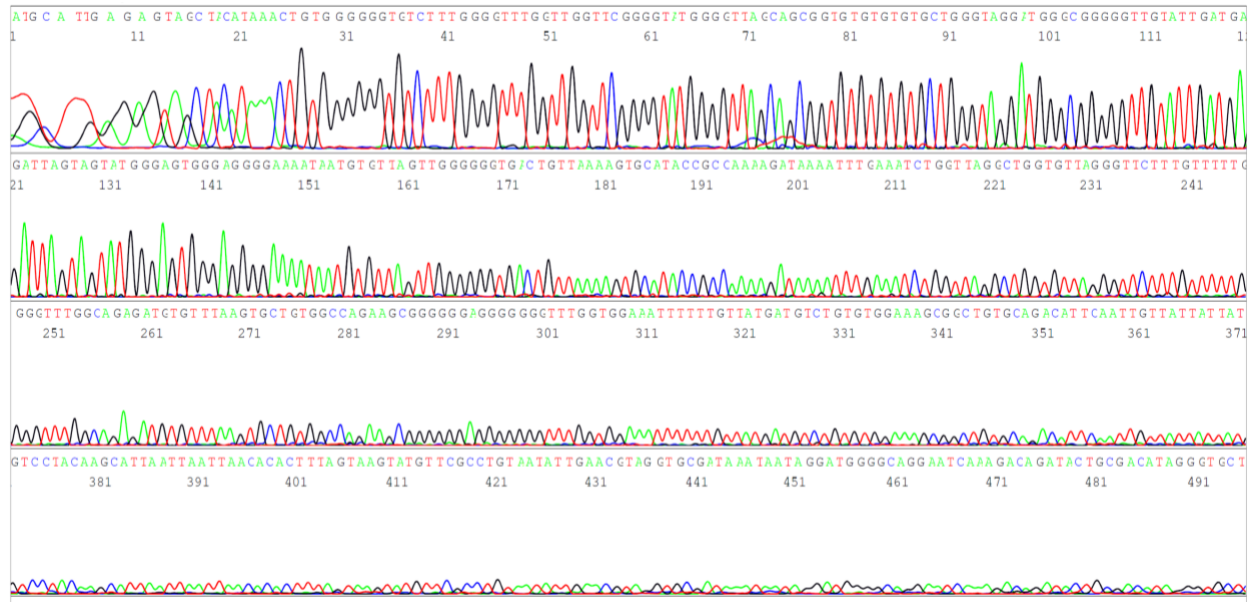


Figure 3.8. Electropherogram of a single random sample capturing the sequence drop of the control region generated using a single reverse primer (R635) (Sequence viewed without the reverse option).

Based on the work done for optimization, methods that were concluded to be used in the next population analysis was PrepFiler® extraction, and Reddymix™ for master mix. The Primers used for the first PCR were forward primer (L15879), and reverse primer (H727). For the BigDye® Terminator reaction additional primers were selected, F15975 as forward primer, and R635 and R240 as reverse primers and the precipitation method was chosen for sequence product purification.

3.5. Populations Analysis using Sanger sequencing

Based on the optimization results, Sanger sequencing approach was applied to analyze the control region for 93 samples. Those samples comprised 33 samples of Emiratis, 30 samples of Pakistanis and 30 samples of Indians residing UAE. All samples were extracted using PrepFiler® and quantitated using Quantifiler™ Kit. The threshold chosen, was 20% for minor variants or minor allele frequency (where the minor peak exceeds 20%).

All samples were successfully extracted and the control region for each sample was sequenced. In case of weak called samples or no called sequence lost samples due to inhibitions or precipitation were, samples were reanalyzed. BioEdit software (Hall, 1999) and 4Peaks software (Nucleobytes, <https://nucleobytes.com/4peaks/index.html>) were utilized for alignment and processing of the sequences.

To identify variations in mtDNA, the conventional protocol of aligning the sequence data obtained from the specimen with the revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999) was followed. While best practice often involved sequencing DNA in both directions to ensure accuracy and detect potential sequencing errors, this was not deemed critical in this study as the primary purpose of the data was to serve as control data. Following alignment and trimming of sequences in the .ab1 file, manual examination of the electropherograms was undertaken to detect all variants and minimise phantom mutations (Kloss-Brandstätter et al., 2005). Table 3.2 summarizes the manual assessment of the Emirati samples. The variations were evaluated in accordance with the criteria outlined by the Scientific Working Group on DNA Analysis Methods (2013). Subsequently, once haplotypes were assigned for each specimen, they underwent quality control assessment through EMPOP query to verify the results (Parson and Dür., 2007, Zimmermann et al., 2011). The Pakistanis and Indians samples underwent the same procedures, and the tables are reported in the Appendices (Appendix V and VI).

Table 3.2. Control region sequences for 33 Emiratis samples with the haplotypes nomenclatures

No.	Samples	Haplotypes
1	UAE_01	16319A 263G 315.1C Transition: 2 Transversion: -
2	UAE_02	16093C 16145A 16176G 16223T 16390A 16519C 73G 152C 263G 315.1C Transition: 8 Transversion: 1
3	UAE_03	16093C 16145A 16176G 16223T 16390A 16519C 73G 152C 263G 315.1C Transition: 8 Transversion: 1
4	UAE_04	16086C 16292T 16519C 73G 237G 263G 315.1C 373G 482C Transition: 8 Transversion: -
5	UAE_05	16224C 16256T 16311C 16519C 73G 263G 497T Transition: 7 Transversion: -
6	UAE_06	16126C 16265G 16355T 16362C 58C 64T 146C 263G 302.1AC 315.1C Transition: 8 Transversion: -
7	UAE_07	16126C 16163G 16186T 16189C 16294T 16519C 73G 152C 195C 263G 309.1C 315.1C Transition: 10 Transversion: -
8	UAE_08	16176T 195C 263G 315.1C Transition: 3 Transversion: -
9	UAE_09	16124C 16223T 16319A 73G 150T 152C 263G 315.1C Transition: 7 Transversion: -
10	UAE_10	16017C 16129A 16163G 16187T 16189C 16209C 16223T 16278T 16293G 16294T 16311C 16360T 16519C 73G 151T 152C 182T 186A 189C 263G 315.1C Transition: 19 Transversion: 2
11	UAE_11	16124C 16223T 16319A 73G 150T 152C 263G 315.1C Transition: 7 Transversion: -
12	UAE_12	16189C 16519C 263G 315.1C Transition: 3 Transversion: -
13	UAE_13	16213A 16224C 16311C 16519C 73G 146C 152C 263G 315.1C Transition: 8 Transversion: -
14	UAE_14	16482G 200G 239C 263G 302.1AC 315.1C 573.1C Transition: 5 Transversion: -
15	UAE_15	16183d 16189C 16223T 16278T 16290T 16294T 16309G 16390A 73G 146C 152C 195C 263G 315.1C

		Transition: 12
		Transversion: -
16	UAE_16	16172C 16298C 16343G 73G 150T 195C 263G 315.1C
		Transition: 7
		Transversion: -
17	UAE_17	16213A 16224C 16311C 16519C 73G 146C 152C 263G 315.1C
		Transition: 8
		Transversion: -
18	UAE_18	16069T 16126C 16207G 16519C 73G 95C 150T 152C 195C 263G 295T 315.1C
		489C
		Transition: 11
		Transversion: 1
19	UAE_19	16189C 16261T 16295T 16519C 73G 263G 271T 315.1C
		Transition: 7
		Transversion: -
20	UAE_20	16129A 16223T 16311C 16391A 16519C 73G 146C 199C 204C 207A 250C
		263G 315.1C
		Transition: 12
		Transversion: -
21	UAE_21	16354T 263G 315.1C
		Transition: 2
		Transversion: -
22	UAE_22	16126C 16355T 16362C 58C 64T 146C 263G 315.1C
		Transition: 7
		Transversion: -
23	UAE_23	16051G 16278T 73G 263G 315.1C 499A
		Transition: 5
		Transversion: -
24	UAE_24	16129A 16189C 16223T 16249C 16311C 16359C 16519C 73G 195C 263G 315.1C
		489C
		Transition: 11
		Transversion: -
25	UAE_25	16223T 16325C 16519C 73G 189G 194T 195C 204C 207A 263G 315.1C
		Transition: 10
		Transversion: -
26	UAE_26	16311C 16343G 16390A 16519C 73G 143A 150T 152C 189G 200G 263G 315.1C
		Transition: 11
		Transversion: -
27	UAE_27	16093C 16300G 16362C 16482G 16519C 152C 239C 263G 315.1C
		Transition: 8
		Transversion: -
28	UAE_28	16069T 16126C 16193T 16300G 16309G 73G 152C 263G 295T 315.1C 462T 489C
		Transition: 11
		Transversion: -
29	UAE_29	16069T 16126C 16193T 16300G 16309G 73G 152C 263G 295T 315.1C 462T 489C
		Transition: 11
		Transversion: -
30	UAE_30	16051G 16129C 16183d 16189C 16362C 16519C 73G 152C 217C
		263G 315.1C 340T 499A 508G
		Transition: 11

		Transversion: 1
31	UAE_31	16168T 16343G 16519C 73G 150T 199C 263G 315.1C Transition: 6 Transversion: -
32	UAE_32	16201T 16220G 16223T 16265G 16497G 16519C 73G 189G 195C 204C 207A 210G 263G 315.1C Transition: 13 Transversion: -
33	UAE_33	16223T 235G 263G 315.1C Transition: 3 Transversion: -

Each haplogroup comprised haplotypes that exhibited a specific set of variant patterns compared to the rCRS, termed a phylogenetic motif. These haplogroups and motifs are documented in Phylotree(mt) (<http://phylotree.org/>, accessed on October 2, 2022), which most recently updated to build 17 (PT17) (Van Oven, 2015). This update included data from 24,275 full mitochondrial genome sequences, identifying 5,435 haplogroup-defining motifs. Since the release of PT17 in 2016, the database of mitochondrial genomes has grown, with many sequences not aligning with the existing tree structure (Dur et al., 2021). To address these discrepancies, Dur et al. revised PT17 by analyzing 26,011 mitochondrial genomes, increasing the recognized haplogroup motifs by 18% to a total of 6,401. This update significantly enhanced the accuracy of haplogroup estimations.

Initially, the alignment and comparison of samples with the Cambridge Reference Sequence (rCRS) in the control region were conducted manually. Subsequently, another online tool mtProfiler (<http://mtprofiler.yonsei.ac.kr>) was utilized for sequence analysis to obtain preliminary data before verification through EMPOP. This tool, equipped with an automated mtDNA nomenclature feature, facilitates sequence alignment and mtSNP

calculation, employing the Parsimonious Smith-Waterman algorithm. Additionally, it included an "mtDNA assembly tool," beneficial for reconstructing the expected sequence from fragmented sequences obtained through multiple amplification reactions, particularly advantageous when dealing with sequences spanning the control region of mtDNA. Furthermore, the tool offered an "mtSNP conversion tool," allowing the conversion of mtSNP data, represented as deviations from the rCRS, into FASTA format for alignment comparison. This tool also enabled users to verify the validity of mtSNP data by comparing the input data with recalculated mtSNP data. Lastly, the "mtSNP concordance-check tool" conducts a concordance analysis between mtSNP data from two independent experiments on the same sample, with the sequence range of input mtSNP data automatically determined. Additionally, it verified the validity of mtSNP data through comparison with recalculated mtSNP data.

In this work, the tool that utilized was the "mtDNA nomenclature", as shown in Figure 3.9, and then was used for all samples. Later, HaploGrep3 performed haplogroup determination by evaluating the phylogenetic significance of polymorphisms found in a haplotype. This evaluation was informed by their occurrence and prevalence within PT17-FU and accounted for both mutation hotspots and super-hotspots. The software provided the top ten possible haplogroup matches for each haplotype, along with a quality score expressed as a percentage, derived from the phylogenetic significance of the matched haplogroup. The highest-ranking haplogroup match was then automatically assigned to the haplotype. Furthermore, each haplotype was visually distinguished by a colour code

reflecting the quality score of its best-matched haplogroup: green indicates a score above 90%, yellow signifies scores between 80% and 90%, and red denotes scores below 80%.

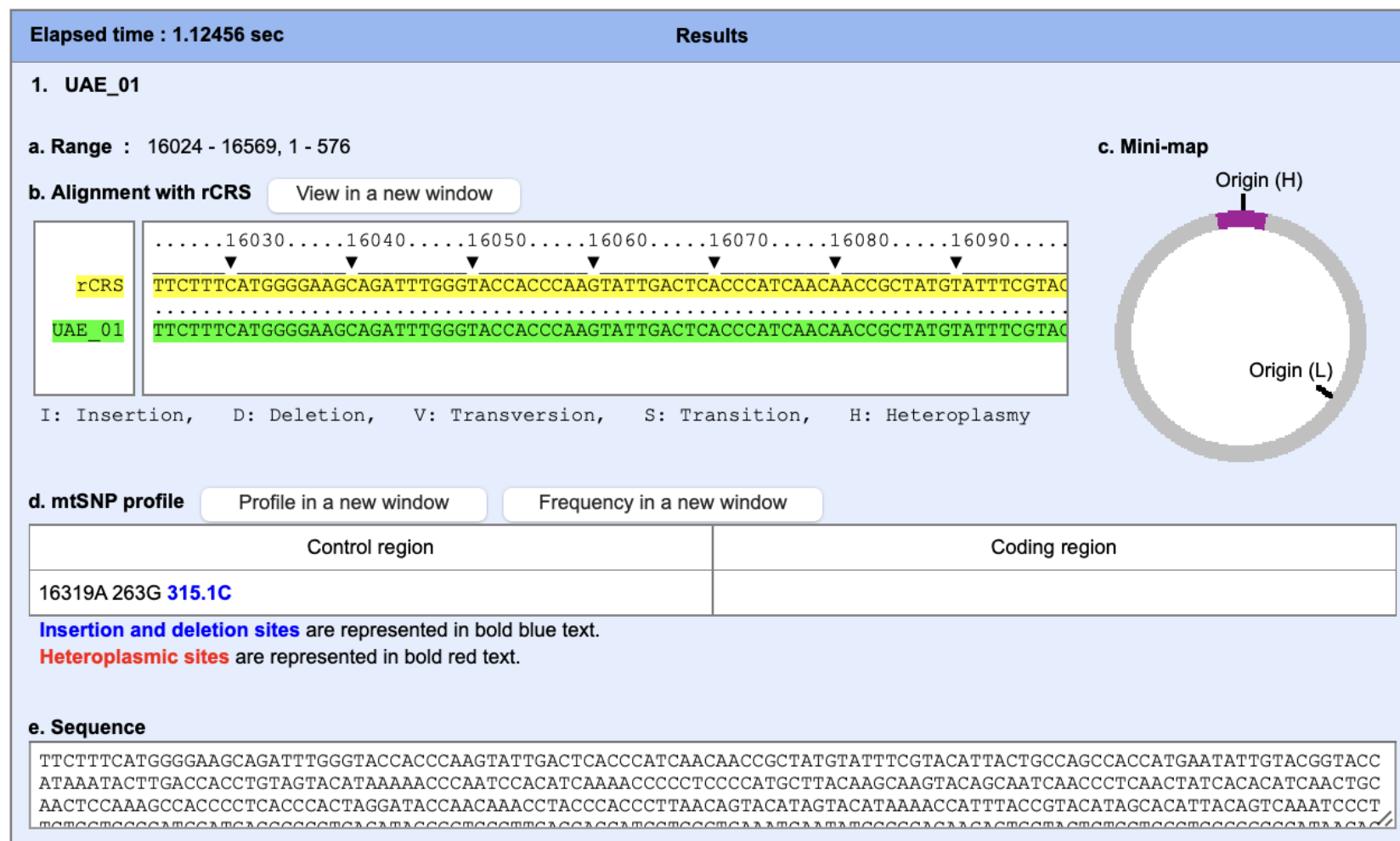


Figure 3.9. Results of the mtDNA control region analysis performed using the mtProfiler tool (<http://mtprofiler.yonsei.ac.kr>)

The variations in each sample were recorded, and full polymorphisms generated the haplotypes, which were determined both by the nomenclature tool and HaploGrep tables from the previous steps (Figure 3.10). During this process, some samples triggered warnings or failed alerts, indicating potential issues with the haplotype assignments. Warnings could arise for several reasons, such as the sample containing undetermined variants (N), having more than two global private mutations that are not recognized by Phylotree, or containing local private mutations associated with other haplogroups. Failed alerts typically indicated a low detected haplogroup quality (quality $\leq 80\%$) or missing more than two expected polymorphisms.

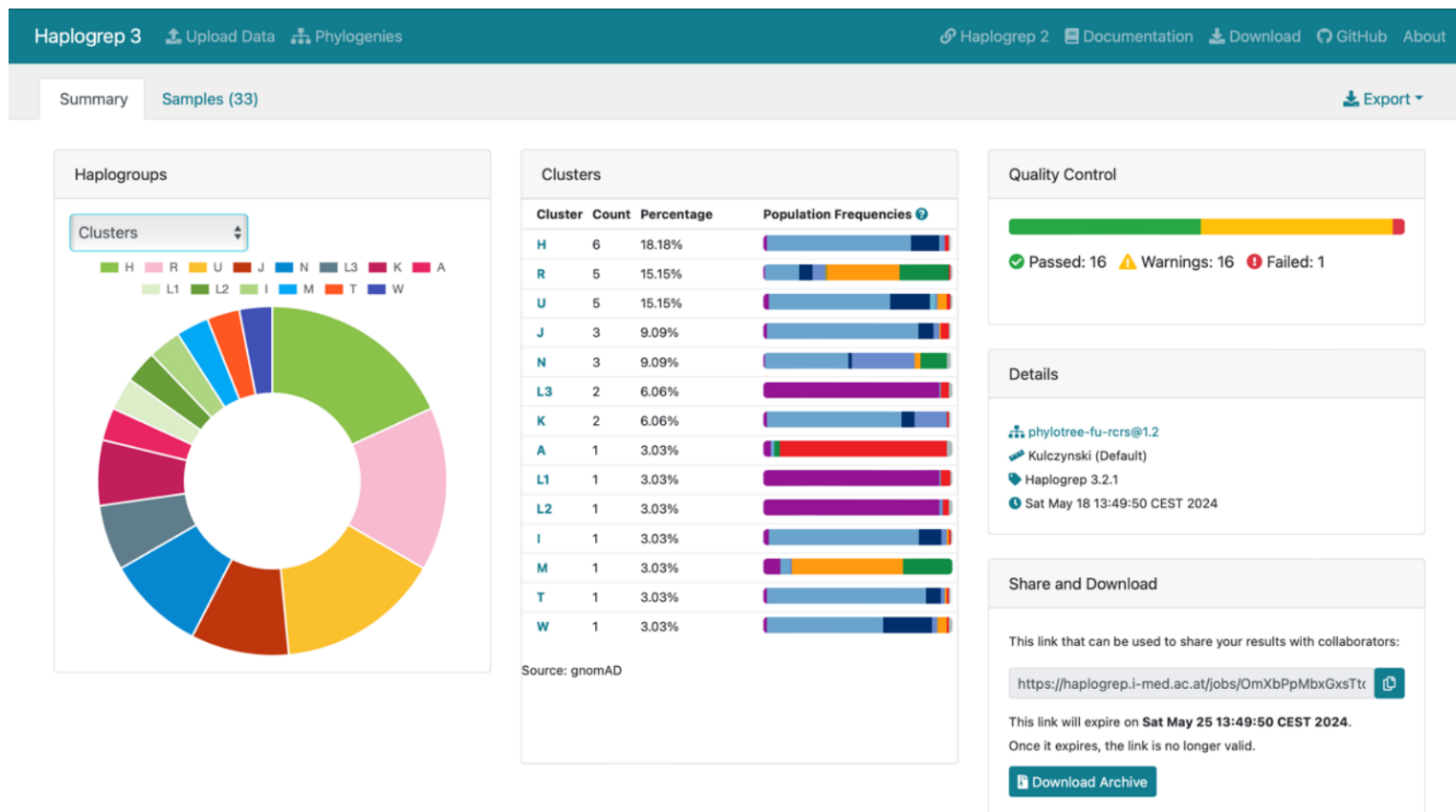


Figure 3.10. HaploGrep results screen of a batch entry for Emirati samples, showing haplogroups, clusters, and quality control with Haplogroup H as the most prevalent (Parson and Bandelt, 2007). The Figure also highlights warnings and failed samples. Warnings may indicate issues such as undetermined variants, global private mutations unknown to Phylotree, or local private mutations associated with other haplogroups. Failed samples are marked in red and generally occur when the detected haplogroup quality is low (quality $\leq 80\%$) or when the sample is missing more than two expected polymorphisms.

Finally, all the information gathered from both tools was structured for the haplogroups of the three ethnicities. To ascertain the haplogroup of the samples, a procedure that entailed the identified mtDNA variants to be cross-referenced with a reference database was used using HaploGrep (Weissensteiner et al., 2016). This encompassed all established mtDNA haplogroups and moreover, provided an online automated approach for haplogroup identification by navigating the underlying PhyloTree (van Oven and Kayser., 2009). This comparison involved aligning the identified mtDNA variants with those cataloged in the classification system, and if any of these variants matched haplogroup-defining variants, the haplogroup was deduced. However, the accuracy of this method for assignment may be affected by various factors, including rare variants absent from the reference database, and the quality of sequence data. To address these potential challenges, assigned haplogroups were submitted to the EMPOP database for verification. While both EMPOP and HaploGrep used similar datasets, EMPOP provided real-time updates and instant comparison with the most contemporary data available. This was done to ensure more up-to-date verification of the detected variants compared to HaploGrep, which was not updated as frequently (Parson and Bandelt., 2007). Haplogroups were confirmed using EMPOP, and the reported results were based on this final step of data analysis, shown in Table 3.3.

Table 3.3. Mitochondrial DNA control region polymorphic variants in the three ethnic groups

No.	Sample ID	Ethnicities	Haplogroups	Haplogroup Frequencies	Haplotype Frequencies	PM
1	UAE_01	Emirati	H13a2c	0.0303	0.03030303	0.00091827
2	UAE_02	Emirati	N1b1	0.0606	0.06060606	0.00367309
3	UAE_03	Emirati	N1b1	0.0606	0.06060606	0.00367309
4	UAE_04	Emirati	R30b2	0.0303	0.03030303	0.00091827
5	UAE_05	Emirati	K1a*1	0.0303	0.03030303	0.00091827
6	UAE_06	Emirati	R0a1a	0.0606	0.03030303	0.00091827
7	UAE_07	Emirati	T1a1'3	0.0303	0.03030303	0.00091827
8	UAE_08	Emirati	H3c2	0.0303	0.03030303	0.00091827
9	UAE_09	Emirati	L3d1a1a	0.0606	0.06060606	0.00367309
10	UAE_10	Emirati	L1c3b1a	0.0303	0.03030303	0.00091827
11	UAE_11	Emirati	L3d1a1a	0.0606	0.06060606	0.00367309
12	UAE_12	Emirati	H1+16189	0.0303	0.03030303	0.00091827
13	UAE_13	Emirati	K2a	0.0606	0.03030303	0.00091827
14	UAE_14	Emirati	H6	0.0303	0.03030303	0.00091827
15	UAE_15	Emirati	L2a1b1a	0.0303	0.03030303	0.00091827
16	UAE_16	Emirati	U3	0.0303	0.03030303	0.00091827
17	UAE_17	Emirati	K2a	0.0606	0.03030303	0.00091827
18	UAE_18	Emirati	J2a	0.0303	0.03030303	0.00091827
19	UAE_19	Emirati	R12'21	0.0303	0.03030303	0.00091827
20	UAE_20	Emirati	I2'3	0.0303	0.03030303	0.00091827
21	UAE_21	Emirati	H2a1	0.0303	0.03030303	0.00091827
22	UAE_22	Emirati	R0a1a	0.0606	0.03030303	0.00091827
23	UAE_23	Emirati	U9a	0.0303	0.03030303	0.00091827
24	UAE_24	Emirati	M1a1	0.0303	0.03030303	0.00091827
25	UAE_25	Emirati	W6	0.0303	0.03030303	0.00091827
26	UAE_26	Emirati	U3a2a1	0.0303	0.03030303	0.00091827
27	UAE_27	Emirati	H6b	0.0303	0.03030303	0.00091827
28	UAE_28	Emirati	J1d1a	0.0606	0.03030303	0.00091827
29	UAE_29	Emirati	J1d1a	0.0606	0.03030303	0.00091827
30	UAE_30	Emirati	U2e1	0.0303	0.03030303	0.00091827
31	UAE_31	Emirati	U3b3	0.0303	0.03030303	0.00091827

32	UAE_32	Emirati	N1a3a	0.0303	0.03030303	0.00091827
33	UAE_33	Emirati	A	0.0303	0.03030303	0.00091827
34	PAK_01	Pakistani	U2a2	0.033	0.03333333	0.00111111
35	PAK_02	Pakistani	M4b	0.033	0.03333333	0.00111111
36	PAK_03	Pakistani	M5c1	0.033	0.03333333	0.00111111
37	PAK_04	Pakistani	M18a	0.033	0.03333333	0.00111111
38	PAK_05	Pakistani	M30	0.033	0.03333333	0.00111111
39	PAK_06	Pakistani	M6a1a	0.033	0.03333333	0.00111111
40	PAK_07	Pakistani	R30a1c	0.033	0.03333333	0.00111111
41	PAK_08	Pakistani	M5a	0.033	0.03333333	0.00111111
42	PAK_09	Pakistani	R2d	0.067	0.03333333	0.00111111
43	PAK_10	Pakistani	M3d	0.033	0.03333333	0.00111111
44	PAK_11	Pakistani	U2c1a	0.033	0.03333333	0.00111111
45	PAK_12	Pakistani	U7a	0.033	0.03333333	0.00111111
46	PAK_13	Pakistani	HV12b1	0.033	0.03333333	0.00111111
47	PAK_14	Pakistani	U2c1b*1a	0.033	0.03333333	0.00111111
48	PAK_15	Pakistani	R2d	0.067	0.03333333	0.00111111
49	PAK_16	Pakistani	M30+16234	0.033	0.03333333	0.00111111
50	PAK_17	Pakistani	L5a1	0.033	0.03333333	0.00111111
51	PAK_18	Pakistani	M30d1	0.033	0.03333333	0.00111111
52	PAK_19	Pakistani	M45	0.033	0.03333333	0.00111111
53	PAK_20	Pakistani	M3c2	0.033	0.03333333	0.00111111
54	PAK_21	Pakistani	M5a2	0.033	0.03333333	0.00111111
55	PAK_22	Pakistani	U1a1c1	0.033	0.03333333	0.00111111
56	PAK_23	Pakistani	U5b2	0.033	0.03333333	0.00111111
57	PAK_24	Pakistani	W3a1	0.033	0.03333333	0.00111111
58	PAK_25	Pakistani	U2b2	0.033	0.03333333	0.00111111
59	PAK_26	Pakistani	M33a2a	0.033	0.03333333	0.00111111
60	PAK_27	Pakistani	F1c1a2	0.033	0.03333333	0.00111111
61	PAK_28	Pakistani	M3	0.033	0.03333333	0.00111111
62	PAK_29	Pakistani	M37e2	0.033	0.03333333	0.00111111
63	PAK_30	Pakistani	M5a2a1a1	0.033	0.03333333	0.00111111
64	IND_01	Indian	M	0.033	0.03333333	0.00111111
65	IND_02	Indian	R8b1	0.033	0.03333333	0.00111111

66	IND_03	Indian	N1a1b1	0.067	0.03333333	0.00111111
67	IND_04	Indian	HV2a	0.033	0.03333333	0.00111111
68	IND_05	Indian	R30b	0.033	0.03333333	0.00111111
69	IND_06	Indian	U7	0.033	0.03333333	0.00111111
70	IND_07	Indian	U2a1a	0.067	0.03333333	0.00111111
71	IND_08	Indian	U2c'd	0.033	0.03333333	0.00111111
72	IND_09	Indian	U9a1	0.033	0.03333333	0.00111111
73	IND_10	Indian	M3	0.033	0.03333333	0.00111111
74	IND_11	Indian	M5	0.067	0.03333333	0.00111111
75	IND_12	Indian	U7a	0.033	0.03333333	0.00111111
76	IND_13	Indian	A17*	0.033	0.03333333	0.00111111
77	IND_14	Indian	M2a'b	0.033	0.03333333	0.00111111
78	IND_15	Indian	M5	0.067	0.03333333	0.00111111
79	IND_16	Indian	U2a	0.033	0.03333333	0.00111111
80	IND_17	Indian	JT	0.033	0.03333333	0.00111111
81	IND_18	Indian	M36a	0.033	0.03333333	0.00111111
82	IND_19	Indian	U2a1a	0.067	0.03333333	0.00111111
83	IND_20	Indian	M6b	0.033	0.03333333	0.00111111
84	IND_21	Indian	M39b1	0.033	0.03333333	0.00111111
85	IND_22	Indian	H3b6	0.033	0.03333333	0.00111111
86	IND_23	Indian	W3a1	0.033	0.03333333	0.00111111
87	IND_24	Indian	M2a1	0.067	0.03333333	0.00111111
88	IND_25	Indian	M44a	0.033	0.03333333	0.00111111
89	IND_26	Indian	M65a+16311	0.033	0.03333333	0.00111111
90	IND_27	Indian	M2a1	0.067	0.03333333	0.00111111
91	IND_28	Indian	N1a1b1	0.067	0.03333333	0.00111111
92	IND_29	Indian	M2a1a	0.033	0.03333333	0.00111111
93	IND_30	Indian	M30c1	0.033	0.03333333	0.00111111

3.6. Discussion

In forensic genetics, the accuracy and reliability of mtDNA haplotypes are of paramount importance. In this study, the number of haplotypes was calculated for each set.

Haplotypes are a set of genetic variations that are inherited together from one parent to the offspring. The number of haplotypes, denoted by H , is an important parameter used to measure genetic diversity in a population. The results were summarized in Table 3.3. In Emiratis set, 23 unique haplotypes were recorded while 10 reoccured (non-unique). For the Pakistanis and indians set, all samples were unique haplotypes.

Next, haplotype diversity (H_d) was calculated. Haplotype diversity (H_d) is a measure of the genetic diversity within a population, based on the distribution of haplotypes (i.e., specific combinations of genetic variants) among individuals (Nei and Tajima, 1981). It considers both the sum of haplotypes present and their frequencies in the population.

Following is the formula for determining haplotype diversity:

$$H_d = \left[\frac{n}{n-1} \right] \left[1 - \sum_{i=1}^k (P_i)^2 \right]$$

H_d is the gene diversity (heterozygosity), n is the total number of individuals in the population, P_i is the frequency of the i -th allele in the population. $\sum_{i=1}^k (P_i)^2$ is the sum of the squared allele frequencies, where k is the total number of different alleles. The formula first calculates the proportion of all possible pairs of haplotypes that differ from each other, which is expressed as $\left[\frac{n}{n-1} \right]$. It then subtracts from this value the sum of the squared haplotype frequencies, multiplied by the sample size. The formula also known as heterozygosity, which measures the probability that two randomly selected alleles from a population will be different. This formula accounts for the bias in small sample sizes by including the term $\left[\frac{n}{n-1} \right]$. The term $\sum_{i=1}^k (P_i)^2$ represents the sum of the squared

frequencies of each allele, which is then subtracted from 1 to provide the probability of selecting two different alleles.

Probability of Matching (PM) and Probability of Discrimination (PD) were calculated to assess the effectiveness of mitochondrial DNA haplogroups in distinguishing individuals within the population. The PM is calculated by considering the likelihood that two randomly chosen individuals from the population will have identical genetic profiles for a given marker or set of markers. The equation used is $PM = \sum_{i=1}^n p_i^2$, where p_i is the frequency of the i -th haplotype in the population. The summation runs over all n haplotypes. The PD is calculated based on the PM and represents the likelihood that two randomly selected individuals from the population will have different genetic profiles. The equation for PD is $PD = 1 - PM$. The PM values, as detailed in Table 3.3, represented the likelihood that two randomly selected individuals will share the same mtDNA haplotype. These calculations are crucial for understanding the genetic diversity captured by the haplogroups and for evaluating the utility of mtDNA in forensic. Expressed in percentages, the PD values for Emiratis, Pakistanis, Indians were 96%, 97%, 97% respectively. This high PD value underscores the robust discriminatory power of mitochondrial haplotypes indicates the probability that two randomly selected individuals will have different mtDNA haplogroups. Such discrimination is crucial for applications where individual genetic differentiation is required, such as in forensic case work.

In addition to PD, the Probability of Identity (PI) was calculated to assess the likelihood that two randomly selected individuals from the population would have identical

mitochondrial DNA haplogroups. The PI is calculated using the equation $PI = 1 - PD$. The calculated values are in Table 3.4. Those values when converted to percentages for Emiratis, Pakistanis, Indians were 4%, 3%, and 3%, respectively. This percentage resembled the chance that two randomly selected individuals from the studied population will have the same mtDNA haplotype, based on the data analysed. This highlights the potential for mtDNA haplogroup overlap in a given population.

Table 3.4. Number of Haplotypes (Ht), number of Haplogroups (Hg), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for the three populations sets.

Set	Ht	Hg	Hd	PD	PI
Emiratis (n=33)	31	27	0.9715	0.96235078	0.03764922
Pakistanis (n=30)	30	29	0.9586	0.96666667	0.03333333
Indians (n=30)	30	26	0.9793	0.96666667	0.03333333

All polymorphisms were imported into EMPOP, an online software that uses PhyloTree Build, and used for assignment of mitochondrial haplogroups (Parson et al. 2014; Parson and Dür 2007; Zimmermann et al 2011; Huber et al 2018). This tool also allowed for quality check of the alignment of sequence as per the PolyTree analysis approach based on the evolution of mtDNA.

All variants were uploaded and searched against EMPOP database to check for alignment and identify macro-haplogroups of the sample populations. Emirati samples was the most diverse set. The set showed haplogroups H, R, U, J, N, L3, K, A, L1, L2, I, M, T, and W (figure 3.11). This also was typical results for Arabs populations in the regions as reported previously (Alshamali et al. 2008). For Pakistani samples, haplogroups were identified to

be M, U, R, H, F, W, and L5 (figure 3.12). The result is typical to South Asian populations as it has been found previously (Quintana-Murci et al. 2004). Indian samples represented haplogroups M, U, R, N, A, W, H, J and HV (figure 3.13). The observed haplogroups in the Pakistanis' set and Indians' set shared very high similarity, which was expected and seen in previous studies, which serves as a validation to the reported findings.

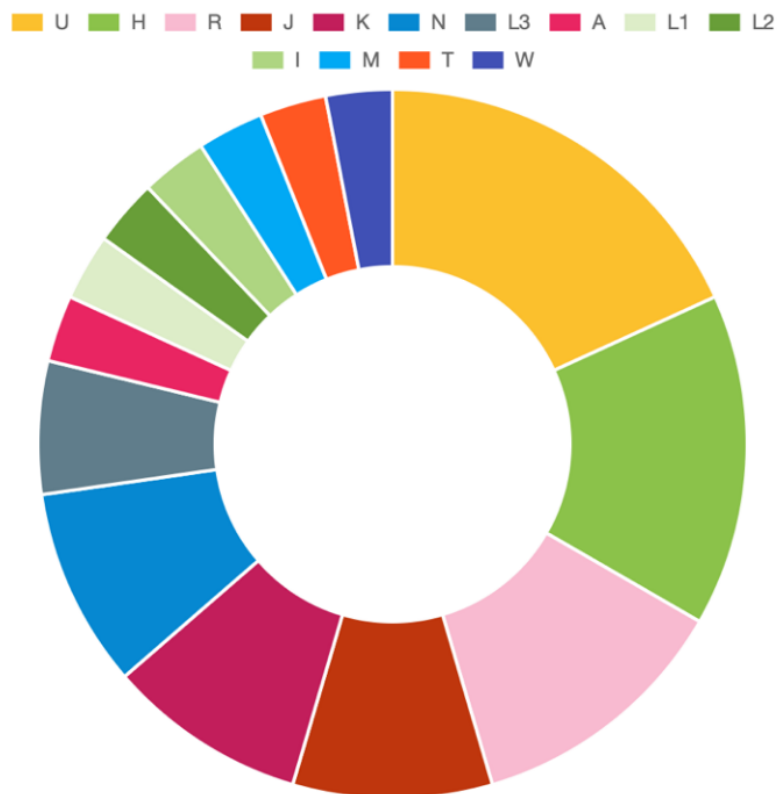


Figure 3.11. Pie chart showing haplogroup clusters of the Emiratis set

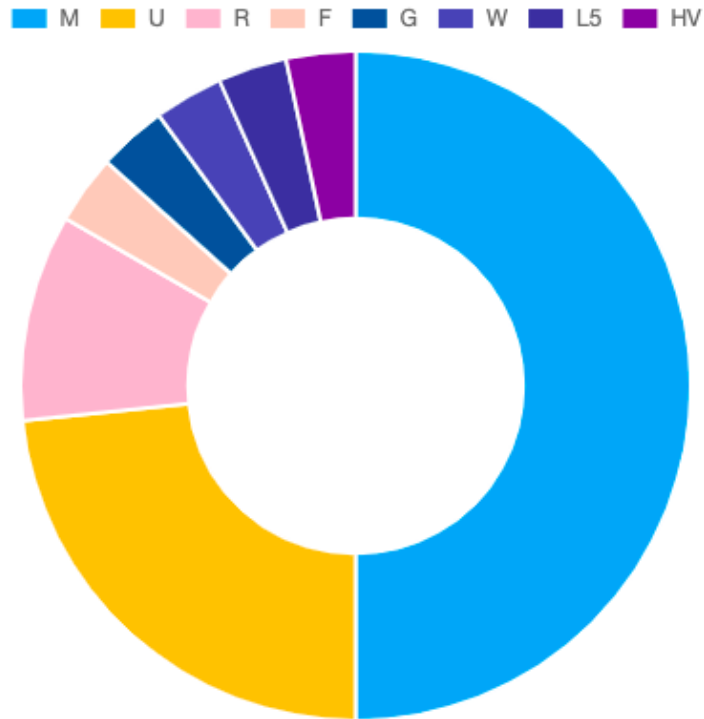


Figure 3.12. Pie chart showing haplogroup clusters of the Pakistanis set

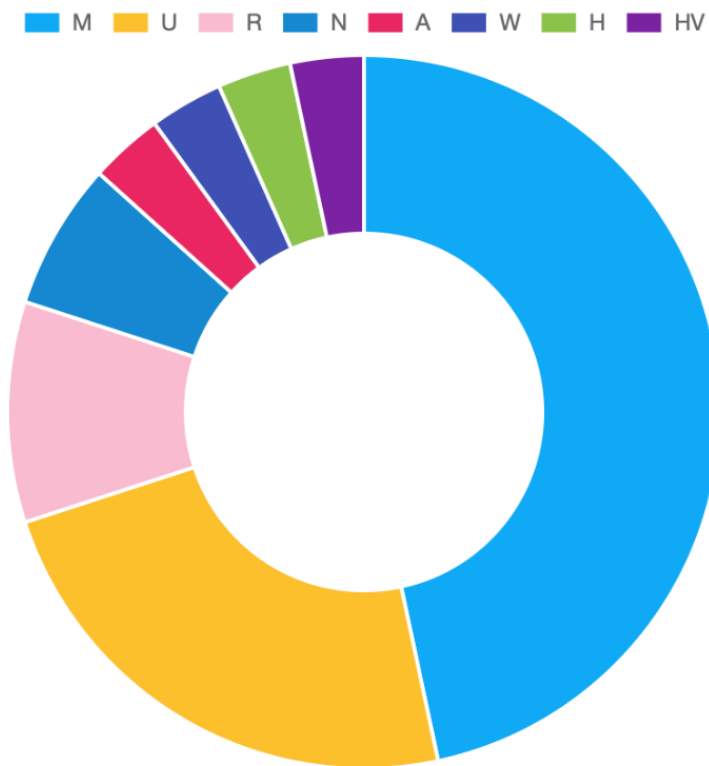


Figure 3.13. Pie chart showing haplogroup clusters of the Indians set

3.7. Conclusion

The Sanger sequencing analysis conducted in this study provided a foundational dataset for evaluating concordance with Massively Parallel Sequencing (MPS) data in subsequent analyses. A total of 93 samples were analyzed: 33 from Emiratis, 30 from Indians, and 30 from Pakistanis. Each sample was sequenced to obtain detailed mitochondrial DNA (mtDNA) haplotypes for the control region (CR), facilitating the identification of unique single nucleotide polymorphisms (SNPs), haplotype assignments, and haplogroup classifications.

A limitation of the Sanger sequencing data was that much of the sequence was obtained on only one strand, which is not ideal for comprehensive coverage. However, the quality was deemed sufficient for the purpose of cross-validating MPS results (Tagliabracci & Turchi, 2020). This approach aligns with standard forensic practices, where Sanger sequencing of the mtDNA control region has been traditionally employed for analyzing degraded samples remains (Santos et al., 2024).

While the current dataset provided valuable insights, it is important to note that the sample size was relatively limited, which restricts the extent of population genetics findings. Nevertheless, this initial analysis has illustrated the genetic diversity within and between the Emirati, Indian, and Pakistani populations. These preliminary results have laid the groundwork for more extensive analysis using MPS, a technology that has transformed genetic data generation and human identification using mtDNA (Taylor et al., 2022).

Chapter 4

4. Massive Parallel Sequencing of Whole mtDNA Genome

4.1.Introduction

The advent of massive parallel sequencing (MPS) technologies has significantly advanced genetics research, including applications in mitochondrial DNA (mtDNA) analysis for forensic purposes (Just and Irwin, 2019). While Sanger sequencing remains widely used and essential in many forensic laboratories, MPS holds promise for generating high-quality mitogenome haplotypes. These haplotypes can provide valuable evidence for forensic applications, such as identifying human remains, solving cold cases, and analyzing degraded samples. For example, the study by Taylor et al. (2020) demonstrates the potential of MPS by generating a comprehensive reference dataset of 1327 ‘platinum-quality’ mtDNA haplotypes from various U.S. populations. This reference dataset enhances the capacity of forensic community to conduct accurate and reliable mtDNA analyses, supporting the ongoing development of forensic methodologies.

Given the genetic diversity and unique population structure of the United Arab Emirates (UAE), implementing a comprehensive mitochondrial DNA (mtDNA) population study in the region is both feasible and highly beneficial. The rich tapestry of ethnicities in the UAE and significant proportion of expatriate residents necessitate a detailed mtDNA reference database for accurate forensic applications. By adapting the methodology from Taylor et al. (2020) to generate high-quality mitogenome haplotypes, this study aims to capture

the genetic diversity within the UAE. Such a localized reference database would significantly enhance forensic investigation capabilities, enabling precise identification and comparison of genetic material. Furthermore, it would support robust interpretation of mtDNA evidence, thereby strengthening the forensic framework in the UAE and improving the accuracy and reliability of forensic analyses.

The advancement of MPS technologies has increased the reproducibility of whole mitochondrial genome sequencing approaches while decreasing human error. The technology used in this study was ion torrent. MPS was tested to check its reproducibility in forensic workflow to test low-level variations, which can be challenging to be detected in Sanger sequencing. Also, MPS allows the detection of heteroplasmy at low levels. In summary, in the area of forensic applications, MPS technologies are capable of providing the whole mtDNA genome in a robust, reliable and efficient way (Ballard et al., 2020).

The SWGDAM published updated validation guidelines for mtDNA assessment and comparisons of mtDNA evidence samples and known samples used in forensic analyses. In turn, this specified which evidence should be included and excluded in mtDNA forensic analysis (Scientific Working Group on DNA Analysis Methods, 2013).

4.2. Chapter Aim

This chapter was designed to process extensive sets of samples and execute Massively Parallel Sequencing (MPS) utilizing Ion Torrent technologies, to generate high-quality genetic data that is suitable for detailed forensic data analysis. This includes rigorous

quality control measures through the use of Standard Reference Materials (SRM), Control DNA, and GEDNAP proficiency testing samples to validate and maintain the integrity and accuracy of the sequencing results. In addition, the samples which were sequenced in Chapter 3 were also used to assess the quality of data produced.

4.3. Chapter Objectives

- To quantify mitochondrial DNA from each sample to ensure optimal input for library preparation.
- To utilize the Precision ID Whole mtDNA Genome Panel for library preparation, tailored to the requirements of MPS.
- To load prepared libraries onto sequencing chips using Ion Torrent technologies (TSS and Ion Chef™) with adherence to manufacturer guidelines.
- To perform MPS to prepared Chips including setup and monitoring of the sequencing run to ensure error-free operation and optimal data generation.
- To conduct comprehensive quality checks of the sequencing data to evaluate the read quality, coverage uniformity, and error rates. This step is crucial to validate the usability of the sequencing data for further analysis.
- To utilize standard metrics and software tools for assessing data integrity and quality, ensuring that only high-quality data is used for downstream analyses.
- To analyze Standard Reference Materials (SRMs), Controls, and GEDNAP Samples to validate the sequencing procedures using MPS system. This analysis helps in benchmarking the system's performance against known standards.

- To evaluate GEDNAP proficiency testing samples, which are part of an external quality assessment scheme, to ensure the laboratory's compliance with international forensic standards. This is essential for forensic application of the MPS data.

4.4. Methods

Detailed methodology is outlined in chapter 2.

4.5. Massive Parallel Sequencing Emirati Reference Data

A total of 510 whole mtGenome Emirati population were obtained, extracted and sequenced. Those samples were more representative of the seven different Emirates in the UAE from previous studies. The samples also included the previous 33 samples that had been processed using Sanger sequencing approach. Initially, the MPS started with those 33 samples to be the first set analysed. Per run, the instruments load 2 chips with a total of 32 samples per ion 530 chip. The Torrent Suite™ Software (TSS) was set using the default parameters. Table 4.1 summarized the parameters obtained from the TSS. The raw data were obtained in different forms depending on the analysis software used. For Converge™ Software .bef files were used, while for Integrative Genomics Viewer (IGV) .bam and .bai format files were also available. Variant calling analysis with Converge™ Software was done using the revised Cambridge Reference Sequence (NC_012920.1). Through Converge™ Software, using a built in algorithm, the sequences of variant status were reported according to EMPOP database. As such, it flagged each

variant detected either confirmed, likely, unclear, possible or false variation. All samples were anonymized and searched in EMPOP for haplotype alignment and haplogroup assignment confirmation.

Table 4.1. Torrent Suite™ Software (TSS) default analysis parameters obtained from the MITO Genotyper

Options	Configurations	
Coverage	Min total read coverage per position	20
	Min variant coverage to call	20
	Coverage threshold to mark region	20
	Min coverage percent compared to the median of the amplicon	5
General	Show input BAM in IGV	✓
	Remove contaminant reads	⊗
	Include variants flagged as NUMT or DEGRAD	✓
Reporting	Threshold for recording detailed coverage stats	2
Thresholds	Threshold for confirmed call	96
	Threshold for PHP call	10
	Threshold for insertion call	20
	Threshold for deletion call	30

4.6. Whole mtDNA Variant Calling Sequencing Analysis

Recent advancements in next-generation sequencing (NGS) have significantly enhanced the resolution at which mtDNA could be studied, necessitating precise variant analysis

tools. Lee et al. (2023) study evaluated the efficacy of two variant calling programs, Converge Software (CS) and Torrent Variant Caller (TVC), both of which were integral to the NGS workflows for mtDNA. This comparative study of CS and TVC provided essential insights into the capabilities and limitations of current variant calling software for mitochondrial DNA, highlighting the critical considerations necessary for accurate genetic analysis. The analysis revealed around 2,300 mtDNA variants with a high consistency rate of 90% between the two software. CS showed a slight advantage in forensic applications due to its tailored features for mtDNA analysis.

Given the high stakes of forensic analysis, understanding the performance, advantages, and limitations of these tools was critical. The choice of analysis software could significantly impact the outcomes of genetic research. As such, ongoing advancements in NGS technology and bioinformatics tools would be vital in harnessing the full potential of mtDNA studies in forensic and medical sciences.

The variant calling in this study was performed using both CS and TVC, with a focus on overcoming all limitations from one software to another, as well as the effectiveness of the software in detecting and accurately calling mtDNA variants. The necessity to choose the variant analysis tools wisely for forensic application was based on the need to cover all possible challenges in the mtDNA analysis. CS for example, outperformed TVC in handling forensic-specific mtDNA complexities, such as the co-amplification of NUMTs and nuanced variant notations.

4.6.1. DNA sequences Reconstruction

DNA sequences were reconstructed by pooling all barcoded libraries to respective samples and trimming adapter sequences 20 bases from the 3' and 5' end, using Torrent Suite™ software (Applied Biosystems, CA, USA). Sequences were formatted to the human mitochondrial genome by alignment to the revised Cambridge Reference Sequence (rCRS). Sequence variants, SNPs, insertions and deletions (INDELs), were reported using the Ion S5™ System: Torrent Variant Caller V5.0 (Applied Biosystems, CA, USA) plugin as variant caller files. Specifically, the alignment was done to an rCRS + 80 reference genome, which accounted for the Precision ID mtDNA Whole Genome Panel's overlapping design.

4.6.2. Alignment and Reference Genome

A read depth threshold of 20 reads was employed, ensuring sufficient coverage for reliable variant calling, as per the parameters determined by Thermo Fisher Scientific to provide accurate data. Additionally, a threshold of 20 was set for detecting heteroplasmy, meaning that at least 20 reads supporting a minor variant were required to identify it as a true heteroplasmic site, which served as the point heteroplasmy threshold for reference work (Table 4.1). In casework, this threshold was decreased to 10, to override possible contaminations.

In conjunction with the overlapping design of the panel, all polymorphisms were imported into EMPOP, an online software that used PhyloTree Build, and used for assignment of mitochondrial haplogroups (Parson et al. 2014; Parson and Dür 2007; Zimmermann et al

2011; Huber et al 2018). This tool also enabled a quality check of the alignment of sequence as per the PolyTree analysis approach which was based on the evolution of mtDNA.

4.7. Quality assessment and Output Files

The mtDNA sequencing data was processed using the Torrent Suite Software (TSS), with the HID Genotyper Plugin being employed for detailed variant analysis and coverage evaluation. Upon completion of sequencing run, TSS automatically generated a series of output files, which were then compressed into a ZIP file. The extracted ZIP file contained a specific coverage graph file found within the different files, typically labelled as "coverage_graph". This graph is normally a circular plot of the coverage graph, illustrating both the sequencing coverage depth and the variant distribution across the mitochondrial genome (Figure 4.1). A circular format was adopted by this graph, which was particularly well-suited for representing the circular structure of mtDNA, facilitating a continuous and intuitive visualization of the genetic data. The depth of coverage at different points along the mitochondrial genome was indicated by radial peaks, which were crucial for assessing the reliability of the genetic data. Positions of genetic variants are marked by coloured squares—green, yellow, and red—categorized by their confirmation status as confirmed/likely, unclear/possible, and false, respectively. Figure 4.4. summarized the colour codes used.

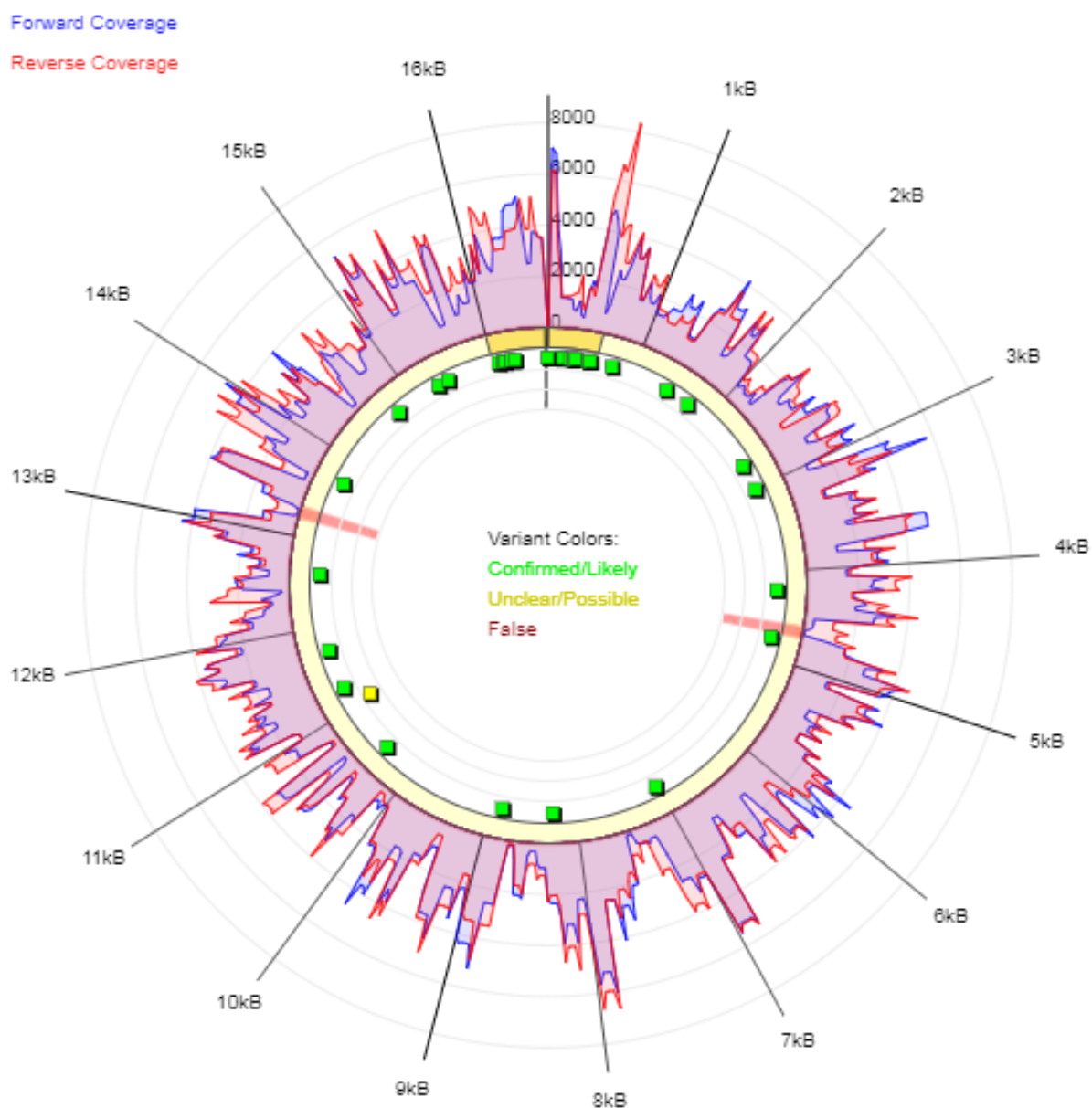


Figure 4.1. Circular plot of the mtGenome sequencing coverage for the forward and reverse pools summary of extracted reference blood samples developed by TSS.

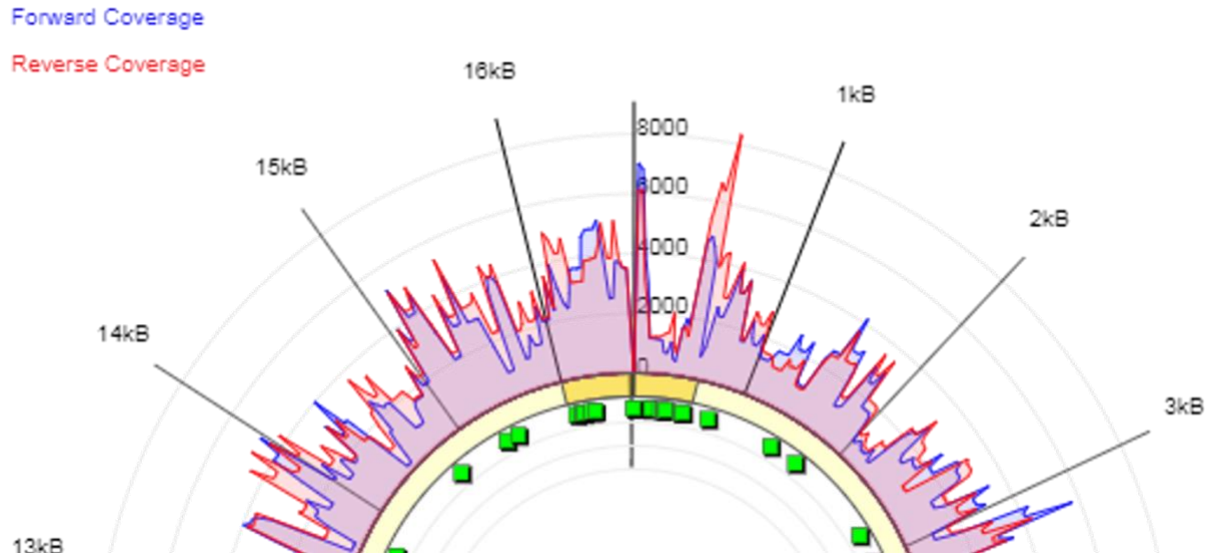


Figure 4.2. A closer image of the mtGenome sequencing coverage chart, the green colored squares flags confirmed variants in the sequence. The blue color indicates forward coverage in sequencing data. It shows how many reads from the sequencing process cover each position in the forward direction of the DNA strand. Conversely, the red color represents reverse coverage. The kilobase (kB) marks indicates the positions along the mtDNA that is around 16.5 kB in length, which help in pinpointing specific locations on the mtDNA. The numeral intervals (2000, 4000, 6000, 8000) are marks to provide visual assess of the coverage reads measurement.

The file named `variants_colored.xlsx` an excel spreadsheet specifically was designed to list and to categorize genetic variants from sequencing data. It was also derived from the analysis performed using the Torrent Suite Server (TSS) with the HID Genotyper Plugin. The excel file was colour-coded and it was utilized for visualizing and managing variant data derived from mtDNA sequencing. These files included multiple sheets (total of 9) starting with the “Variant”. This sheet included all the variant data with a crucial header information at the top of the spreadsheet (Figure 4.3).

Position	Ref	Sampl	Varia	Var Freq	Major Freq	Type	Read Cover	Read Cover	Allele Cov	Allele	Allele	G%	A%	T%	C%	N%	ins%	del%	Polymorph	Control Ref	State	Frequency	Artefact	Var Strand	Read Strand	EMPOP
73	A	G	G	100	100	SNP	1425	1159	2584	1425	1159	100	0	0	0	0	3.6	0	73G	73G	confirmed	100	True variant	0.5	0.6	unchecked
152	T	C	C	98.4	98.4	SNP	436	905	1319	435	884	0	0.1	0.1	98.4	0	1	1.5	152C	152C	confirmed	98.4	True variant	0.5	0.7	unchecked
263	A	G	G	100	100	SNP	55	36	91	55	36	100	0	0	0	0	0	0	263G	263G	confirmed	100	True variant	0.5	0.6	unchecked
291	A	A	-	38.2	58.8	DEL	31	37	26	11	15	0	58.8	2.9	0	0	0	38.2	291a	291a	unclear	38.2	Length Het	0.5	0.5	unexpected
309	C	C	-	32.4	67.6	DEL	30	41	23	14	9	0	0	0	67.6	0	0	32.4	309c	309c	likely	32.4	Length Het	0.7	0.6	unexpected
315	G	+	C	82.6	86.8	INS	29	40	57	20	37	0	0	0	0	0	85.5	0	315.1C	315.1C	likely	82.6	True variant	0.6	0.6	confirmed
750	A	G	G	95.7	95.7	SNP	453	321	741	430	311	95.7	2.6	0	0	0	0.9	1.7	750G		likely	95.7	True variant	0.5	0.6	unchecked
1438	A	G	G	99.4	99.4	SNP	243	93	334	241	93	99.4	0.6	0	0	0	0	0	1438G		confirmed	99.4	True variant	0.5	0.7	unchecked
1598	G	A	A	98.7	98.7	SNP	953	557	1490	938	552	0.5	98.7	0	0	0	0.1	0.9	1598A		confirmed	98.7	True variant	0.5	0.6	unchecked
1703	C	T	T	99.2	99.2	SNP	32	100	131	32	99	0	0.8	99.2	0	0	0	0	1703T		confirmed	99.2	True variant	0.5	0.8	unchecked
1719	G	A	A	98.5	98.5	SNP	32	100	130	32	98	1.5	98.5	0	0	0	0	0	1719A		confirmed	98.5	True variant	0.5	0.8	unchecked
2232	A	A	-	34.7	65.3	DEL	6	95	35	1	34	0	65.3	0	0	0	0	34.7	2232a		unlikely	34.7	Length Het	0.7	0.9	unknown ir
2639	C	T	T	98.8	98.8	SNP	450	549	987	444	543	0	0.8	98.8	0.4	0	0	0	2639T		confirmed	98.8	True variant	0.5	0.5	unchecked
2706	A	G	G	99.9	99.9	SNP	615	232	846	614	232	99.9	0.1	0	0	0	0.4	0	2706G		confirmed	99.9	True variant	0.5	0.7	unchecked
3921	C	A	A	99.7	99.7	SNP	1186	684	1864	1183	681	0.2	99.7	0.1	0	0	1.9	0.1	3921A		confirmed	99.7	True variant	0.5	0.6	confirmed
4611	A	-	-	50.8	50.8	DEL	93	615	360	8	352	0.3	48.7	0.1	0	0	1.6	50.8	4611del		unclear	50.8	Length Het	0.9	0.9	unknown ir
4769	A	G	G	99.9	99.9	SNP	349	550	898	348	550	99.9	0	0	0	0	0.1	0.1	4769G		confirmed	99.9	True variant	0.5	0.6	unchecked
4874	A	A	-	38.3	60	DEL	94	81	67	60	7	1.1	60	0	0.6	0	0	38.3	4874a		unclear	38.3	Length Het	0.9	0.5	unknown ir
4904	C	T	T	100	100	SNP	92	81	173	92	81	0	0	100	0	0	0	0	4904T		confirmed	100	True variant	0.5	0.5	unchecked
4960	C	T	T	99.8	99.8	SNP	2184	1262	3439	2179	1260	0	0	99.8	0.1	0	0.2	0	4960T		confirmed	99.8	True variant	0.5	0.6	unchecked
5237	G	A	A	100	100	SNP	188	57	245	188	57	0	100	0	0	0	1.2	0	5237A		confirmed	100	True variant	0.5	0.8	unchecked
5471	G	A	A	99.6	99.6	SNP	208	247	453	208	245	0	99.6	0	0	0	0.2	0.4	5471A		confirmed	99.6	True variant	0.5	0.5	unchecked
6272	A	G	G	98.8	98.8	SNP	1058	971	2005	1038	967	98.8	0.5	0	0.6	0	0.9	0	6272G		confirmed	98.8	True variant	0.5	0.5	unchecked
7028	C	T	T	99.7	99.7	SNP	709	563	1268	706	562	0	0	99.7	0.3	0	0.1	0	7028T		confirmed	99.7	True variant	0.5	0.6	unchecked
8251	G	A	A	99.4	99.4	SNP	388	1572	1948	385	1563	0.2	99.4	0	0.2	0	0.2	0.3	8251A		confirmed	99.4	True variant	0.5	0.8	unchecked
8472	C	T	T	98.2	98.2	SNP	0	56	55	0	55	0	0	98.2	1.8	0	0	0	8472T		likely	98.2	True variant	0.5	1	unchecked
8836	A	G	G	99.5	99.5	SNP	97	449	543	97	446	99.5	0	0	0.5	0	0.9	0	8836G		confirmed	99.5	True variant	0.5	0.8	unchecked
8860	A	G	G	100	100	SNP	96	451	547	96	451	100	0	0	0	0	0.5	0	8860G		confirmed	100	True variant	0.5	0.8	unchecked
9335	C	T	T	98.6	98.6	SNP	328	229	549	324	225	0	0	98.6	1.3	0	0.4	0.2	9335T		confirmed	98.6	True variant	0.5	0.6	unchecked
10238	T	C	C	95.9	95.9	SNP	229	706	897	228	669	0	0	4.1	95.9	0	2.6	0	10238C		likely	95.9	True variant	0.5	0.8	unchecked
11038	A	A	-	38.4	61.1	DEL	334	54	149	142	7	0.5	61.1	0	0	0	18.3	38.4	11038a		unclear	38.4	Length Het	0.8	0.9	unknown ir
11362	A	G	G	99	99	SNP	372	448	812	371	441	99	0.2	0	0	0	0.2	0.7	11362G		confirmed	99	True variant	0.5	0.5	unchecked
11719	G	A	A	99.7	99.7	SNP	726	605	1327	724	603	0.3	99.7	0	0	0	0.1	0	11719A		confirmed	99.7	True variant	0.5	0.5	unchecked
12501	G	A	A	99.7	99.7	SNP	150	166	315	149	166	0.3	99.7	0	0	0	0.3	0	12501A		confirmed	99.7	True variant	0.5	0.5	unchecked
12705	C	T	T	100	100	SNP	284	270	554	284	270	0	0	100	0	0	0.4	0	12705T		confirmed	100	True variant	0.5	0.5	unchecked
12822	A	G	G	99.9	99.9	SNP	249	548	796	249	547	99.9	0.1	0	0	0	0.1	0	12822G		confirmed	99.9	True variant	0.5	0.7	unchecked
14766	C	T	T	100	100	SNP	8	91	99	8	91	0	0	100	0	0	0	0	14766T		confirmed	100	True variant	0.5	0.9	unchecked
15326	A	G	G	100	100	SNP	76	43	119	76	43	100	0	0	0	0	0	0	15326G		confirmed	100	True variant	0.5	0.6	unchecked
16093	T	C	C	98.2	98.2	SNP	455	103	548	448	100	0	0.2	1.6	98.2	0	0.4	0	16093C	16093C	confirmed	98.2	True variant	0.5	0.8	unchecked
16145	G	A	A	99.3	99.3	SNP	19	127	145	19	126	0	99.3	0	0	0	0.7	0.7	16145A	16145A	confirmed	99.3	True variant	0.5	0.9	unchecked
16176	C	G	G	100	100	SNP	19	130	149	19	130	100	0	0	0	0	0	0	16176G	16176G	confirmed	100	True variant	0.5	0.9	confirmed
16223	C	T	T	99	99	SNP	142	168	307	139	168	0	0	99	0.3	0	0	0.6	16223T	16223T	confirmed	99	True variant	0.5	0.5	unchecked
16390	G	A	A	98.4	98.4	SNP	611	464	1058	602	456	1.4	98.4	0	0.2	0	0.1	0	16390A	16390A	confirmed	98.4	True variant	0.5	0.6	unchecked
16509	T	C	C	95	95	SNP	435	876	1245	412	833	0	0.2	4.7	95	0	0.5	0.2	16509C	16509C	likely	95	True variant	0.5	0.7	unexpected
16519	T	C	C	99.2	99.2	SNP	433	876	1298	424	874	0.2	0.1	0.2	99.2	0	0	0.3	16519C	16519C	confirmed	99.2	True variant	0.5	0.7	unchecked

Figure 4.3. An example of the 'Variant' sheet in the `variants_colored.xlsx` file.

The excel file included the date of the sequencing run, the BAM file extension was also listed, linking the data directly to the specific sequence data file used for the analysis. The name of the sample was prominently displayed, allowing for easy identification of the dataset. The closest haplogroup to which the sample belongs was noted, providing insights into the genetic ancestry and phylogenetic placement of the sample. Finally, the region of extraction or study was indicated, reflecting the specific area of the mitochondrial genome that was covered and successfully analysed. Figure 4.3 is a snapshot of the 'Variant' sheet. Table 4.2 is a detailed breakdown of the columns typically included in the 'Variant' sheet. Figure 4.4 explained the colour codes used in the sheet.

Table 4.2. Detailed list of the columns of the 'Variant' sheet in the `variants_colored.xlsx` file

Columns	Information
Position	Indicates the specific nucleotide position of the variant within the mitochondrial genome
Reference	Shows the nucleotide present at that position in the reference genome
Sample	Lists the sequence reading of the sample, which includes detailed information about the specific genetic sequence obtained from the sample at each listed position
Variant	Describes the actual variant observed at the position in the sample
Variant Frequency	Indicates the frequency of the variant within the dataset
Major Frequency	Shows the frequency of the most common allele at that position
Type	Classifies the variant, such as SNP (single nucleotide polymorphism), insertion, or deletion
Read Coverage +	Represents the read coverage from the forward strand
Read Coverage –	Represents the read coverage from the reverse strand
Allele Coverage	Total coverage of the variant allele
Allele Coverage +	Coverage of the variant allele on the forward strand
Allele Coverage –	Coverage of the variant allele on the reverse strand
G%	Percentage of guanine bases at that position
A%	Percentage of adenine bases at that position

T%	Percentage of thymine bases at that position
C%	Percentage of cytosine bases at that position
N%	Indicates the percentage of nucleotides at that position that could not be confidently called
Ins%	Percentage of insertions relative to the total reads at that position
Del%	Percentage of deletions relative to the total reads at that position
Polymorphism	Indicates whether the variant is considered a polymorphism
Control Region	Specifies if the variant is within the mitochondrial control region
State	Describes the state of the variant, such as validated, likely, or questionable
Frequency	Provides additional details on the frequency of the variant across different populations or datasets
Artefact	Indicates whether the variant is likely to be an artifact
Var Strand	Shows the strand bias of the variant calls
Read Strand	Indicates the predominant read strand for that position
EMPOP	Links to the EMPPOP database (a mitochondrial DNA database) for comparing population data
Score	A score assigned based on the quality and reliability of the variant data

Color Legend						
State	false	unlikely	unclear	possible	likely	confirmed
Artefact	Artifact	True variant	Point Heteroplasmy	Numt	Length Heteroplasmy	Degradation
EMPOP state	CONFIRMED	NEW	UNCHECKED	UNKNOWN	UNEXPECTED	KNOWN
Frequency	0	10	11	90	91	100
Coverage	0	1	5	10	100	200
Base Coverage Percentage (GATC,	0	80	81	90	91	100
Strand Bias	0.0	0.7	0.8	0.81	0.91	1.0

Figure 4.4. Colour codes used in the 'Variant' sheet in the `variants_colored.xlsx` and the different scores for the listed columns.

The "Summary Info" sheet helped to get an overview of the findings in the sequence. It included legends and explanations to the scores. For example, in the number of (non-indel) heteroplasmic variants, point heteroplasmy (PHP) were expected to be rare, and observed at a maximum of 2 points if observed in a sample. A number larger than 2-3 could indicate contamination (including degradation or NUMTS that were not detected) or a mixture. While degradation alone did not directly increase the number of heteroplasmic variants,

it could complicate the analysis and interpretation of the data, potentially leading to misidentification of variants or false positives.

From the TSS, the Variant Caller Plugin was used after alignment to generate variant calls. This plugin identified differences between the sample sequence and the reference genome, producing variant call format (VCF) output files. The VCF files were then processed through mitoSAVE, a tool that translates these variants into haplotype calls using standard forensic nomenclature. This step was crucial for interpreting the mitochondrial DNA profile in a manner consistent with forensic databases and standards. The performance metrics were evaluated through the assesment of various metrics. The read depth ensured adequate sequencing coverage. Strand balance checked the balance of sequencing reads from both DNA strands. Noise assessed the level of background noise in the sequencing data. These metrics were critical for evaluating the quality and reliability of the sequencing results generated from the MPS process.

The Integrative Genomic Viewer (IGV), another tool utilised from the TSS, was used to visually inspect the aligned Binary Alignment Map (BAM) files. IGV helped in verifying the alignment and identifying any anomalies or errors in the sequencing data.

For phylogenetic analysis, Converge™ software and EMPOP were employed for phylogenetic analysis of the haplotype calls. They helped in classifying the mitochondrial DNA into specific haplogroups and checking for consistency with known mitochondrial DNA databases.

4.8.Mitochondrial genome coverage

Prior to diving into data analysis, it was crucial to review the run summary metrics generated by the sequencing platform (TSS). This was to ensure that the data quality met the recommended parameters for accurate analysis. Upon sequencing completion, the summary report was navigated from the TSS for the desired run. To better understand the interpretation of sequencing data, it was important to consider the following example. Key metrics such as Ion Sphere Particles (ISP) Loading (which was ideally 80% or higher) and Key Signal (ideally 80+) were compared to ensure optimal chip utilization and strong signal strength. The ISP density plot was analysed, (see Figure 4.5). Next, the ISP Summary was verified: for an Ion 530 chip, at least 10 million reads were expected with around 30% usable and ideally 99-100% enrichment. High clonality and a low percentage of low-quality reads and adapter dimers were found to be desirable.

The sequencing run had generated a total of 2.58 giga base pairs as shown in Figure 4.6 indicating the number of filtered and trimmed base pairs documented in the output BAM file. The average key signal for library ISPs was found to be 88, reflecting the average signal intensity for all library Ion Sphere Particles (ISPs) with the library key (TCAG). ISP loading had reached 82%, the percentage of chip wells filled with an ISP. This high loading efficiency was visually supported by the ISP Loading Density (Figure 4.5).

The read histogram was checked, a centered peak around 126 bp was ideal, as shown in Figure 4.6. Once these parameters were confirmed, and alignment and raw accuracy met

expectations (>95% and 99% respectively), this ensured that high-quality data were ready for further exploration.

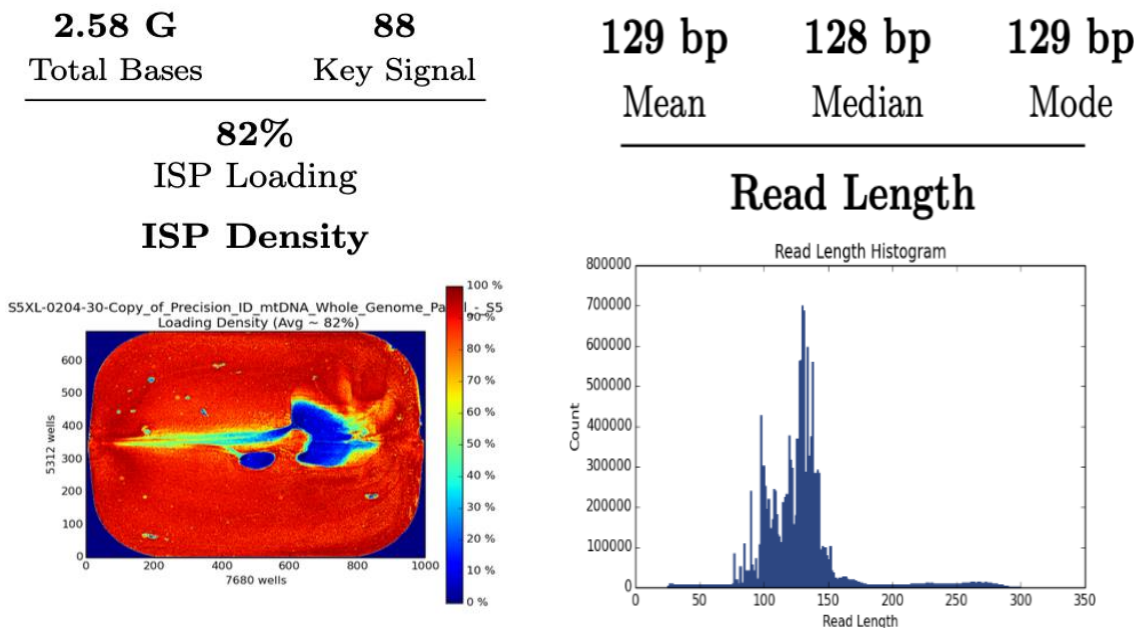


Figure 4.5. ISP density colour gradient, where Red indicates areas of high ISP loading, while blue is the lowest.

Figure 4.6. Read Length Histogram.

A total of 17,991,451 reads were obtained as shown in Figure 4.7, with 59% classified as usable reads. This signified the percentage of library ISPs passing the polyclonal, low-quality, and primer-dimer filters. Furthermore, ISP summary metrics revealed an 82% loading rate, with 18% of wells remaining empty, and a perfect 100% enrichment rate, indicating that all loaded wells contained live ISPs matching the library or test fragment key signals. Clonality was determined to be 70%, demonstrating the proportion of ISPs derived from a single original template while polyclonally is present as 30% and is in the expected range, the final library accounted negligible adapter dimer interference and a 16% discard rate due to low quality. Out of a total of 30,749,394 addressable wells, the

filtering process detected and excluded polyclonal, low-quality, or adaptor dimer-containing samples, resulting in 17,991,451 wells that met the requisite quality standards. This filtering phase was critical for ensuring the reliability and accuracy of downstream analysis. After filtering, approximately 58.52% of the original wells remained, implying that 41.48% were eliminated due to their low quality. This strict quality control method was critical for preserving the integrity of genetic data and ensuring that only high-quality samples could be used in subsequent analysis.

The predicted read length distribution parameters were set to 126 bp as the mean, 138 bp as the median, and 130 bp for the mode by the manufacturer protocol. The observed findings were in line with expected values, with a mean of 129 bp, a median of 129 bp, and a mode of 129 bp. This consistency suggested that the FUPA reagent step, which was designed to digest all primers to a maximum of 200 bp during the library preparation stage was effective. As a result, nearly no read lengths more than 200 bp were detected in the read length histogram as shown in Figure 4.7, demonstrating the precision and reliability of the library preparation procedure. The analysis of individual sample results provided by the Torrent software system followed next.

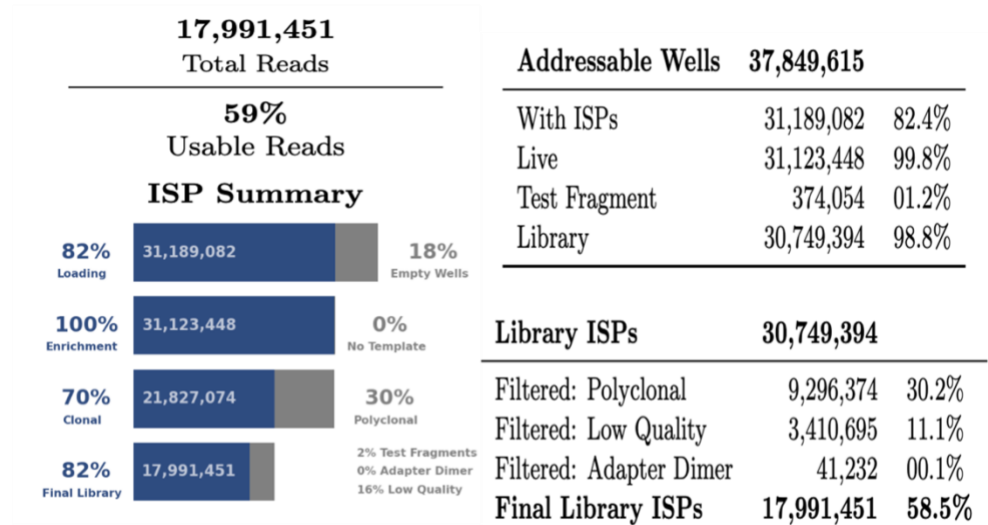


Figure 4.7. Summary of Ion Sphere Particle (ISP) Sequencing Results.

The remaining sequencing chips for the Emirati, Pakistani, and Indian samples yielded similar quality results, as shown in Figures 4.8 to 4.14 and Tables 4.3 to 4.5. These results are consistent with the previously discussed example, confirming the effectiveness of the sequencing protocol across different population samples.

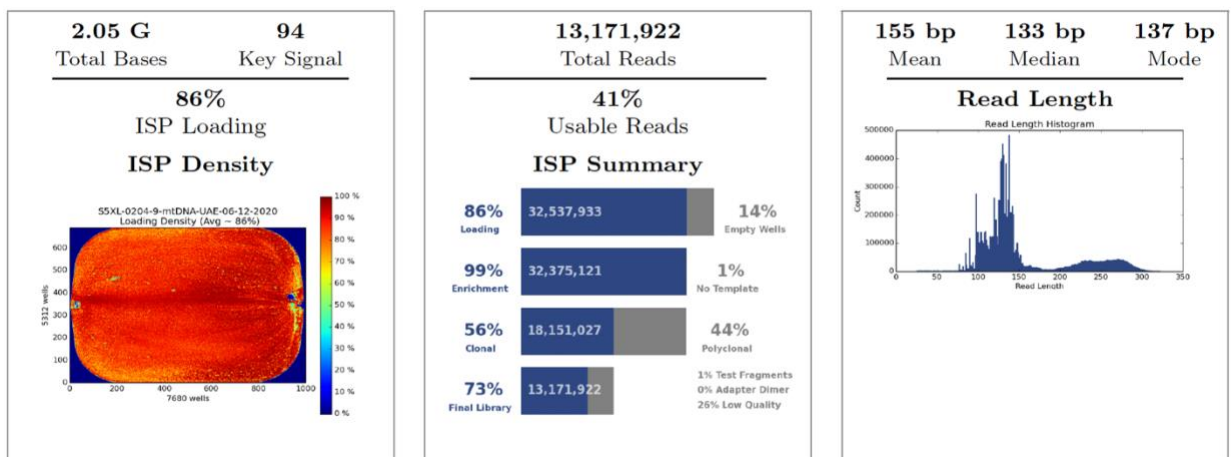


Figure 4.8. Emirati samples Chip 1 TSS quality check

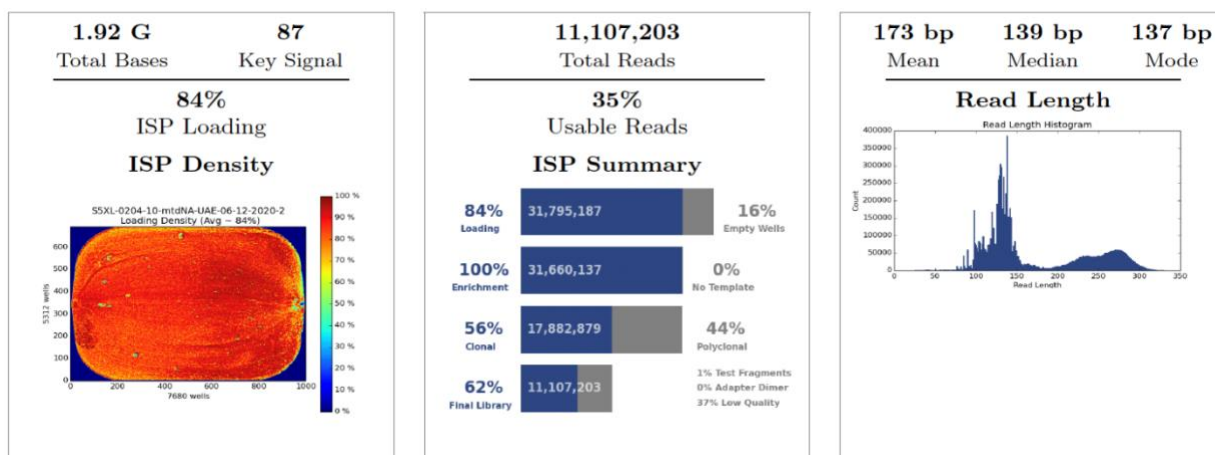


Figure 4.9. Emirati samples Chip 2 TSS quality check

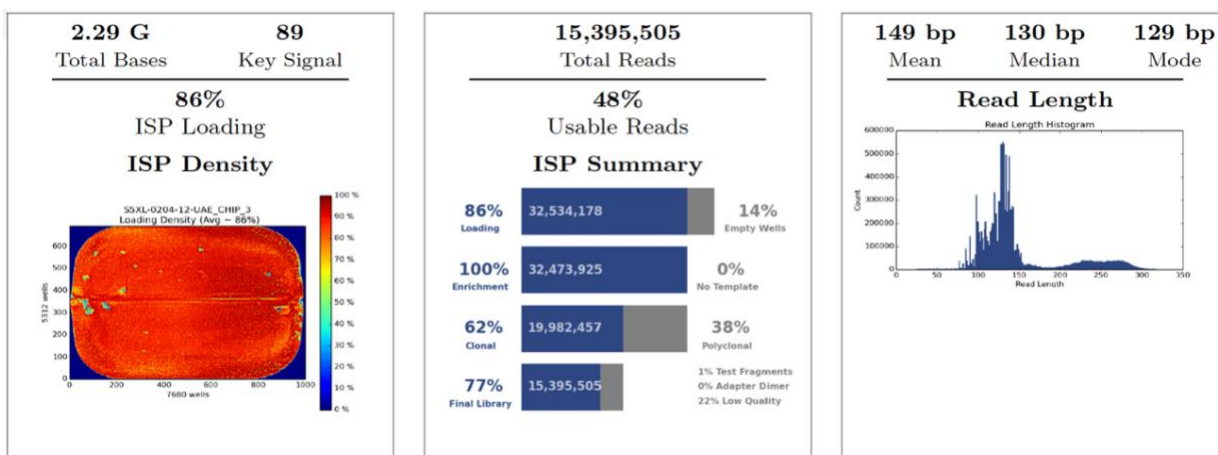


Figure 4.10. Emirati samples Chip 3 TSS quality check

Table 4.3. Quality Control Results of UAE samples

Parameters	Reads		
Samples no.	32	32	32
Chip number	154	155	156
Chip type	530v1	530v1	530v1
Chip loading (%)	86%	84%	86%
Enrichment (%)	99%	100%	100%
Polyclonal (%)	44%	44%	38%
Low quality (%)	26%	37%	22%
adapter Dimer (%)	0.1%	0.0%	0.1%
Usable reads (%)	41%	35%	48%
Mapped usable reads (%)	40.9%	35.2%	47.7%
Mean Read length (%)	155 bp	173 bp	149
Total bases	2.05 G	1.92 G	2.29 G

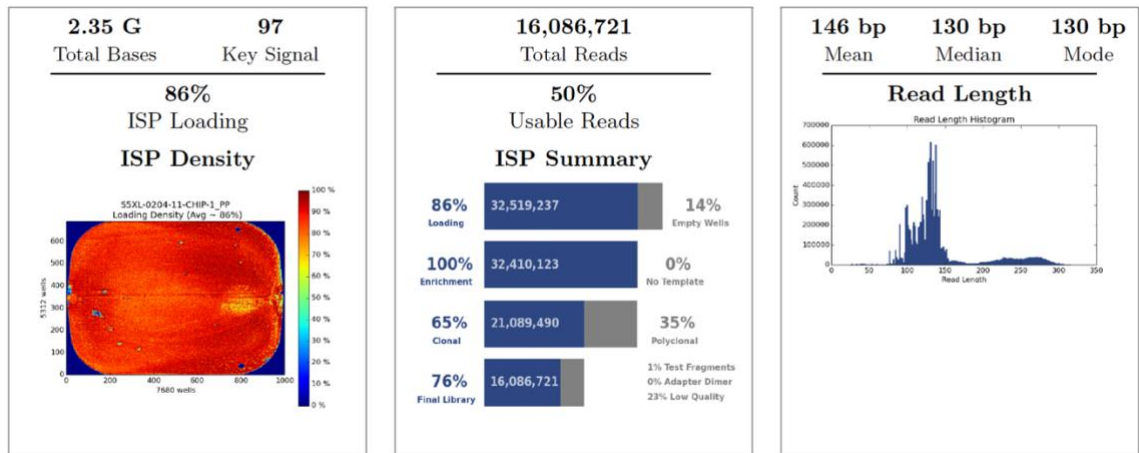


Figure 4.11. Pakistani samples Chip 1 TSS quality check

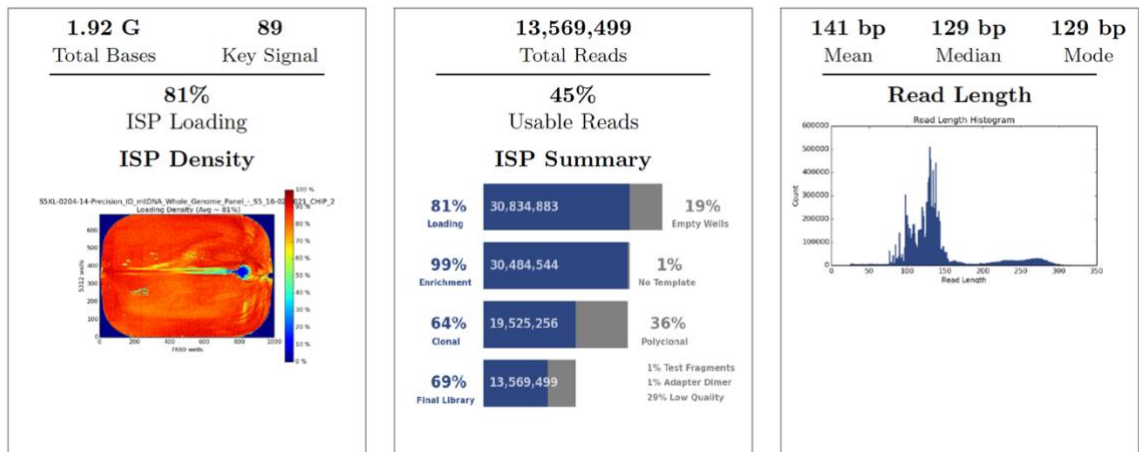


Figure 4.12. Pakistani samples Chip 2 TSS quality check

Table 4.4. Quality Control Results of Pakistanis population samples

Parameters	Reads	
Samples no.	18	32
Chip number	163	162
Chip type	530v1	530v1
Chip loading (%)	81%	86%
Enrichment (%)	99%	100%
Polyclonal (%)	36%	35%
Low quality (%)	29%	23%
adapter Dimer (%)	0.5%	0.2%
Usable reads (%)	45%	50%
Mapped usable reads (%)	44.9%	50.0%
Mean Read length (%)	141 bp	146 bp
Total bases	1.92 G	2.35 G

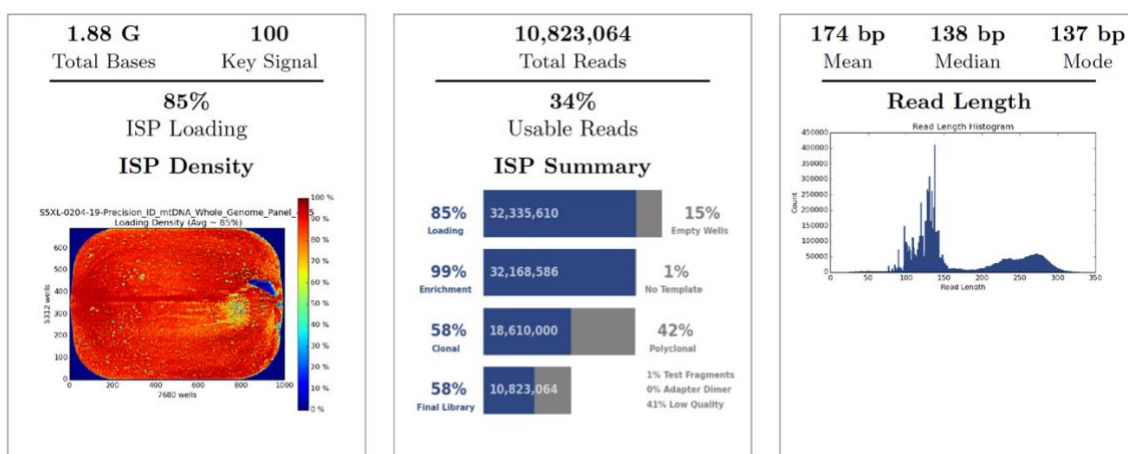


Figure 4.13. Indian samples Chip 1 TSS quality check

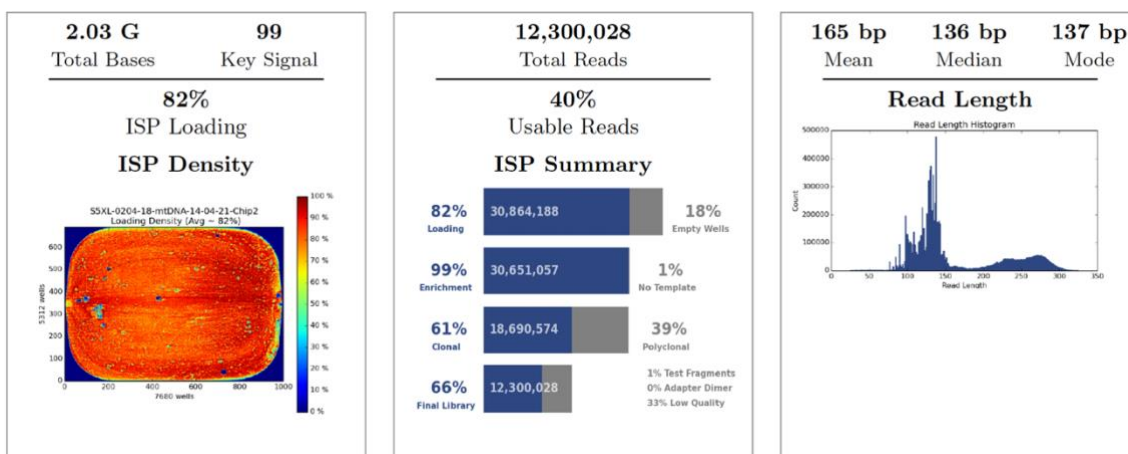


Figure 4.14. Indian samples Chip 2 TSS quality check

Table 4.5. Quality Control Results of Indian population samples

Parameters	Reads	
Samples no.	18	32
Chip number	170	171
Chip type	530v1	530v1
Chip loading (%)	82%	85%
Enrichment (%)	99%	99%
Polyclonal (%)	39%	42%
Low quality (%)	33%	41%
Adapter Dimer (%)	0.1%	0.0%
Usable reads (%)	40%	34%
Mapped usable reads (%)	40.3%	33.9%
Mean Read length (%)	165 bp	174 bp
Total bases	2.03 G	1.88 G

4.9.MPS Data Generation

The study implemented rigorous Quality Control (QC) measures, to ensure the authenticity of the data. These QC procedures were essential to address common issues such as nuclear mitochondrial DNA segments (NUMTs) interference, which could compromise the accuracy of mtDNA analyses (Taylor et al., 2020). The results showed that some NUMTs interference was detected, yet overall impact on data quality was negligible.

4.10. Mixtures Detection

In MPS data, each position in the mitochondrial genome is usually represented by a single base at each position in a single-source sample. However, the detection of two or more states at single nucleotide positions within the mtDNA is indicative of a mixture. This occurs because the MPS captures all the DNA present in the sample, including variations from different individuals. These multiple signals represent different nucleotides at the same position, indicating the presence of DNA from more than one individual.

4.11. Validation of Results Using Standards

To ensure the robustness and accuracy of the MPS workflow, various standards and controls from different manufacturers were sequenced. The sequencing included the 2800M Control DNA and 9947A from Promega, which were widely recognized for their consistency and reliability in genetic analysis. The 2800M Control DNA, specifically, was designed to serve as a male human genomic standard, containing a known allele frequency

that facilitated the calibration of genetic testing systems. Similarly, the 9947A control, also from Promega, provided a female DNA profile, often used to validate new methods or to benchmark performance across different laboratories.

Additionally, the 9947A DNA sample from Origene was included to compare with Promega's version. This provided an opportunity to assess inter-manufacturer variability and ensure that the MPS workflow was robust across different sources of the same control.

Moreover, the Human Mitochondrial DNA Standard Reference Material (SRM-2392) from the National Institute of Standards and Technology (NIST) was sequenced. This reference material was critical for ensuring precision in mitochondrial DNA analyses, providing a standardized mitochondrial sequence that aided in the evaluation of techniques and methodologies in forensic and genetic research.

Through the careful sequencing of these standards and controls, the MPS workflow was assessed and validated for accuracy and reproducibility. This approach involved using reference standards and controls to evaluate and optimize the sequencing process. While the fine-tuning was based on the performance metrics obtained from these standards and controls, it was not necessarily the sole basis for adjustments but rather a key component in ensuring adherence to high standards of genetic analysis.

Two standards supplied by Promega, 2800M Control DNA at a concentration of 10 ng/μl, and 9947A were sequenced. The results from the 2800M Control DNA were compared to

the reference data processed with the Promega PowerSeq® Whole Mito Amp and Prep Kit. Although minor differences might be expected due to differing techniques, the results were a 100% match with the reference, as shown in Table 4.6. For the 9947A standard, the sequencing reads were compared to the reference, and the results are presented in Table 4.7. Most of the polymorphisms matched, except for two variants, 309.2C and 3107c, which were not detected in this study's run. The variant 3107c was present but not detected as a variant since it was a true known variant confirmed by the manufacturer after correcting the errors of rCRS (i.e. the 3017c was incorporated into the reference genome and so not identified as a variant) (HumanMitoSeq, n.d.). Figure 4.15 shows the absence of nucleotide C at position 3107 of this standard, which was confirmed to be present, while 309.2C was manually confirmed.

Origene provided the 9947A standard. The sequencing reads were compared to the reference, and the results are presented in Table 4.8. Most of the polymorphisms matched, except for one variant, 309.2C which was an insertion. It was then manually confirmed by the converge software.

In addition to the sequencing of various standards, a No Template Control (NTC) was also run as an essential part of our quality control procedures for each batch of sequencing. The NTC, containing no DNA, served to ensure that there was no contamination across samples and that all reagents were free from DNA that could lead to false positives or other artifacts in the sequencing results. This control was critical for confirming the reliability of the current sequencing data, as it helped to identify any issues related to

contamination or cross-contamination, thereby upholding the standards required for forensic analysis.

Table 4.6. Comparison of the obtained sequence of 2800M promega to the reference

Polymorphism sites	References	This Study
152	152C	152C
263	263G	263G
309	309c	309c
315	315.1C	315.1C
477	477C	477C
750	750G	750G
1438	1438G	1438G
3010	3010A	3010A
4769	4769G	4769G
8860	8860G	8860G
15326	15326G	15326G
16519	16519C	16519C

Table 4.7. Comparison of Sequencing Results with Promega 9947A Standard

Polymorphism sites	References	This Study
93	93G	93G
195	195G	195G
214	214G	214G
263	263G	263G
309.1	309.1C	309.1C
309.2	309.2C	Manually confirmed
315	315.1C	315.1C
750	750G	750G
1438	1438G	1438G
3107	3107c	Manually confirmed
4135	4135C	4135C
4769	4769G	4769G
7645	7645C	7645C
7861	7861Y	7861Y
8448	8448C	8448C
8860	8860G	8860G
9315	9315C	9315C
13572	13572C	13572C
13759	13759A	13759A
15326	15326G	15326G
16311	16311C	16311C
16519	16519C	16519C

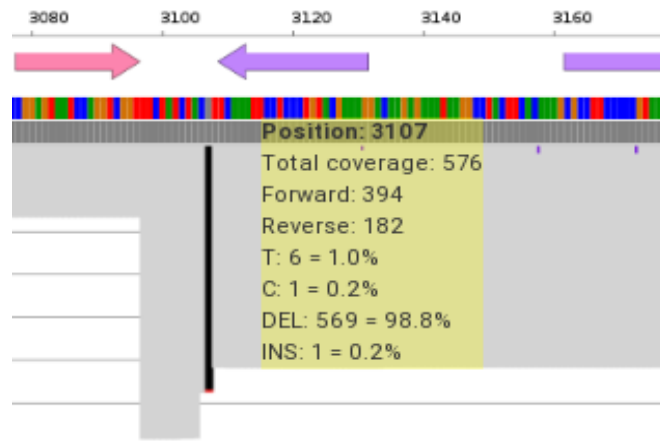


Figure 4.15. Deletion at position 3107 with 98.8% in Promega 9947 standard.

Table 4.8. Comparison of Sequencing Results with Origene 9947A Standard

Polymorphism sites	References	This Study
93	93G	93G
195	195C	195C
214	214G	214G
263	263G	263G
309.1	309.1C	309.1C
309.2	309.2C	Manually confirmed
315.1	315.1C	315.1C
750	750G	750G
1438	1438G	1438G
3107	3107c	3107c
4135	4135C	4135C
4769	4769G	4769G
7645	7645C	7645C
7861	7861C	7861C
8448	8448C	8448C
8860	8860G	8860G
9315	9315C	9315C
13572	13572C	13572C
13759	13759A	13759A
15326	15326G	15326G
16311	16311C	16311C
16519	16519C	16519C

Additionally, the NIST Human Mitochondrial DNA SRM-2392 standards were analyzed.

This SRM included three components. Component A, extracted DNA from cell culture line CHR, containing 60 μL of DNA at approximately 1 ng/ μL . Component B, extracted DNA from cell culture line GM09947A, containing 60 μL of DNA at approximately 1 ng/ μL . Component C, cloned DNA from the CHR HV1 region containing the C-stretch, with 10 μL of DNA at approximately 100 ng/ μL . These components were analysed, and the results were compared with the findings, as detailed in the Tables 4.9, 4.10, and 4.11. The use of

these well-characterized standards allowed for cross-validation of the work, ensuring the reliability and accuracy of the MPS results. Its worth mentioning here that non-detection of expected variant is normally common in position 3107c where it is no longer considered a variation as it appears as a known variation in all samples. Thus, the detection of this variation is not reported as variation but confirmed manually using the IGV view. Variant 14470C in Component CHR (A) was verified using IGV view, (Figure 4.16). Position 309.2C (Component GM09947A (B)) was also viewed via IGV view to verify the presence of the variation, (Figure 4.17).

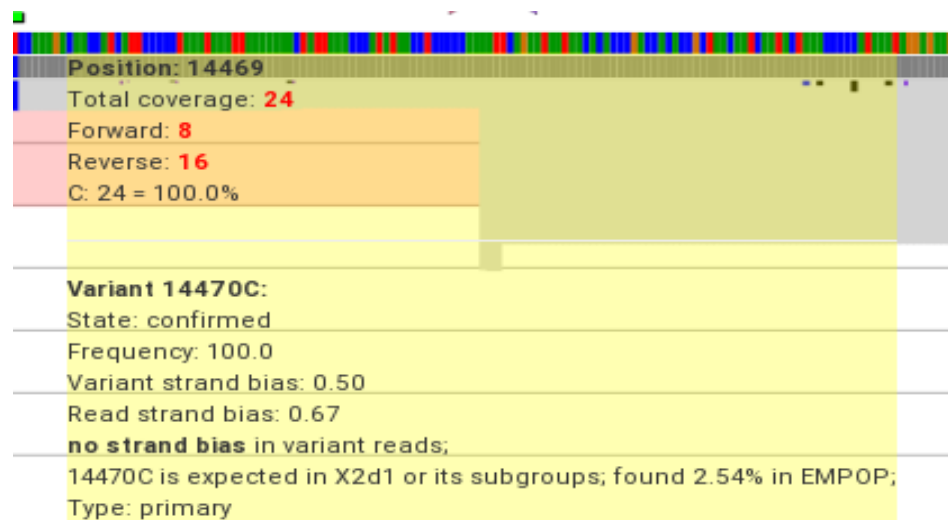


Figure 4.16. Position 14470C in Component CHR (A) with 100% confirmed cytosine.

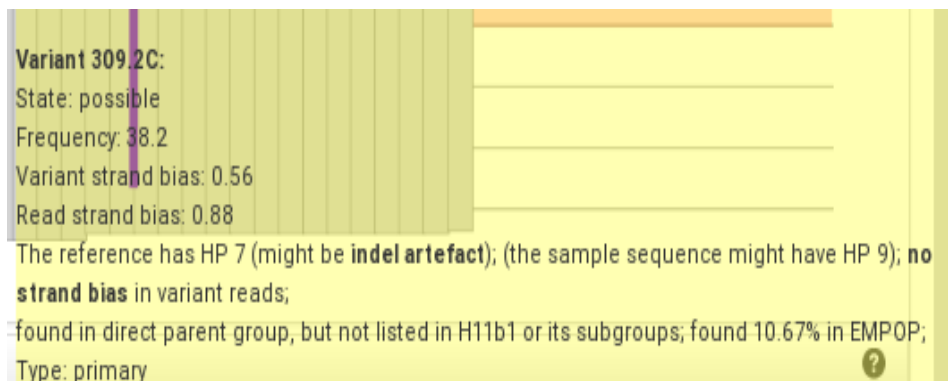


Figure 4.17. Position 309.2C of Component GM09947A (B) with 80.6% confirmed insertion.

In Component CHR clone (C), the 16193.1C, and was confirmed with NIST that this position could be missing out during the sequencing process due to its position in the poly-C stretch. This type of sequence, where a single nucleotide repeats multiple times in a row, is normally referred to as a homopolymeric or polymeric tract. In mtDNA, such regions are common and can be prone to sequencing errors or variations in the number of repeated bases among different individuals.

Table 4.9. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component CHR (A).

Polymorphism sites	References	This Study
64	64Y	64Y
73	73G	73G
195	195C	195C
204	204C	204C
207	207A	207A
263	263G	263G
309.1	309.1C	309.1C
315.1	315.1C	315.1C
709	709A	709A
750	750G	750G
1438	1438G	1438G
1719	1719A	1719A
3107	3107c	Manually confirmed
4769	4769G	4769G
4929	4929T	4929T
5186	5186G	5186G
6221	6221C	6221C
6371	6371T	6371T
6791	6791G	6791G
7028	7028T	7028T
8503	8503C	8503C
8860	8860G	8860G
11719	11719A	11719A
11878	11878C	11878C
12612	12612G	12612G
12705	12705T	12705T

13708	13708A	13708A
13966	13966G	13966G
14470	14470C	Manually confirmed
14766	14766T	14766T
15326	15326G	15326G
16183	16183C	16183C
16189	16189C	16189C
16223	16223T	16223T
16278	16278T	16278T
16519	16519C	16519C

Table 4.10. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component GM09947A (B).

Polymorphism sites	References	This Study
93	93G	93G
195	195C	195C
214	214G	214G
263	263G	263G
309.1	309.1C	309.1C
309.2	309.2C	Manually confirmed
315.1	315.1C	315.1C
750	750G	750G
1438	1438G	1438G
3107	3107c	Manually confirmed
4135	4135C	4135C
4769	4769G	4769G
7645	7645C	7645C
7861	7861C	7861C
8448	8448C	8448C
8860	8860G	8860G
9315	9315C	9315C
13572	13572C	13572C
13759	13759A	13759A
15326	15326G	15326G
16311	16311C	16311C
16519	16519C	16519C

Table 4.11. Comparison of Sequencing Results with NIST Human Mitochondrial DNA SRM-2392 Standard Components; Component CHR clone (C).

Polymorphism sites	References	This Study
16183	16183a	16183a
16189	16189C	16189C
16193.1	16193.1C	-
16223	16223T	16223T
16278	16278T	16278T
16519	16519C	16519C

4.12. Proficiency Testing

Validation and proficiency testing remain critical to ensuring the reliability of forensic DNA analyses. NIST (2024) reports emphasize the role of human factors in DNA interpretation, underscoring the importance of robust standards and training in forensic laboratories. In addition to the sequencing of various DNA standards and controls, the examination of GEDNAP (German DNA Profiling Group) proficiency tests was included in this study. These tests are internationally renowned for their role in evaluating the quality and performance of DNA profiling laboratories. By incorporating GEDNAP proficiency testing, the aim was to adhere to global standards and ensure the methods met international benchmarks.

For this round of the proficiency testing, this study utilized samples from two sets - 66 and 67. Each set had 3 reference blood samples labelled as Person A, Person B, and Person C, amounting a total 6 reference samples. These samples represented anonymized

individual profiles that were specifically prepared by GEDNAP to challenge and verify the accuracy and reliability of mtDNA sequencing techniques being employed.

The results from these tests demonstrated a successful 100% match across all individuals, affirming the high precision and dependability of the MPS techniques employed in this study. Tables 4.12 and 4.13 summarized the results of GEDNAP 66 and GEDNAP 67, respectively.

The MPS technology allowed for detailed genetic analysis of this proficiency test samples, providing detailed insights into the level of detection, mixtures, and overall data integrity achievable with our setup. By successfully analyzing these samples, the laboratory at Dubai Police Forensic Department demonstrated its capability to produce reliable and consistent results, further establishing adherence to best practices in forensic DNA analysis.

Through the application of GEDNAP proficiency tests, this study not only validated the employed Ion torrent technology methodology for MPS, but also benchmarked the performance of the workflow against other leading laboratories in the field. This is a crucial practice in forensic laboratories taken as continual assessment that helps identifying areas of improvement and in maintaining a high standard of forensic DNA profiling.

The testing of reference sample and proficiency test samples have provided a high degree of confidence in the data obtained and illustrated that the processing and analysis streams are providing robust data.

Table 4.12. Comparison between the examination reference and the results of GEDNAP 66 sequences of Person A, Person B, and Person C, obtained using Ion Torrent Technologies and Precision ID Whole mtDNA genome panel.

References	Person A	References	Person B	References	Person C
A73G	73G	A73G	73G	A73G	73G
G185A	185A	C150T	150T	A263G	263G
G228A	228A	A263G	263G	-315.1C	315.1C
A263G	263G	-315.1C	315.1C	C497T	497T
C295T	295T	C16192T	16192T	A16183C	16183C
-309.1C	309.1C	T16311C	16311C	T16189C	16189C
-315.1C	315.1C			T16224C	16224C
C462T	462T			T16311C	16311C
T489C	489C			T16519C	16519C

Table 4.13. Comparison between the examination reference and the results of GEDNAP 67 sequences of Person A, Person B, and Person C, obtained using Ion Torrent Technologies and Precision ID Whole mtDNA genome panel.

References	Person A	References	Person B	References	Person C
A263G	263G	T152C	152C	T146C	146C
-315.1C	315.1C	A263G	263G	T239C	239C
C16234T	16234T	-309.1C	309.1C	A263G	263G
T16311C	16311C	-315.1C	315.1C	-309.1C	309.1C
		C456T	456T	-315.1C	315.1C
		A523DEL	523del	T16362C	16362C
		C524DEL	524del		
		T16304C	16304C		

4.13. Discussion

The findings in this chapter highlighted the successful implementation of MPS using Ion Torrent technology for the analysis of Standard Reference Materials (SRM), control DNA, and GEDNAP proficiency testing samples. The validation of MPS through SRM and control DNA ensured the accuracy, reproducibility, and integrity of sequencing results. SRMs serve as benchmark materials for forensic DNA testing, allowing laboratories to assess sequencing performance against known reference standards, which is essential for ensuring comparability across different forensic laboratories (Guo et al., 2024; NIST, 2024). The results obtained reaffirm that the use of SRMs is the gold standard practice in forensic mtDNA analysis, providing a robust foundation for quality control.

The inclusion of GEDNAP proficiency testing further strengthened the validation process. Proficiency testing is a critical aspect of forensic accreditation, ensuring that laboratories consistently produce high-quality and reliable mtDNA sequencing results. The successful processing of GEDNAP samples in this study confirmed the laboratory's practical competency in handling forensic casework samples using MPS, supporting previous findings that emphasized the importance of external proficiency assessments in forensic genomics (Faith, 2018).

One of the most significant advantages observed was the sensitivity of MPS in detecting low-level heteroplasmy. The ability to identify minor variant populations is essential for forensic DNA analysis, as heteroplasmy enhances individual discrimination and kinship

analysis. This observation is consistent with findings from Parson et al. (2023), who demonstrated that MPS provides superior resolution for mtDNA heteroplasmy detection in forensic casework.

4.14. Conclusion

This chapter has demonstrated that MPS is a highly reliable technology for mtDNA sequencing. The results confirm that MPS offers high accuracy and reproducibility, making it a dependable sequencing approach for forensic applications. In forensic mtDNA analysis, the use of SRM and control DNA is the gold standard practice, ensuring that sequencing results maintain high integrity and comparability across forensic laboratories (NIST, 2024).

Additionally, GEDNAP proficiency testing plays a vital role in validating laboratory competency in forensic sequencing workflows. The successful processing of GEDNAP samples in this study underscores the laboratory's ability to produce high-quality mtDNA results, reinforcing the significance of external proficiency assessments as a forensic accreditation requirement (Faith, 2018).

In conclusion, the results confirmed that MPS is a reliable approach for mtDNA analysis, offering high-resolution sequencing, and reliability. By incorporating SRM validation, control DNA assessments, and GEDNAP proficiency testing, this supports the adoption of MPS as a routine forensic mtDNA sequencing technology.

Chapter 5

5. Massive Parallel Sequencing Data Analysis

5.1. Introduction

In Chapter 3, Sanger sequencing was employed to sequence the mtDNA control region (CR), followed by a analysis to obtain haplotypes and an assessment of genetic diversity in the different populations. In this Chapter, the focus shifts to an in-depth analysis of mtDNA haplotypes detected in 510 Emirati samples, 50 Indian samples, and 50 Pakistani samples using Massive Parallel Sequencing (MPS). The primary objective of this research was to provide high-quality genetic data specific to the Emirati population, with an emphasis on haplotypes, which are crucial for forensic and population genetics studies. While haplogroups are also examined, they play a secondary role in this study. The analysis of haplogroups provides additional insights into the broader phylogenetic relationships and evolutionary history of the populations studied. These data can improve the accuracy and reliability of forensic analyses in the region. Moreover, they can also contribute to the understanding of population genetics, including genetic diversity, migration patterns, and evolutionary history. By examining mtDNA haplotypes and haplogroups. The aims of this chapter of the study provide scientifically valuable and practically applicable data for forensic contexts.

In this chapter high-quality data were to provided using advancement of the MPS technology and its application to mitochondrial DNA (mtDNA) analysis in forensic genetics, which the primary aim of this PhD study.

5.2. Chapter Objectives

- To evaluate and compare the accuracy and reliability of mtDNA profiles generated by Sanger sequencing, aiming to identify discrepancies and consistencies between Sanger sequencing and MPS methods to enhance the forensic analysis of mtDNA.
- To provide a dataset of mtDNA haplotypes and haplogroups in Emirati, Indian, and Pakistani populations, contributing valuable data to the global mtDNA database.
- To investigate the presence and frequency of heteroplasmy and nuclear mitochondrial DNA segments (NUMTs) in the analysed samples to ensure accurate interpretation of mtDNA data.

5.3. Methods

Note that detailed methodology can be obtained in Chapter 2 for comparison.

5.4. Results: Section 1

5.5. MPS Data Analysis: Emiratis Samples Set

The results of haplotype analysis from 510 Emirati samples using MPS revealed that each sample exhibited a set of SNPs variations; including insertion and deletions that defined individual haplotypes. Certain haplotypes were found to be shared among multiple

samples, indicating common genetic lineages within the population, while others were rare. The SNP variations were the basis of the haplogroup assignment for each sample.

5.5.1. Haplotypes Generating and Assessment

Table 5.1 summarizes the 33 samples used for concordance between Sanger sequence and MPS, while the remaining samples are included in the Appendices. The data, generated using MPS, were fully concordant with the results obtained by Sanger sequence. The results aligned with existing standards and established methodologies for mtDNA sequencing, demonstrating high accuracy and reproducibility. Over the past decade, MPS technology played a transformative role in forensic genetics, offering high concordance rates with traditional sequencing methods while surpassing them in sensitivity and throughput. These advancements paralleled innovations in other forensic disciplines, where technological improvements significantly enhanced methodologies and best practices across various sub-fields (Breitinger et al., 2024).

Table 5.1. Whole mtDNA genome sequences for 33 Emirati samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing (as shown in Table 3.2) are indicated in italics, and insertions/deletions (indels) are highlighted in red.

No.	Samples	Haplogroups	Haplotypes
1	UAE_01	H14b	263G <i>315.1C</i> 750G 1438G 3197C 4769G 7645C 8860G 10217G 10685A 15326G 16319A
2	UAE_02	N1b1a2	73G 152C 263G <i>315.1C</i> 750G 1438G 1598A 1703T 1719A 2639T 2706G 3921A 4769G 4904T 4960T 5237A 5471A 6272G 7028T 8251A 8472T 8836G 8860G 9335T 10238C 11362G 11719A 12501A 12705T 12822G 14766T 15326G 16093C 16145A 16176G 16223T 16390A 16519C
3	UAE_03	N1b1a2	73G 152C 263G <i>315.1C</i> 750G 1438G 1598A 1703T 1719A 2639T 2706G 3921A 4769G 4904T 4960T 5237A 5471A 6272G 7028T 8251A 8472T 8836G 8860G 9335T 10238C 11362G 11719A 12501A 12705T 12822G 14766T 15326G 16093C 16145A 16176G 16223T 16390A 16519C
4	UAE_04	R30b2a	73G 237G 263G <i>315.1C</i> 373G 482C 750G 1438G 2414T 2706G 4769G 6290T 7028T 7280T 7843G 8584A 8590T 8639G 8860G 11719A 13539G 14000A 14560A 14766T 15148A 15326G 16086C 16292T 16519C
5	UAE_05	K1a4c	73G 263G 497T <i>524.1ACAC</i> 750G 1189C 1438G 1811G 2706G 3480G 4769G 7028T 8860G 9055A 9698C 10398G 10550G 11299C 11467G 11485C 11719A 12308G 12346T 12372A 12612G 12969T 13827G 14070G 14167T 14766T 14798C 15326G 16224C 16256T 16311C 16519C
6	UAE_06	R0a1a	58C 64T 146C 263G <i>315.1C 302.1AC 523del 524del</i> 750G 827G 1438G 2442C 2706G 3847C 4769G 7028T 8292A 8860G 11761T 13188T 14727C 14766T 15326G 16126C 16265G 16355T 16362C

7	UAE_07	T1a1	73G 152C 195C 263G 315.1C 709A 750G 1438G 1888A 2706G 4216C 4769G 4917G 7028T 8860G 9025A 9286C 9899C 10463C 11251G 11719A 12633A 13368A 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16163G 16186T 16189C 16294T 16519C
8	UAE_08	H3c2b	195C 263G 315.1C 750G 1438G 4769G 6776C 8860G 12957C 14305A 15326G 16176T
9	UAE_09	L3d1a1a	73G 150T 152C 263G 315.1C 523del 524del 750G 921C 1438G 1503A 2706G 4048A 4203G 4769G 5147A 5471A 6680C 7028T 7242G 7648T 8616C 8701G 8860G 9540C 10398G 10640C 10873C 10915C 11719A 11887A 12705T 13105G 13886C 14284T 14766T 15301A 15326G 16124C 16223T 16319A
10	UAE_10	L1c3b1a	73G 151T 152C 182T 186A 189C 263G 315.1C 523del 524del 629C 750G 769A 825A 1018A 1438G 2283T 2394del 2706G 2758A 2885C 3210T 3434G 3594T 3666A 4104G 4755C 4769G 5951G 6071C 6221A 6917A 7028T 7146G 7256T 7389C 7521A 8027A 8251A 8417T 8655T 8860G 9072G 9540C 10398G 10586A 10688A 10810C 10873C 11302T 11317G 11719A 12542T 12705T 12810G 13105G 13485G 13506T 13789C 13981T 14000A 14178C 14766T 14794T 14911T 15226G 15326G 15905C 15978T 16017C 16129A 16163G 16187T 16189C 16209C 16223T 16278T 16293G 16294T 16311C 16360T 16519C
11	UAE_11	L3d1a1a	73G 150T 152C 263G 315.1C 523del 524del 750G 921C 1438G 1503A 2706G 3200C 4048A 4203G 4769G 5147A 5471A 6680C 7028T 7424G 7648T 8618C 8701G 8860G 9540C 10398G 10640C 10873C 11719A 11887A 12705T 13105G 13886C 14284T 14766T 15301A 15326G 16124C 16223T 16319A
12	UAE_12	HV18	263G 315.1C 750G 1438G 2706G 4769G 7028T 8860G 9039A 9899C 15326G 16189C 16519C
13	UAE_13	K2a2a	73G 146C 152C 263G 315.1C 709A 750G 1438G 1811G 2706G 3480G 4561C 4769G 7028T 8860G 9055A 9698C 9716C 10550G 11299C 11348T 11467G 11719A 12308G 12372A 14167T 14766T 14798C 15326G 16213A 16224C 16311C 16519C

14	UAE_14	H6a1a1	239C 263G 315.1C 302.1AC 567.1C 573.1C 750G 1438G 3915A 4727G 4769G 5460A 8860G 9380A 11253C 15326G 16362C 16482G
15	UAE_15	L2a1b1a	73G 146C 152C 195C 263G 263G 315.1C 750G 769A 1018A 1438G 2416C 2706G 2789T 3594T 4104G 4769G 5090C 7028T 7175C 7256T 7274T 7521A 7771G 8206A 8701G 8860G 9221G 9540C 10115C 10143A 10398G 10873C 11719A 11914A 11944C 12693G 12705T 13803G 14566G 14766T 15301A 15326G 15735T 15784C 16183del 16189C 16223T 16278T 16290T 16294T 16309G 16390A
16	UAE_16	U3b	73G 150T 195C 263G 315.1C 750G 1438G 1811G 2706G 2833G 3316A 4188G 4640A 4769G 5773A 7028T 8895C 9656C 11467G 11719A 12308G 12372A 12738C 13743C 14139G 14766T 15326G 15454C 16172C 16298C 16343G
17	UAE_17	K2a2a	73G 146C 152C 263G 315.1C 523del 524del 709A 750G 1438G 1811G 2706G 3480G 4561C 4769G 7028T 8860G 9055A 9698C 9716C 10550G 11299C 11348T 11467G 11719A 12308G 12372A 14167T 14766T 14798C 15326G 16213A 16224C 16311C 16519C
18	UAE_18	J2a2e	73G 95C 150T 152C 195C 263G 315.1C 295T 489C 750G 1438G 2706G 4216C 4769G 4890G 6671C 7028T 7476T 8860G 10398G 10499G 10685A 11002G 11251G 11377A 11719A 12570G 12612G 14364A 14766T 15257A 15326G 15452A 15679G 16069T 16126C 16207G 16519C
19	UAE_19	J1b1b3	73G 263G 271T 295T 315.1C 750G 1438G 2706G 3010A 4216C 4769G 5460A 7028T 8860G 10398G 11251G 11719A 12612G 14766T 15326G 15941C 16189C 16261T 16295T 16519C
20	UAE_20	I*	73G 146C 199C 204C 207A 250C 263G 315.1C 750G 1120T 1438G 1719A 1900G 2706G 4529T 4769G 4772C 5895T 7028T 8251A 8860G 10034C 10238C 10398G 11719A 12501A 12705T 13515T 13780G 13983T 14766T 15043A 15190T 15326G 15924G 16129A 16223T 16311C 16391A 16519C
21	UAE_21	H2a1	263G 315.1C 750G 951A 8860G 15326G 15401G 16354T

22	UAE_22	R0a1a1	58C 64T 146C 263G 315.1C 750G 827G 1438G 2442C 3847C 4769G 7028T 8292A 8860G 11761T 13188T 13708A 14766T 15326G 16126C 16355T 16362C
23	UAE_23	U9a	73G 263G 315.1C 499A 750G 1438G 1811G 2706G 3290C 3434G 3531A 3834A 4113A 4769G 5999C 6386T 7028T 8860G 11467G 11719A 12308G 12372A 12880C 14094C 14766T 15326G 15718C 16051G 16278T
24	UAE_24	M1a1b1b1	73G 195C 263G 315.1C 489C 750G 813G 930A 1438G 2706G 3705A 4769G 6446A 6671C 6680C 7028T 7853A 8701G 8860G 9053A 9540C 10398G 10400T 10873C 11719A 12346T 12705T 12950C 13152G 14110C 14766T 14769G 14783C 15043A 15301A 15326G 16129A 16189C 16223T 16249C 16311C 16359C 16519C
25	UAE_25	W6b1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1438G 2706G 4093G 4646C 4769G 5046A 5460A 6297C 7028T 7269A 8251A 8349T 8614C 8705C 8860G 8994A 11674T 11719A 11947G 12705T 14766T 15326G 15884C 15930A 16223T 16325C 16519C
26	UAE_26	U3a2a1	73G 143A 150T 152C 189G 200G 263G 315.1C 750G 1393A 1438G 1811G 2294G 2706G 4703C 4769G 6050C 6518T 7028T 8860G 9266A 10506G 11050C 11467G 11719A 11935C 12308G 12372A 13149C 13934T 14139G 14766T 15226G 15326G 15454C 16311C 16343G 16390A 16519C
27	UAE_27	H6b	152C 239C 263G 315.1C 750G 1438G 4769G 8860G 11032d 13231d 14040A 15326G 16093C 16300G 16362C 16482G 16519C
28	UAE_28	J1d1a1	73G 152C 263G 295T 315.1C 462T 489C 750G 1007A 1438G 2706G 3010A 4216C 4769G 5704T 7028T 7789A 7963G 8860G 10398G 11251G 11719A 12612G 13392C 14207A 14766T 15326G 15452A 16069T 16126C 16193T 16300G 16309G
29	UAE_29	J1d1a1	73G 152C 263G 295T 315.1C 462T 489C 750G 1007A 1438G 2706G 3010A 4216C 4769G 5460A 7028T 7789A 7963G 8860G 10398G 11251G 11719A 12612G 13392C 14766T 15326G 15452A 16069T 16126C 16193T 16300G 16309G

30	UAE_30	U2e1f1	73G 152C 217C 263G 315.1C 340T 499A 508G 524.1C 750G 1438G 1811G 2706G 3720G 4769G 6045T 6152C 6755A 7028T 8155A 8860G 9101C 10876G 11467G 11719A 12308G 12358G 12372A 13020C 13676G 13734C 14766T 15326G 15907G 16051G 16129C 16183d 16189C 16362C 16519C
31	UAE_31	U3b3	73G 150T 199C 263G 315.1C 750G 1438G 1811G 2706G 3531A 4188G 4640A 4769G 6962A 7028T 8860G 9424T 9656C 11467G 11719A 12308G 12372A 13470G 13743C 14139G 14485T 14766T 15326G 15454C 16168T 16343G 16519C
32	UAE_32	N1a3a	73G 189G 195C 204C 207A 210G 263G 315.1C 750G 1438G 1719A 2706G 4769G 6908C 7028T 8222C 8860G 10238C 11025C 11437C 11719A 11914A 12501A 12705T 13637G 13780G 13933G 14766T 15326G 16201T 16220G 16223T 16265G 16497G 16519C
33	UAE_33	HV	235G 263G 315.1C 523d 524d 750G 1438G 2706G 3397G 4655A 4769G 7028T 8860G 14207A 15326G 16223T

After retrieving the haplotypes and concordance alignment, haplotypes assessment was carried out based on the findings summarized in Table 5.2.

Table 5.2. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for the three populations sets.

Set	Ht	Hd	PD	PI
Emiratis (n=510)	445	0.9989	0.9970	0.00302

5.5.2. Haplogroups Assignments

Full concordance was achieved between the mtDNA profiles generated by Sanger sequencing and those obtained through MPS. This concordance supported the reliability and accuracy of the MPS method in analyzing mtDNA, demonstrating its effectiveness in providing consistent and reproducible results. Following this quality control step, the SNP variations were used to assign each sample to its closest haplogroup. Figure 5.1 summarized the results obtained, highlighting the genetic diversity and potential evolutionary insights specific to the Emirati population. Haplogroup assignment in mtDNA involves the process of classifying mitochondrial DNA sequences into distinct haplogroups based on specific variations. Normally, haplogroups are assigned by comparing an individual's mtDNA sequence to rCRS, on the PhyloTree, which catalogs known mtDNA haplogroups and their defining mutations. In forensic genetics, mtDNA haplogroup assignment is particularly valuable for determining maternal lineage and identifying genetic links between individuals or populations.

Cluster	Count	Percentage	Population Frequencies
H	79	15.49%	
U	70	13.73%	
J	56	10.98%	
M	55	10.78%	
R	55	10.78%	
N	44	8.63%	
T	25	4.9%	
K	24	4.71%	
HV	23	4.51%	
L3	21	4.12%	
L2	12	2.35%	
L1	9	1.76%	
X	9	1.76%	
L0	8	1.57%	
I	7	1.37%	
W	7	1.37%	
B	2	0.39%	
D	2	0.39%	
F	2	0.39%	

Figure 5.1. An overview of 19 haplogroups in 510 samples of the Emirati population, detailing sample counts, frequencies, and percentages. The coloured bar illustrates population frequencies for variants observed across different global populations. Each colour corresponds to a specific population category as follows: Purple: African/African American, Light Blue: European (non-Finnish), Dark Blue: European (Finnish), Teal: Amish, Lavender: Ashkenazi Jewish, Gold: Middle Eastern, Orange: South Asian, Green: East Asian, Red: Latino/Admixed American and Grey: Other. This representation was generated using Haplogrep 3 (Schönherr et al., 2023).

Haplogroup H

Haplogroup H emerges as the most prevalent in this study, comprising 15.49% of the population with a total of 79 samples. Primarily associated with the Gravettian culture in European (non-Finnish) populations, haplogroup H is estimated to have emerged around 25,000 years ago (Richards et al., 2000). It is also found in European (Finnish), African/African American, Latino/Admixed American, and Ashkenazi Jewish populations, with rare occurrences in the Amish community. Haplogroup H is defined by major SNPs 263G, 750G, 1438G, 4769G, 8860G, and 15326G. Within this study, haplogroup H encompassed numerous subhaplogroups (Figure 5.2), each characterized by specific SNPs. These include H1V, H35a, H5m, H13a1a1a, H17a, H13a1a1, H1a5, H5'36, H6c, H11a, H5a1a, H1b, H1u, H14a, H2a2a1a, H83, H10a, H13b1+200, H17b, H49a1, H5r, H5, H5a1n, H6a1a, H14b3, H26a1, H33, and H107. Additionally, the subclade Haplogroup HV, comprising 4.51% of this study population, and it was characterized mainly by 2706G and 7028T. Its subhaplogroups included HV+16311, HV18, HV6, and HV+73.

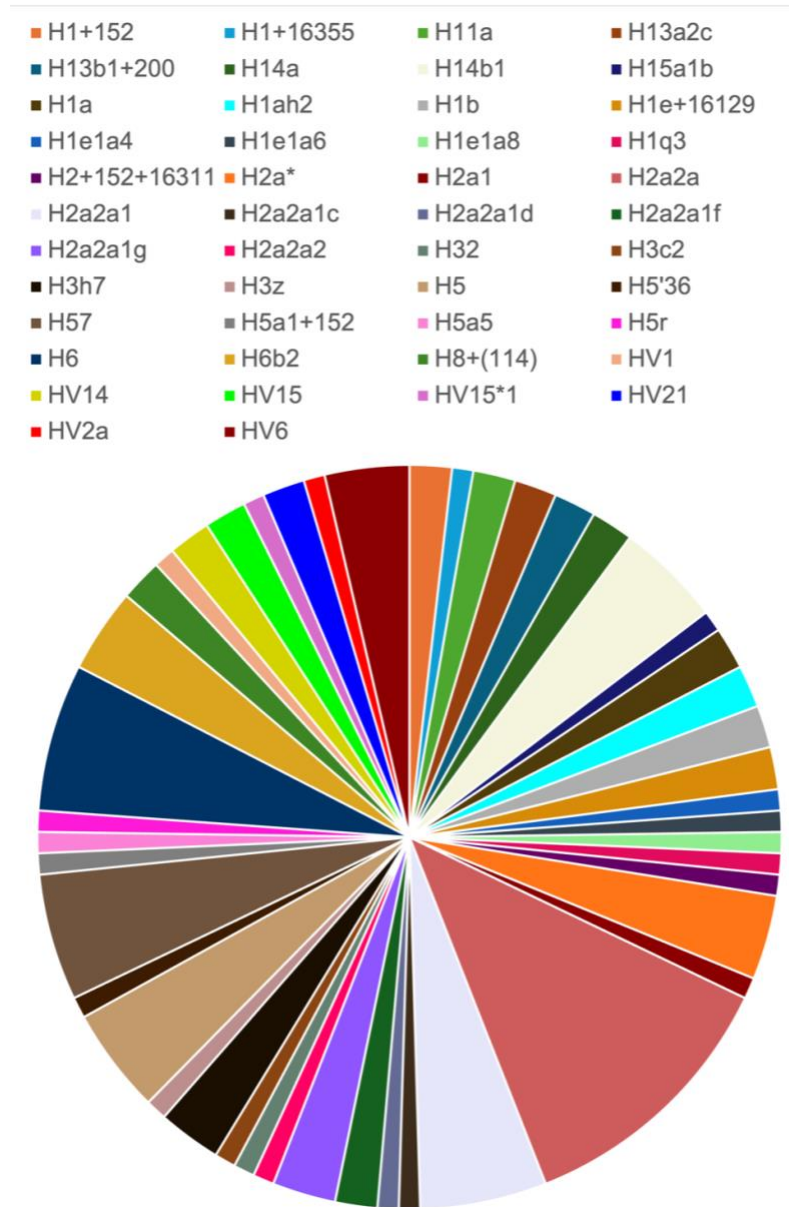


Figure 5.2. The distribution of haplogroup H within the population is illustrated in the accompanying chart. The colour-coded haplogroups are as follows: H1+152 (red), H13b1+200 (dark brown), H1a (brown), H1e1a4 (blue-grey), H2+152+16311 (purple), H2a* (light beige), H2a2a1 (white), H2a2a1c (light grey), H2a2a2 (pink), H3h7 (dark brown), H57 (brown), H6 (green), HV14 (golden yellow), HV2a (red-orange), H1+16355 (deep blue), H14a (teal), H1ah2 (cyan), H1e1a6 (grey-blue), H2a1 (dark green), H2a2a1d (olive green), H32 (light grey), H5 (light brown), H5a1+152 (brown), H6b2 (yellow-brown), HV15 (bright green), HV6 (lime green), H11a (light yellow-green), H14b1 (pale yellow), H1b (light grey), H1e+16129 (dark grey), H1q3 (pink), H2a2a1f (green), H3c2 (light olive), H5'36 (brownish pink), H5a5 (light pink), H5r (hot pink), H8+(114) (light brown), HV1 (magenta), HV15*1 (soft pink), and HV21 (navy blue).

Haplogroup U

In this study, haplogroup U constituted approximately 13.73% of the population, based on 50 samples. Originating from Europe or the Near East around 55,000 years ago, haplogroup U was closely linked to ancient European hunter-gatherers and was part of macrohaplogroup R, a descendant of haplogroup N (Kivisild et al., 2006). Haplogroup U had diversified into several subclades, including U1, U2, U3, U4, U5, U6, U7, and U8. It is widespread across Europe, with subclades like U5 being particularly common among ancient hunter-gatherers. U4 and U8 were also well-represented in European populations, with U4 additionally found in Siberian populations, indicating ancient migrations across the Eurasian steppe. Subclades U1, U3, and U7 are prevalent in the Near East, suggesting early human migrations and population mixing. U2 was significant in South and Central Asia, pointing to ancient migrations into the Indian subcontinent, while U6 was predominantly found in North Africa, particularly among Berber populations, with traces in the Iberian Peninsula. Specific SNPs defined Haplogroup U and its subclades: 11467G and 12308A for Haplogroup U; 10238C and 16311T for U1; 16051T and 16189C for U2; 16343G and 16390C for U3; 16356A and 195T for U4; 12308G and 1811G for U5; 3348C and 7768A for U6; 9410A and 11204G for U7; and 9055A and 12372T for U8 (Figure 5.3).

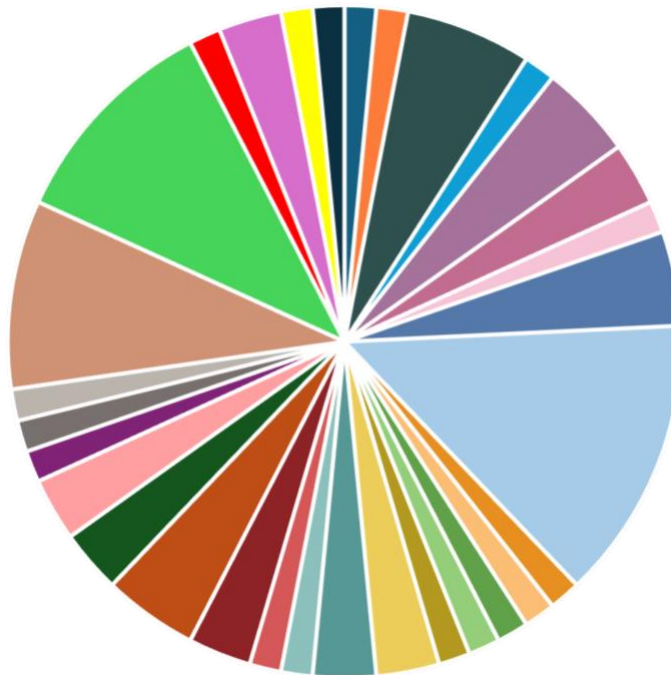
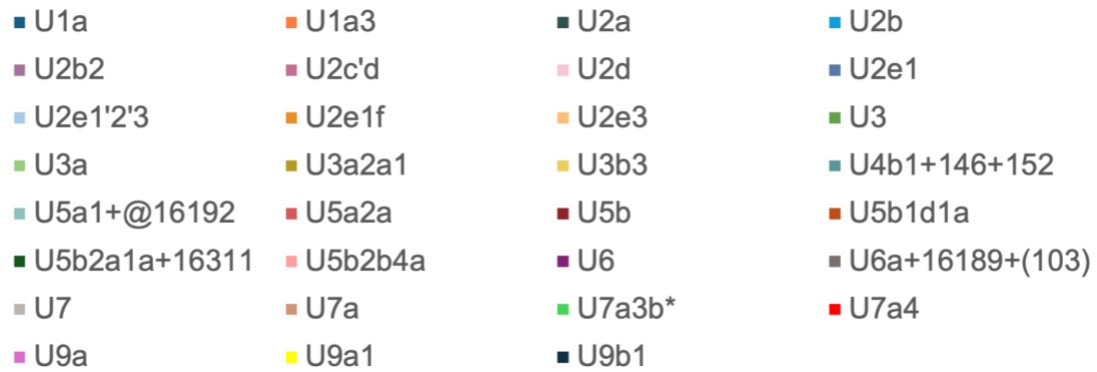


Figure 5.3. The distribution of haplogroup U within the population is illustrated in the accompanying chart. The colour-coded haplogroups are as follows: U1a (light blue), U2b2 (lavender), U2e1'2'3 (sky blue), U3a (green), U5a1+@16192 (pale green), U5b2a1a+16311 (dark green), U7 (white), U9a (pink), U1a3 (orange), U2c'd (light pink), U2e1f (gold), U3a2a1 (yellow), U5a2a (mustard yellow), U5b2b4a (olive), U2a (blue), U2b (dark blue), U2d (pale pink), U2e3 (orange), U3b3 (yellow-orange), U5b (red), U6 (purple), U7a3b (pale green), U9a1 (yellow), U9b1 (navy blue), U2e1 (light grey), U3 (light green), U4b1+146+152 (dark green), U5b1d1a (brown), U6a+16189+(103) (grey), and U7a4 (bright red).

Haplogroup J

Haplogroup J accounted for 10.98% of this study population, totaling 56 samples, and it was predominantly found in European (non-Finnish) populations. It was also observed in European (Finnish), Ashkenazi Jewish, Latino/Admixed American, South Asian, and African/African American populations, albeit rarely (Haplogrep 3 Schönherr et al., 2023). Characterized by key SNPs 295T, 498C, 10398G, 12612G, 13708A, and 16069T, haplogroup J encompassed a diverse range of subhaplogroups within this study (Figure 5.4). These included J2a2b, J1c5, J1b1a1, J1c, J1c2, J1b*, J1b, J2a2, J1c3, J1c4, J2a2e, J1b1b3, J1d1a1, J1b2, J1b3a, J2a2a1, J1d1a, J1b1b1, JT, J1d, and J1d3a. Subclade J1 is characterized by mutations C462T, G3010A, while J2 is defined by T152C!, C7476T, and G15257A.

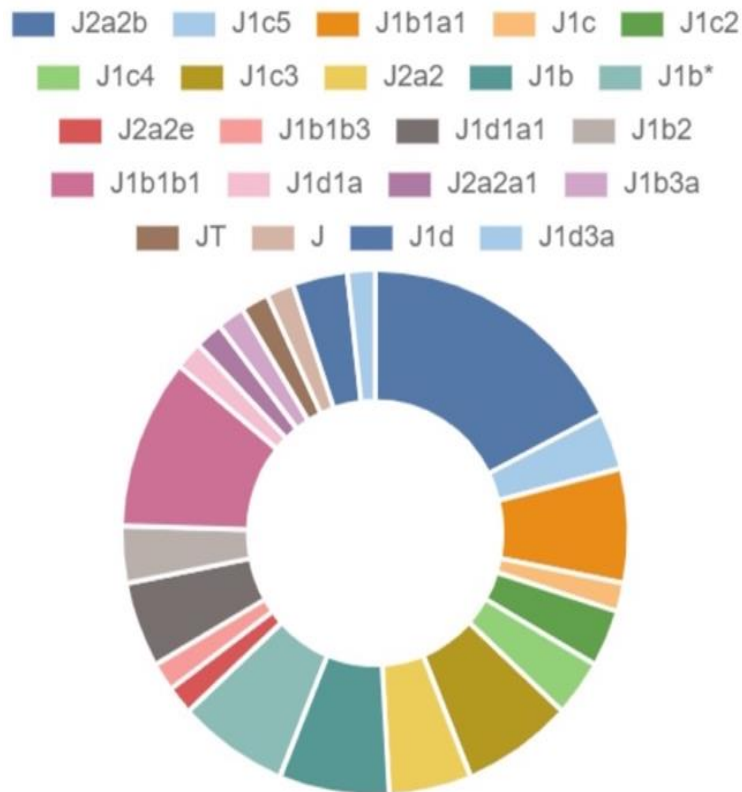


Figure 5.4. The distribution of haplogroup J within is illustrated in the chart, generated using Haplogrep 3 software. The colour-coded haplogroups are: J2a2b (dark blue), J1c5 (light blue), J1b1a1 (dark orange), J1c (peach), J1c2 (dark green), J1c4 (pale green), J1c3 (mustard yellow), J2a2 (pale yellow), J1b (teal), J1b* (light turquoise), J2a2e (red), J1b1b3 (light pink), J1d1a1 (grey), J1b2 (light grey), J1b1b1 (dark pink), J1d1a (light pink), J2a2a1 (lavender), J1b3a (pale purple), JT (light brown), J (pale beige), J1d (royal blue), J1d3a (sky blue).

Haplogroup M

The results reported 55 samples that constituted 10.78% of haplogroup M in the Emirati sets. This haplogroup originated from Asia, around 60,000 years ago, soon after the out-of-Africa migration. This migration led to the emergence of Haplogroups M and N, both of which are descendants of Haplogroup L3. Interestingly, haplogroup M was nearly absent in Europe but prevalent in other regions. It was predominantly found in Asia, Oceania, and Indigenous American populations, reflecting early human migrations (Behar

et al., 2012). In Oceania, various subclades of Haplogroup M were found, particularly through Southeast Asia and the Pacific Islands. In South Asia, it included subclades M2, M3, M4, M5, and M6. On the other hand, East Asia is represented by subclades M7, M8, M9, M10, M11, M12, and M13. Additionally, subclades C and Z were present in Siberia and the Americas, indicating ancient migration routes through Siberia into the Americas. Haplogroup M was characterized by the mutations 10400G and 14783C. Within this haplogroup, subclade M1 was defined by the SNPs 1719G and 195C. Subclade M2 is characterized by mutations 447G and 16319A, while subclade M3 is identified by SNPs 482C and 4580A. Subclade M4 is defined by mutations 16519C and 16311C, and subclade M5 by SNPs 1888A and 16129A. Subclade M6 is characterized by mutations 3537G and 16231C, and subclade M7 by SNPs 6455T and 9824C. Subclade M8 is characterized by the mutations 4715G and 15487T, with notable descendants including subclade C, defined by SNPs 11914C and 13263A, and subclade Z, characterized by mutations 6752G and 15874C. Subclade M9 is identified by SNPs 4491A and 16362C, while subclade M10 is defined by mutations 709A and 4140T. Subclade M11 is characterized by SNPs 8108G and 13074G, subclade M12 by mutations 14569A and 14727C, and subclade M13 by SNPs 6253C and 15924G (Figure 5.5).

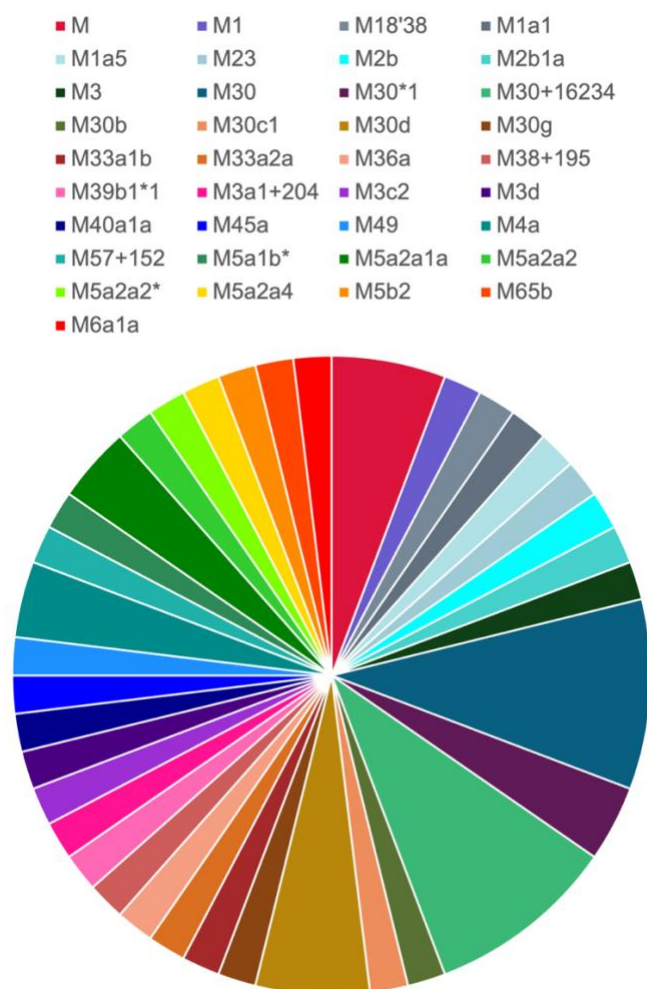


Figure 5.5. The distribution of haplogroup M illustrated in the chart. The colour-coded haplogroups are: M (bright red), M1 (violet), M1a1 (dark gray), M1a5 (light sky blue), M18'38 (medium gray), M23 (lavender), M2b (cyan), M2b1a (light cyan), M3 (dark green), M30 (light brown), M30*1 (peach), M30+16234 (teal), M30b (dark olive), M30c1 (sand brown), M30d (warm brown), M30g (chestnut brown), M33a1b (deep red-brown), M33a2a (salmon pink), M3a1+204 (light orange-brown), M36a (peach-orange), M3c2 (deep purple), M38+195 (brick red), M3d (dark purple), M40a1a (royal blue), M45a (mustard yellow), M49 (deep forest green), M4a (sea green), M5a1b* (olive green), M5a2a1a (teal green), M5a2a2* (lime green), M5a2a4 (bright yellow-green), M5b2 (golden yellow), M57+152 (spring green), M6a1a (crimson red), and M65b (orange).

Haplogroup R

This haplogroup constituted approximately 10.78% of the Emirati population, represented by 55 samples. Haplogroup R, a descendant of haplogroup N, is estimated to have originated around 55,000 years ago, giving rise to numerous sub-haplogroups that spread throughout Europe, Asia, and the Americas. Key subclades of haplogroup R include H, V, J, T, U, K, B, and F (Van Oven & Kayser, 2009). The major SNPs associated with haplogroup R are 12705T and 16223T. This study identified several subhaplogroups within haplogroup R illustrated in Figure 5.6.

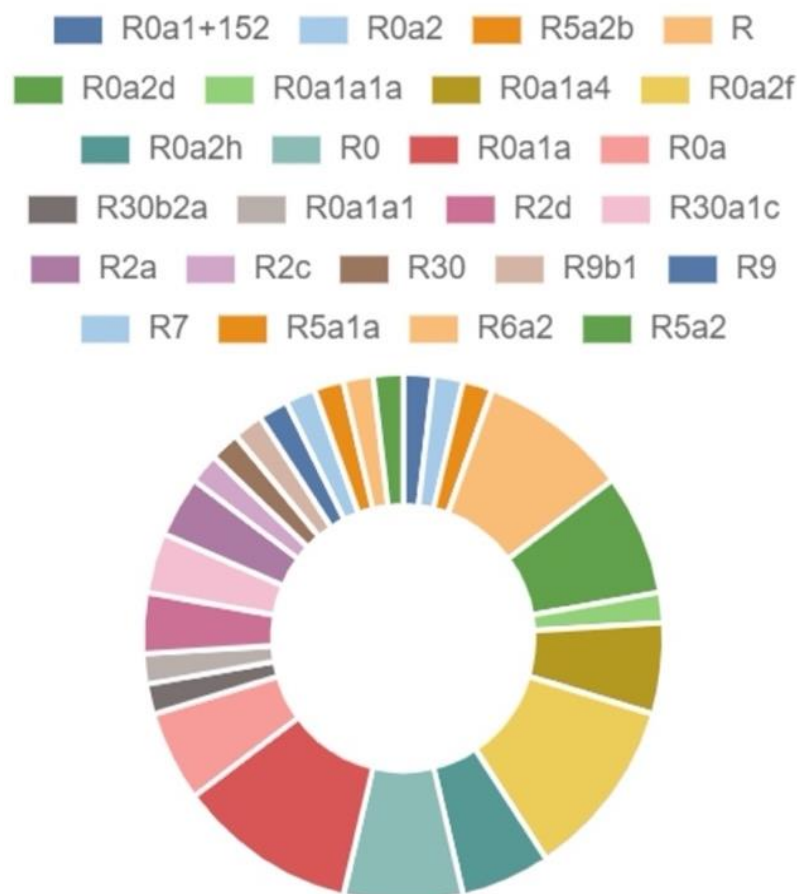


Figure 5.6. The distribution of R subhaplogroups illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: R0a1+152 (dark blue), R0a2 (light blue), R5a2b (dark orange), R (peach), R0a2d (dark green), R0a1a1a (pale green), R0a1a4 (mustard yellow), R0a2f (pale yellow), R0a2h (teal), R0 (light turquoise), R0a1a (red), R0a (light pink), R30b2a (grey), R0a1a1 (light grey), R2d (dark pink), R30a1c (light pink), R2a (lavender), R2c (pale purple), R30 (light brown), R9b1 (pale beige), R9 (royal blue), R7 (sky blue), R5a1a (orange), R6a2 (pale orange), R5a2 (dark green).

Haplogroup N

According to the results in this study, about 8.63% of the Emiratis samples (44 sample) belonged to haplogroup N. Haplogroup N as mentioned previously, is a descendent of haplogroup L3, which originated around 60,000 years ago, haplogroup N spread across Asia and Europe, giving rise to several other haplogroups. Including ancestral to many European, Asian, and Oceanian haplogroups. Key subclades of haplogroup N are N1, N2,

N9, R (Macaulay et al., 2005). It is also present in East Asian, African, and South Asian populations. The major SNPs associated with haplogroup N include 11719A, 12705T, and 16223T. Within this study, haplogroup N encompassed subhaplogroups such as N2a2, N1b1a2, N1a1a3, N1a3a, N1b1a+16129, and N1a1b1. Subclade N1 is characterized by mutations 10238C and 12501A, while N2 is defined by 189G, 709A, 5046A, 11674T, and 12414C (Figure 5.7).

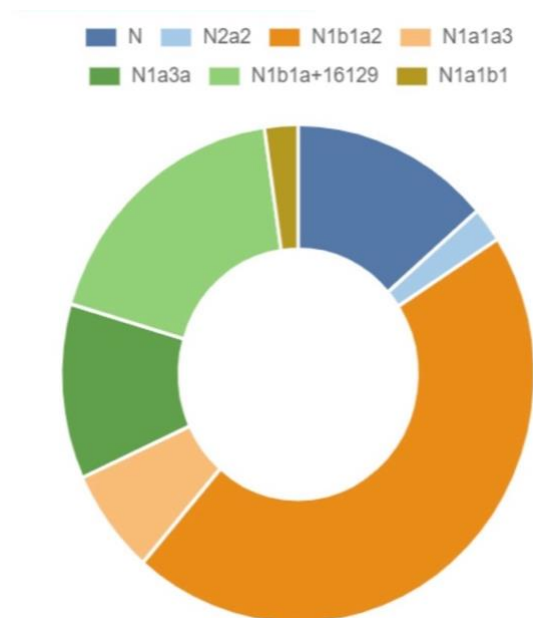


Figure 5.7. The distribution of N subhaplogroups within the population is illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: N (dark blue), N2a2 (light blue), N1b1a2 (dark orange), N1a1a3 (peach), N1a3a (dark green), N1b1a+16129 (pale green), N1a1b1 (mustard yellow).

Haplogroup T

The data from this study haplogroup T appeared in 25 samples, estimated 4.9% of the population (Figure 5.8). They are mainly found in European (non-Finnish) populations based on literature and previous research. It is also present in European (Finnish), Ashkenazi Jewish, Latino/Admixed American, South Asian, and African/African American

populations, although it is rare in some of these groups. Haplogroup T is defined by major SNPs 709A, 1888A, 4917G, 8697A, 10463C, 13368A, 14905A, 15607G, 15928A, and 16294T. Within this study, haplogroup T encompasses several subhaplogroups, each characterized by specific SNPs. These include T1a, T1a1, T1a2a, T2g, T2b, T2e, T1a1m1, T2d2, T2b4+152, and T2c1a2. Subclade T1 is characterized by mutations C12633a, A16163G, and T16189C!, while T2 is defined by A11812G, A14233G.

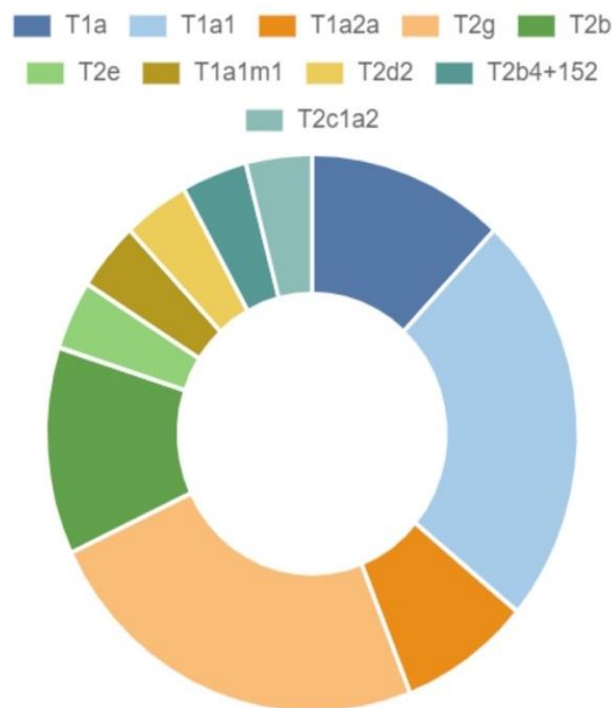


Figure 5.8. The distribution of T subhaplogroups illustrated in a pie chart, generated by Haplogrep 3. The colour-coded haplogroups are as follows: T1a (dark blue), T1a1 (light blue), T1a2a (dark orange), T2g (peach), T2b (dark green), T2e (pale green), T1a1m1 (mustard yellow), T2d2 (pale yellow), T2b4+152 (teal), T2c1a2 (light turquoise).

Haplogroup K

The results from this study, show that haplogroup K accounted for approximately 4.71% of the population, encompassing 24 samples, and predominantly observed in European

(non-Finnish) populations. It was also detected in Ashkenazi Jewish and European (Finnish) populations, with sporadic occurrences in African/African American and Latino/Admixed American populations. Haplogroup K is normally characterized by major SNPs 10550G, 11299C, 14798C, 16224C, and 16311C. Within this research, haplogroup K was shown to include several distinct subhaplogroups, each defined by specific genetic markers. These subhaplogroups encompassed K1a4c1, K1a4f, K2b1b, K1a1a1, K1a2a, K1a3a, K1a4c, K2a2a, K2a5b, and K1a4. Subclade K1 is characterized by mutations T1189C, A10398G! while K2 is defined by T146C!, T9716C (Figure 5.9).

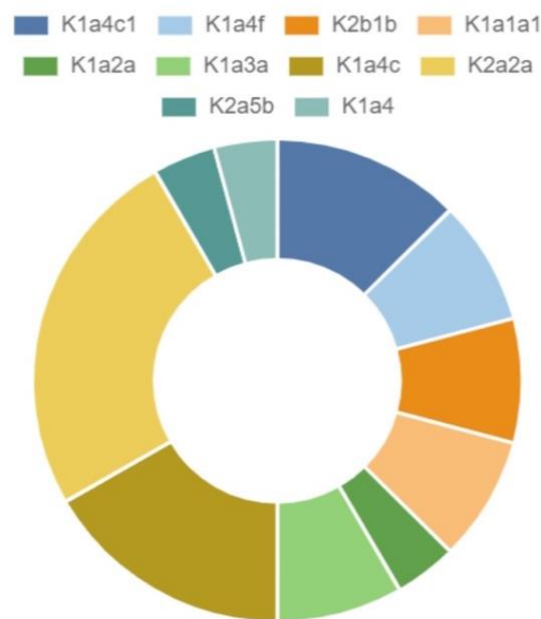


Figure 5.9. The distribution of K subhaplogroups within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: K1a4c1 (dark blue), K1a4f (light blue), K2b1b (dark orange), K1a1a1 (peach), K1a2a (dark green), K1a3a (pale green), K1a4c (mustard yellow), K2a2a (pale yellow), K2a5b (teal), K1a4 (light turquoise).

Haplogroup L

Haplogroups L is the first major split in the mtDNA tree, which is the root of all other haplogroups. It is origination from Sub-Saharan Africa, approximately 150,000 to 200,000 years ago. It represents the root of the human mtDNA tree and the origin of all other haplogroups. Haplogroup L is primarily found in Africa and further divides into several subclades. L0 is found in Southern African populations, L1, L2, and L3, Distributed across Sub-Saharan Africa (Soares et al. 2012). Rare occurrences has been recorded in Latino/Admixed American and European (non-Finnish) populations. They are represented in this study as follows: Haplogroup L0 makes up 1.57% of the study population found in 8 samples (Figure 5.10). It was identified by SNPs 263G!, 1048T, 3516A, 5442C, 6185C, 9042T, 9347G, 10589A, 12007A, and 12720G. Within this study, haplogroup L0 encompassed several subhaplogroups, including L0d2c1a, L0a2a2a, and L0a1a2. Haplogroup L1 constituted 1.76% in this study population set and it was found in 9 samples and identified by SNPs 3666A, 7055G, 7389C, 13789C, 14178C, and 14560A. Haplogroup L1 included diverse subhaplogroups, L1b1a8, L1c2a1a, L1c2b2, and L1b1a3. Haplogroup L2 represented 2.35% of the study population found in 12 samples. Subhaplogroups including L2b1a, L2a1h, L2a1b1, and L2a1+143 wered optimized. Haplogroup L3 makes up 4.12% of Emiratis samples found in 12 samples and is identified by SNPs 8701G, 9540C, 10398G, 10873C, and 15301A. Within the study, L3e2b1a2, L3e1e1, L3e2b, L3e1b2, L3f1b4a1, L3f1b4a, L3d1a1a, L3b1a1a, L3h2, and L3b1a+152 subhaplogroups were identified.

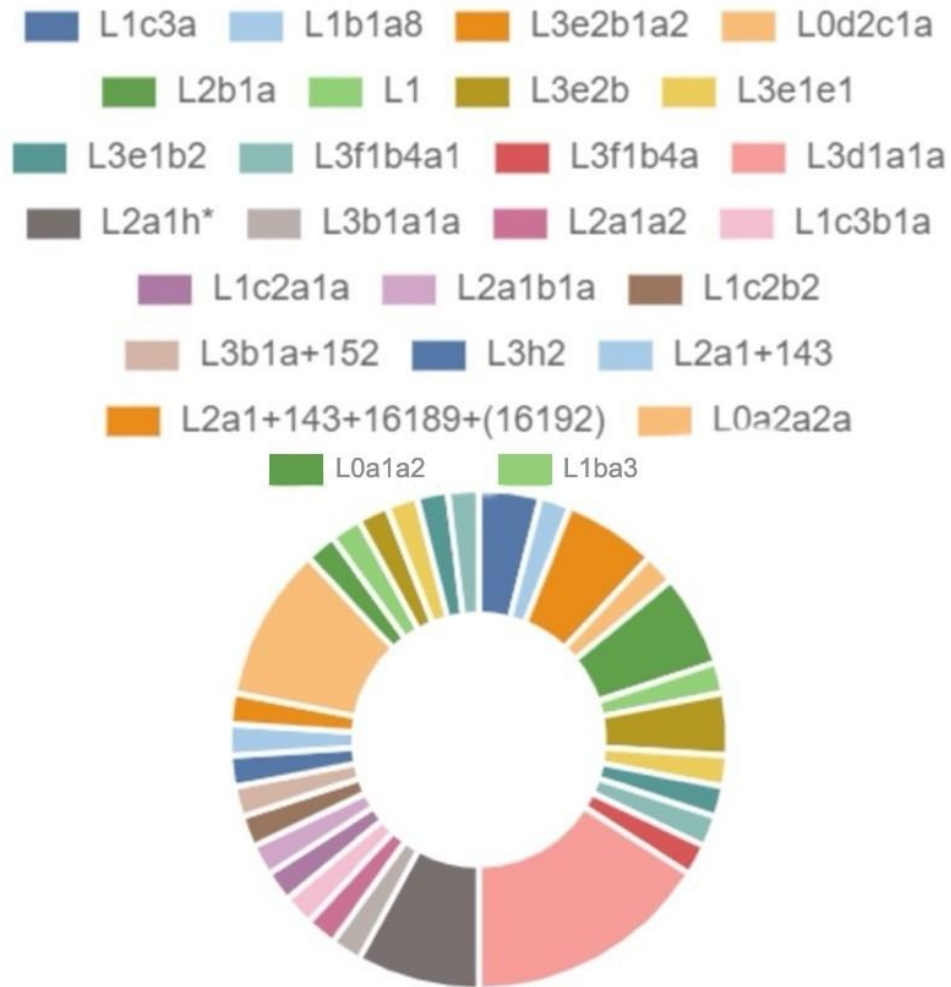


Figure 5.10. The distribution of subhaplogroups of macrohaplogroup L within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: L1c3a (dark blue), L1b1a8 (light blue), L3e2b1a2 (dark orange), L0d2c1a (peach), L2b1a (dark green), L1 (pale green), L3e2b (mustard yellow), L3e1e1 (pale yellow), L3e1b2 (teal), L3f1b4a1 (light turquoise), L3f1b4a (red), L3d1a1a (light pink), L2a1h* (grey), L3b1a1a (light grey), L2a1a2 (dark pink), L1c3b1a (light pink), L1c2a1a (lavender), L2a1b1a (pale purple), L1c2b2 (light brown), L3b1a+152 (pale beige), L3h2 (royal blue), L2a1+143 (sky blue), L2a1+143+16189+16192 (orange), L0a2a2a (pale orange), L0a1a2 (dark green), L1ba3 (yellow green).

Haplogroup X

In this study, haplogroup X accounted for about 1.76 % of the population in 9 samples (Figure 5.11). In this Haplogroup, it is believed to have originated around 30,000 years ago in the Near East or West Asia and belongs to the macrohaplogroup N. It is notable for its wide but uneven distribution across Europe, the Near East, North Africa, and indigenous American populations. Haplogroup X includes two primary subclades: X1, found mainly in North and East Africa and the Near East, and X2, present in Europe, the Near East, the Caucasus, and Native American groups (Reidla et al., 2003). Subclade X2 is further divided into X2a, X2b, X2c, X2d, X2e, X2f, and X2g. Key SNPs defining Haplogroup X are 14470A and 6221C, with subclade X1 characterized by SNPs 146C and 15654C, and subclade X2 by SNPs 195C, 1719A, and 1438G. Although rare, Haplogroup X is significant due to its unique distribution and ancient origins. Its presence in Europe and the Near East indicates ancient migration and population admixture, while X1's prevalence in North and East Africa suggests ancient maternal lineages. X2a's occurrence in Native American populations supports early migrations from Siberia across the Bering land bridge. The unique distribution of Haplogroup X in both the Old and New Worlds is crucial for studying prehistoric human dispersal and genetic interactions, enhancing our understanding of genetic diversity and historical connections among modern populations.

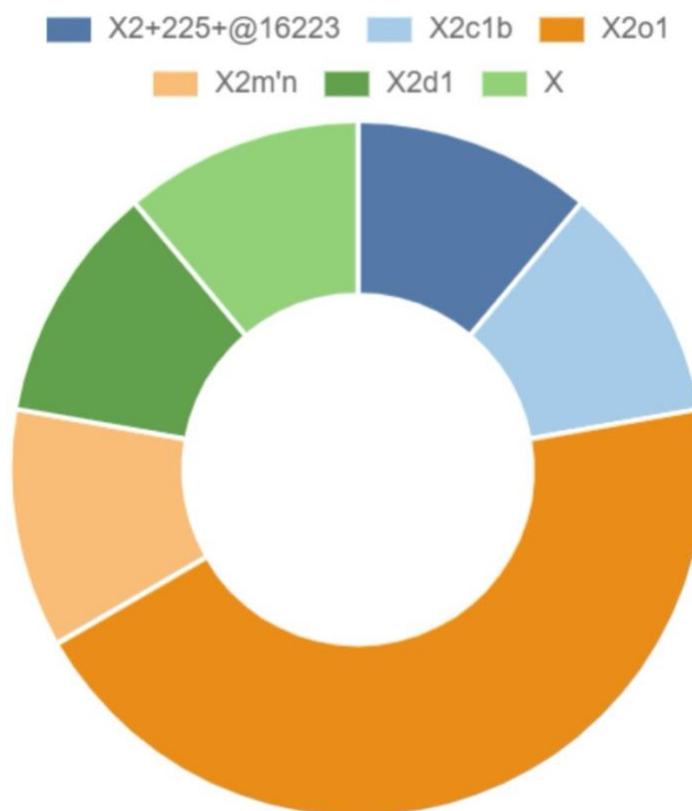


Figure 5.11. The distribution of subhaplogroups of macrohaplogroup X within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: X2+225+@16223 (dark blue), X2c1b (light blue), X2o1 (dark orange), X2m'n (peach), X2d1 (dark green), X (pale green).

Haplogroup I

In this study, haplogroup I represented roughly 1.37% of the population in 7 samples and it was primarily found in European (non-Finnish) populations. It is also present in European (Finnish) populations, with rare occurrences in African/African American, Ashkenazi Jewish, and Latino/Admixed American populations. Haplogroup I is defined by major SNPs 10034C and 16129A. Within this study, haplogroup I encompasses several subhaplogroups, each characterized by specific SNPs. These include I1a1, I6, I*, I1, I1c1, and I5a3 (Figure 5.12).

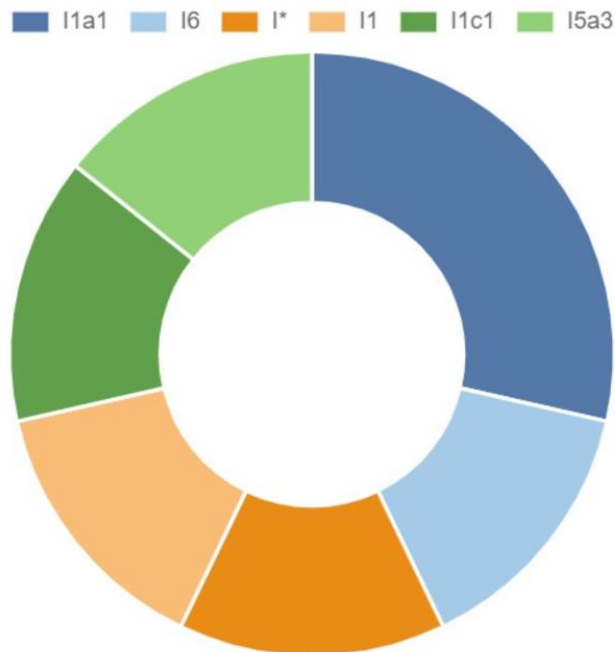


Figure 5.12. The distribution of subhaplogroups of macrohaplogroup I within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: I1a1 (dark blue), I6 (light blue), I* (dark orange), I1 (peach), I1c1 (dark green), I5a3 (pale green).

Haplogroup W

In this study, haplogroup W comprised nearly 1.37% of the population, totaling 7 samples (Figure 5.13), and it was predominantly found in European (non-Finnish) populations. It was also detected in Ashkenazi Jewish populations, with infrequent occurrences noted in African, South Asian, and Latino/Admixed American populations. Haplogroup W is characterized by major SNPs within the study dataset. This haplogroup encompassed several distinct subhaplogroups, each delineated by specific genetic markers. These include W3a1, W6a, W1, W+194, W1+119, W6b1, and W+194. Subclade W1 is characterized by 7864T, W3: 1406C, W6: 4093G, 8614C, 16325C.

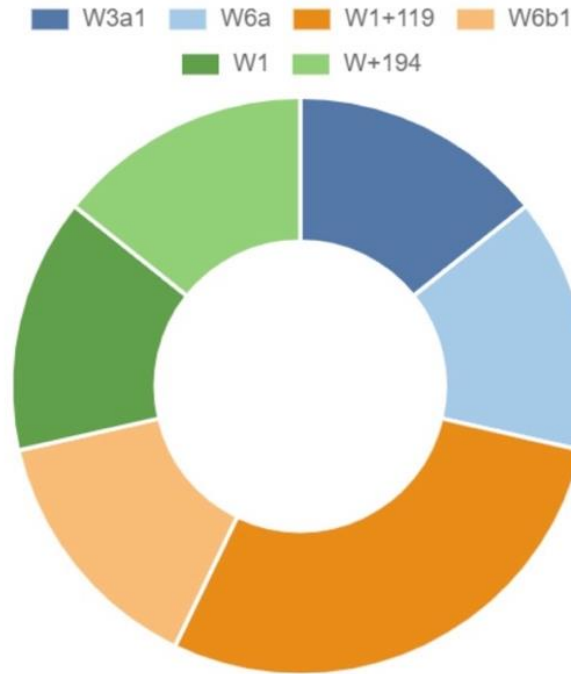


Figure 5.13. The distribution of subhaplogroups of macrohaplogroup W within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: W3a1 (dark blue), W6a (light blue), W1+119 (dark orange), W6b1 (peach), W1 (dark green), W+194 (pale green).

Haplogroup B

In this study, haplogroup B was identified in 2 samples, constituting approximately 0.39% of the population (Figure 5.14). It was estimated to have originated around 50,000 years ago, haplogroup B is primarily observed in Latino and Admixed American populations. East Asian is the origins of B, and also shows presence in African/African American, and, more rarely, European (non-Finnish) populations. Major SNPs associated with haplogroup B included 16189, 16217, and 8281-8289d. Within the scope of this research, haplogroup B encompassed several subhaplogroups, including B5b1c and B4a1a. B4 is characterized by 16217C SNP while B5 is characterized by 709A, 8584A, 9950C, 10398G, 16140C.

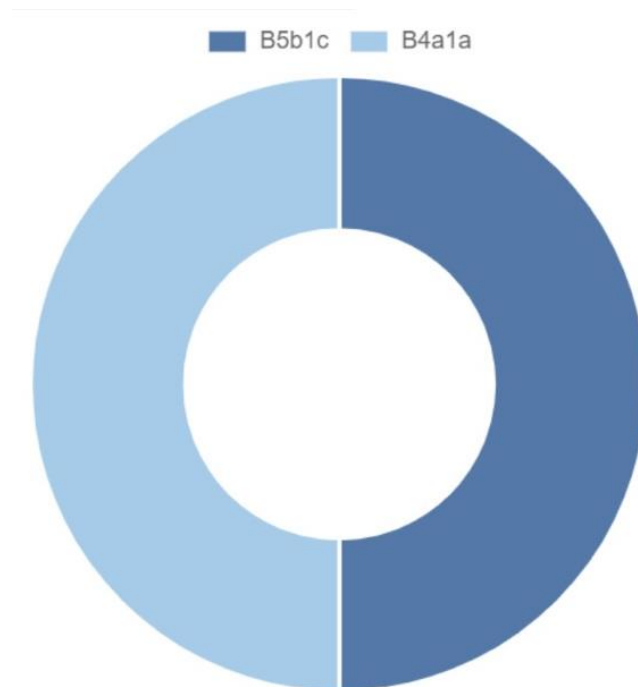


Figure 5.14. The distribution of subhaplogroups of macrohaplogroup B within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: B5b1c (dark blue), B4a1a (light blue).

Haplogroup D

In this study, haplogroup D comprised 0.39% of the population, represented by 2 samples (Figure 5.15). It was estimated to have originated around 48,000 years ago, haplogroup D is predominantly found in East Asian populations. It was also manifested in Latino American, African, Amish, South Asian, and, less frequently, European (non-Finnish) populations. Major SNPs associated with haplogroup D included 5178, 4883, and 16362. Within the context of this research, haplogroup D encompassed two subhaplogroups, D4j and D4i. Subclade D4 is characterized by SNPs in 3010A, 8414T, 14668T.

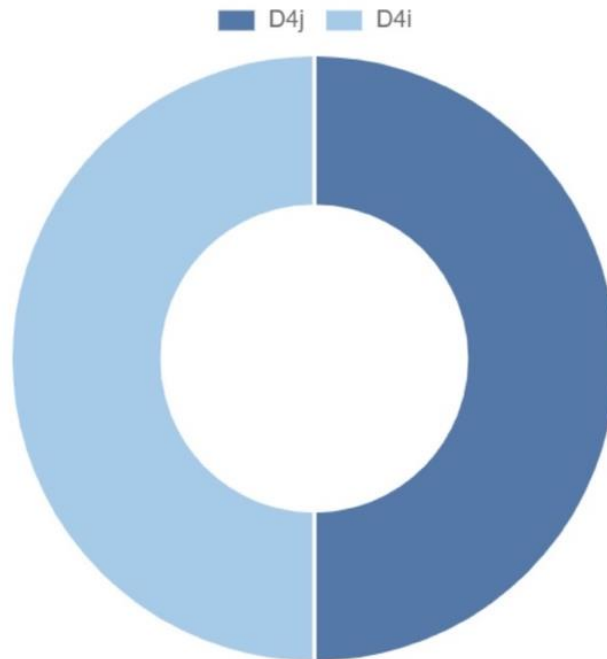


Figure 5.15. The distribution of subhaplogroups of macrohaplogroup D within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: D4j (dark blue), D4i (light blue).

Haplogroup F

In this study, haplogroup F constituted 0.39% of the population in 2 samples (Figure 5.16). Haplogroup F was estimated to have originated around 40,000 years ago and is primarily found in East Asian populations. It also appeared in African, South Asian, European, and, more rarely, Latino American regions. The major SNPs associated with haplogroup F included 16304, 249d, 6392 and 10310. Within this study, haplogroup F encompassed two subhaplogroups F1a1a, and F3b1b1. F1 is characterized by SNPs in 6962A, 10609C, 12406A, 12882T, while F3 is by 3434G, 5585A, 5913A, 5978G, 10320A, 11065G, 16298C, 16304T!, 16362C.

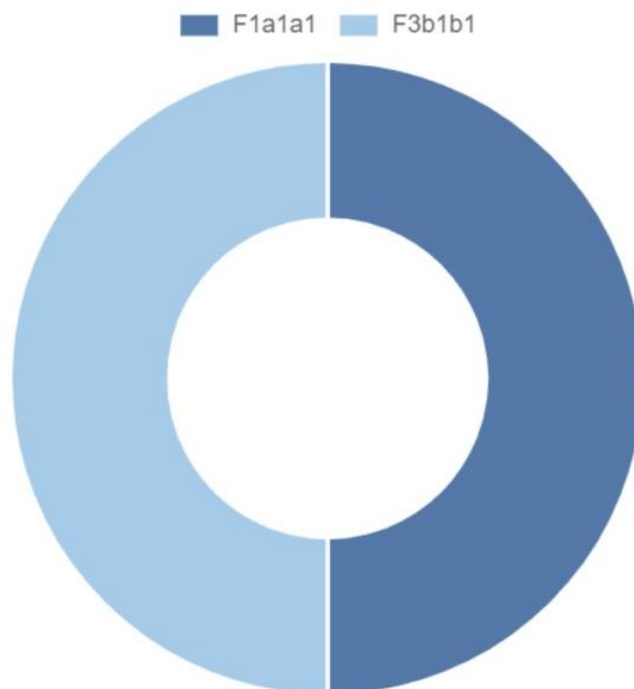


Figure 5.16. The distribution of subhaplogroups of macrohaplogroup F within the population is illustrated in the accompanying chart, generated using Haplogrep 3 software. The colour-coded haplogroups are as follows: F1a1a1 (dark blue), F3b1b1 (light blue).

The results of the alignments, haplogroups observed in this study included H, HV, U, J, M, R, N, T, K, L, X, I, W, B, D and F haplogroups. Due to the history of the migration movement in the Arabian Peninsula, such haplogroups were expected to be present in the Emirati population. A single sample representing H haplogroup was observed on EMPOP map alignment tool and the retained results are shown in Figure 5.17.

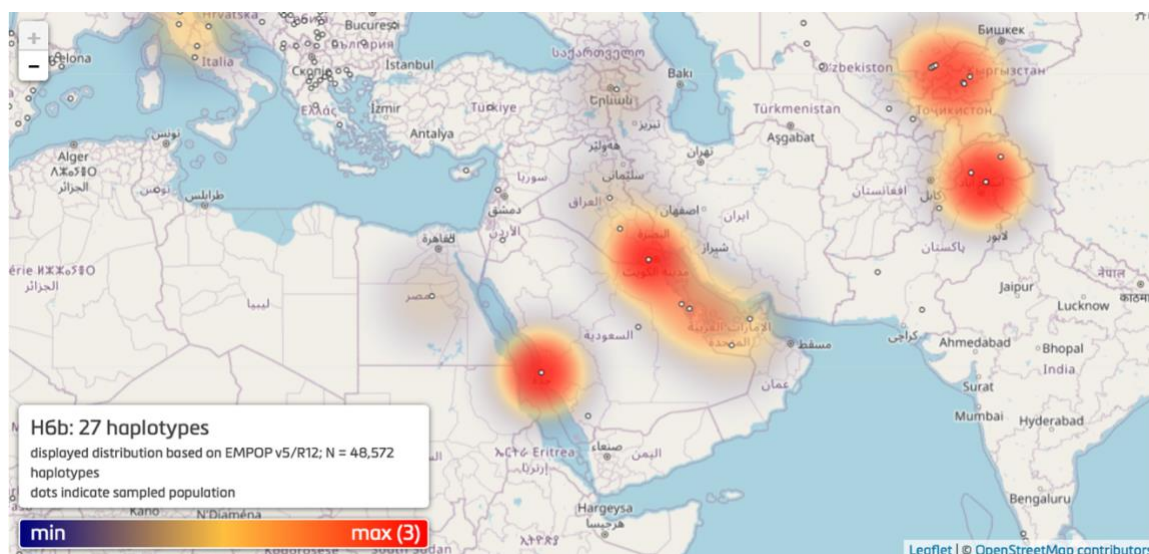


Figure 5.17. An online alignment and haplogroup assignment using EMPOP platform of the haplogroup H6b as a heat map.

As a result, 510 samples 214 haplogroups were present from 16 macro haplogroups. Those results were compared with two studies from Aljasmi et al., 2020, and Alshamali et al., 2008. The distribution and frequency of mtDNA haplogroups were evaluated, to illuminate patterns of maternal lineage, infer historical migration events, and delineate demographic histories shaping these populations' genetic structures. Figure 5.18 represented the 3 studies, haplogroup labels on the x-axis, colour coding to the bars with each colour representing the frequency data from each study. Each coloured bar represents the percentage frequency of a haplogroup within the respective population study. Blue for Aljasmi et al., 2020, red for the current study, and green for Alshamali et al., 2008.

The comparative analysis revealed noteworthy insights into the mtDNA haplogroup distribution across the studies. These findings significantly contribute to our understanding of genetic structure and historical migrations within these populations,

highlighting the complex nature of human evolutionary history as reflected through maternal lineages.

There are variations in the frequencies of specific haplogroups among the populations. For instance, haplogroup H is most prevalent in the current study at approximately 15.49%, compared to Aljasmi et al., 2020 and Alshamali et al., 2008, where it is found at 8.18% and 9.64%, respectively.

Common haplogroups appeared with notable frequencies across all datasets. Their prevalence reflects historical gene flow and shared ancestry, providing insights into past human migrations and demographic events that have shaped genetic diversity. Haplogroup U appeared with high frequencies in the three studies.

The commonality of certain haplogroups across diverse geographical and cultural landscapes underscores a complex tapestry of human history, marked by migrations, expansions, and sometimes isolations. These patterns are crucial for both anthropological genetics and for practical applications in fields like forensic science, where understanding the frequency of haplogroups can aid in identity resolution and ancestry reconstruction.

In this study, the analysis has identified several rare mtDNA haplogroups whose frequencies are consistently low across the three datasets. These haplogroups are of particular interest as they might indicate unique evolutionary paths or restricted migration events that have only affected limited population subsets. Here, "rare" haplogroups are defined as those with a mean frequency of less than 1% across the studies.

Haplogroup B: Exhibits a mean frequency of approximately 0.13%. It is absent in the Aljasmi et al. (2020) and Alshamali et al. (2008) datasets, but appears minimally in the current dataset.

Haplogroup D: Similar to haplogroup B, this haplogroup shows a mean frequency of around 0.13% and is only present in the current study, suggesting a very limited representation.

Haplogroup E: This haplogroup is slightly more frequent than the previous two, with a mean frequency of 0.29%, but it is entirely absent from this study and Alshamali et al. (2008) datasets, only appearing in Aljasmi et al. (2020).

Haplogroup F: Present in all datasets with a mean frequency of about 0.41%, indicating a somewhat broader but still limited distribution compared to other haplogroups.

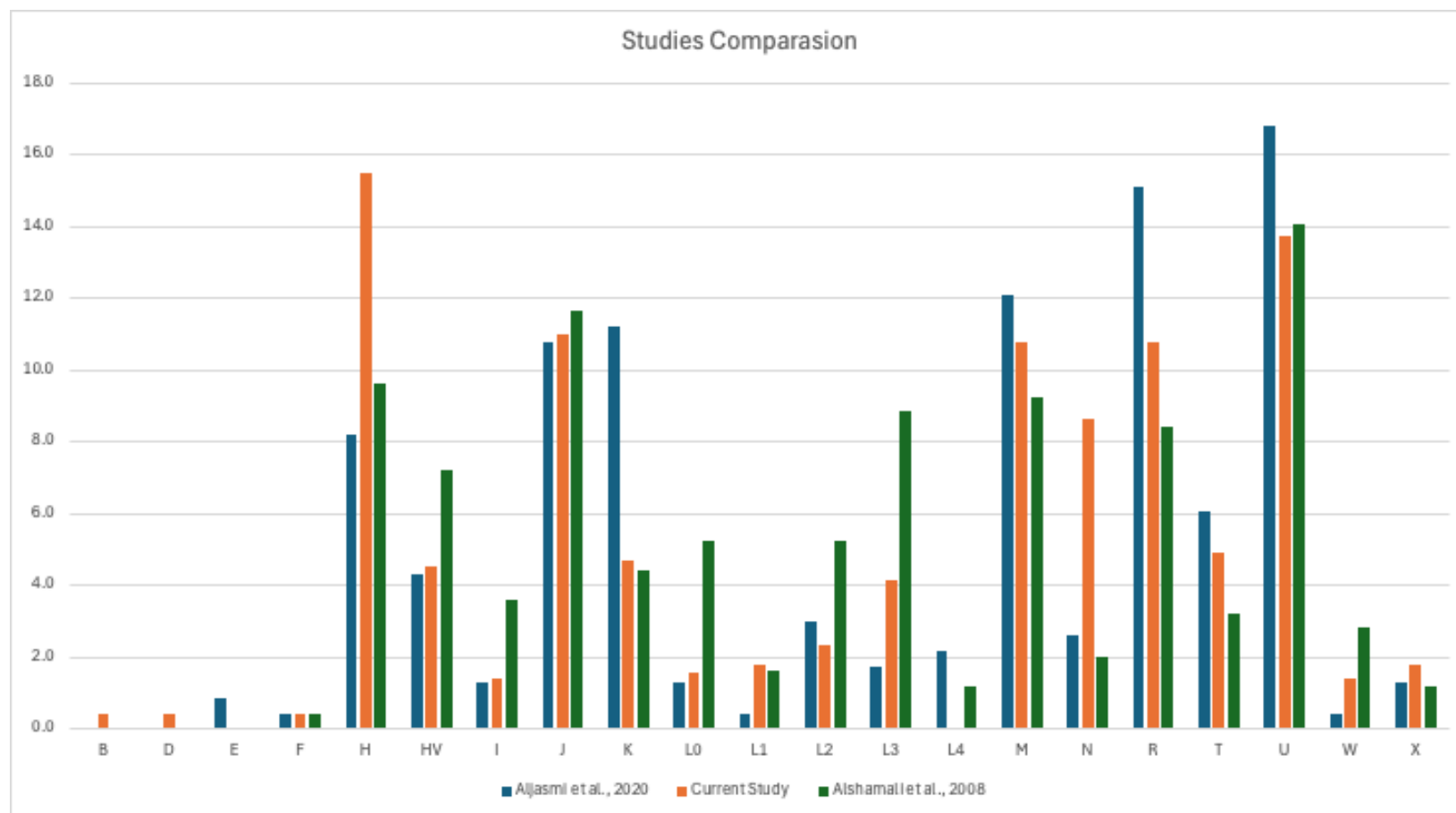


Figure 5.18. Comparative graph of haplogroup percentages (B, D, F, H, HV, I, J, K, L0, L1, L2, L3, L4, M, N, R, T, U, W, X) between three studies on the UAE population. The studies include Aljasmil et al. (2020) with n=232 samples (blue), Alshamali et al. (2008) with n=249 samples (green), and the current study with n=510 samples (orange).

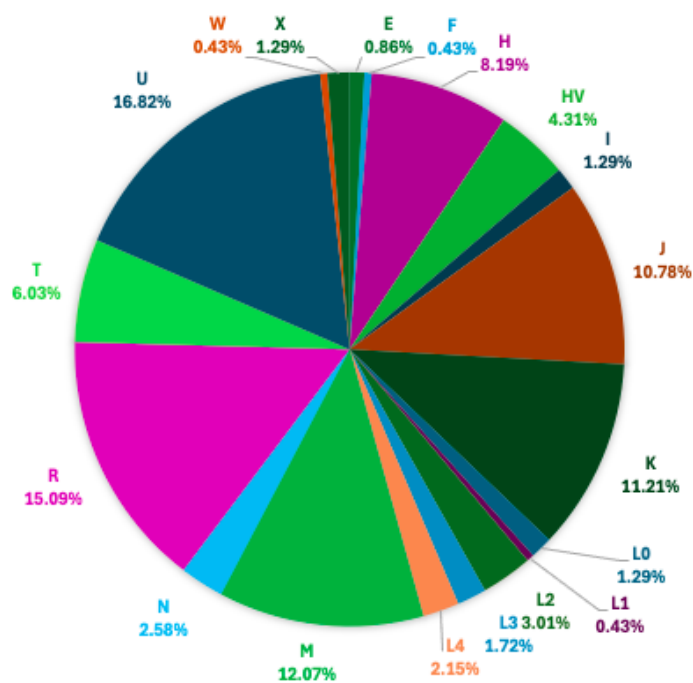


Figure 5.19. Whole mtDNA pie chart representation of the haplogroups frequencies identified in the study conducted by Aljasmí et al., 2020 (n=232). (Chapter 1)

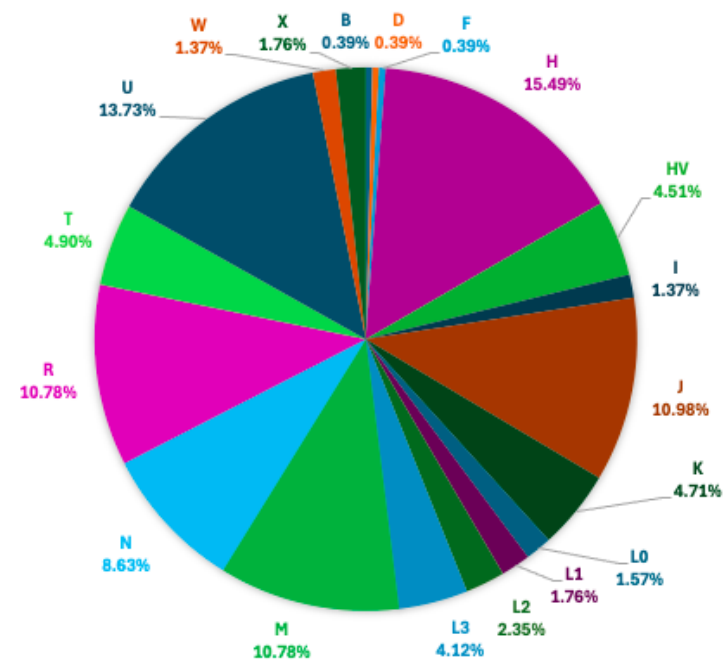


Figure 5.20. Whole mtDNA pie chart representation of the haplogroups frequencies identified by the current study conducted (n=510).

5.6.Results: Section 2

5.7.MPS Data Analysis: Indians Samples Set

Haplotype analysis of 50 Indian samples using MPS revealed that each sample presented a distinct combination of SNP variations, including insertions and deletions in heteroplasmic regions, which characterized individual haplotypes. Some haplotypes were common across multiple samples, suggesting shared genetic ancestry within the population, while others were unique, reflecting rare genetic variations. These SNP variations were instrumental in determining the haplogroup for each sample.

5.7.1. Haplotypes generating

Table 5.3 summarized the 30 samples used to compare Sanger sequencing and MPS, while the remaining samples are detailed in the Appendices.

Table 5.3. Whole mtDNA genome sequences for 30 indian samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing are indicated in italics, and insertions/deletions (indels) are highlighted in red.

No.	Samples	Haplogroups	Haplotypes
1	IND_01	M5a2a	73G 236C 263G 315.1C 489C 709A 742C 750G 1438G 1888A 2706G 2833G 3921T 4454C 4769G 4907C 6062T 6293C 6378C 7028T 8158G 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16223T 16519C
2	IND_02	R8b1	73G 150T 154C 195C 263G 315.1C 455.1T 456T 750G 1438G 1709A 2706G 2746C 2755G 3384G 4769G 5096C 6485G 7028T 7759C 8860G 9449T 9758C 11719A 12007A 13194A 13215C 14766T 15250T 15326G 16172C 16390A 16519C 73G 143A 199C 204C 250C 263G 297G 315.1C 573.CC 710C 750G 1438G 1719A 2706G 4529T 4769G 4947C 6671C
3	IND_03	N1a1b	7028T 8251A 8860G 9804A 10238C 10398G 10790C 10882C 11719A 12501A 12705T 13780G 14766T 15043A 15326G 15924G 15954G 16223T 16311C 16519C
4	IND_04	HV2a2	72C 73G 152C 195C 263G 315.1C 523del 524del 709A 750G 1438G 2706G 4769G 5153G 6366A 7028T 7193C 7861C 8860G 9336G 10306G 11935C 12061T 15326G 16217C
5	IND_05	R30b2a	73G 152C 263G 315.1C 373G 750G 1438G 2706G 2831A 4769G 6290T 7028T 7280T 7843G 8584A 8860G 11719A 11916A 12028C 13539G 14000A 14766T 15148A 15326G 16497G 16519C 16524G 73G 152C 263G 315.1C 523del 524del 750G 980C 1438G 1811G 2706G 3741T 4769G 4851T 5360T 7028T 8137T
6	IND_06	U7b	8684T 8860G 10053T 10084C 10142T 11467G 11719A 12308G 12372A 13500C 14569A 14766T 14971C 15326G 16309G 16318T 16519C
7	IND_07	U2a1a	73G 263G 315.1C 750G 1438G 1811G 2706G 4769G 7028T 8572A 11368C 11467G 11719A 12308G 12372A 13708A 14766T 15326G 16051G 16093G 16154C 16206C 16230G 16311C

8	IND_08	U2c1	73G 152C 239C 247A 263G 315.1C 750G 1438G 1811G 2706G 3915A 4769G 5790A 7028T 8023C 8676T 8860G 9692G 9767T 11467G 11719A 12308G 12361G 12372A 13368A 14766T 14935C 15061G 15326G 16051G 16129A 16179T 16234T 16247G 16519C
9	IND_09	U9a1	73G 200G 263G 315.1C 499A 750G 1438G 1811G 2706G 3290C 3531A 3834A 4769G 5351G 5999C 6386T 7028T 8860G 11467G 11719A 12308G 12372A 14094C 14766T 15077A 15326G 16051G 16193T 16234T 16278T 16357C
10	IND_10	M3a2	73G 263G 315.1C 482C 489C 750G 1438G 2706G 4580A 4769G 5783A 6359G 7028T 8701G 8860G 8950A 9540C 10398G 10400T 10727T 10873C 11719A 12705T 14766T 14783C 15043A 15169G 15301A 15326G 16051G 16126C 16223T 16278T 16519C
11	IND_11	M5a	73G 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4721G 4769G 7028T 8701G 8860G 9540C 9773T 9947A 10398G 10400T 10873C 11719A 12477C 12705T 13708A 14323A 14766T 14783C 15043A 15301A 15326G 15927A 16129A 16223T 16519C
12	IND_12	U7a3a	73G 151T 152C 263G 315.1C 523del 524del 750G 824C 980C 1438G 1811G 2706G 2863C 3741T 4769G 5360T 6620C 7028T 8137T 8684T 8860G 9852G 10142T 11467G 11719A 12308G 12372A 12618A 13500C 14569A 14766T 15326G 16069T 16274A 16318T 16519C
13	IND_13	A17	73G 152C 234G 235G 263G 315.1C 523del 524del 663G 750G 1438G 1736G 2706G 4113A 4248C 4592C 4769G 4824G 5147A 5514G 7028T 8794T 8860G 9126C 11719A 12705T 14766T 15217A 15326G 16172C 16173T 16223T 16235G 16290T 16311C 16319A 16362C 16519C
14	IND_14	M2a'b	73G 143A 195C 263G 315.1C 337G 447G 489C 750G 1007A 1438G 1780C 2706G 3432T 4769G 6647G 7028T 7337A 8212T 8502G 8701G 8860G 9540C 9899C 10398G 10400T 10873C 11083G 11518A 11719A 12705T 14766T 14783C 14861A 15043A 15253G 15301A 15326G 15670C 15721C 16223T 16274A 16319A 16320T 16518T 16519C
15	IND_15	M5a	73G 263G 315.CC 489C 709A 750G 1438G 1888A 2706G 3921T 4314A 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16129A 16223T 16519C

16	IND_16	U2a2	73G 263G 315.1C 750G 1095C 1438G 1760A 1811G 2393T 2706G 3316A 4769G 4970G 5021C 5201C 6116G 7028T 7859A 8860G 11299C 11467G 11719A 12308G 12372A 12477C 12561A 14182C 14766T 14883T 15326G 16051G 16206C
17	IND_17	H	73G 263G 315.1C 750G 1438G 4769G 5120G 5216T 8860G 15326G 15439T 16051G 16126C 16519C
18	IND_18	M36a	73G 152C 239C 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3783T 4769G 6320C 6917A 7028T 7271G 8701G 8860G 9540C 10398G 10400T 10873C 11608G 11719A 12346T 12705T 13831T 14203G 14302C 14766T 14783C 15043A 15110A 15301A 15326G 16193T 16223T 16519C
19	IND_19	U2a1a	73G 151T 263G 315.1C 750G 1438G 1811G 2706G 4769G 7028T 8572A 11368C 11467G 11719A 12308G 12372A 13708A 14766T 15326G 16051G 16093G 16154C 16206C 16230G 16311C
20	IND_20	M6b	73G 146C 152C 263G 315.1C 461T 489C 523del 524del 750G 1438G 2706G 3254A 3444T 4216C 4417G 4769G 5301G 5558G 7028T 8281del 8282del 8283del 8284del 8285del 8286del 8287del 8288del 8289del 8701G 8860G 9540C 10321C 10398G 10400T 10640C 10667C 10873C 11719A 12634G 12705T 13161C 14128G 14696G 14766T 14783C 15043A 15301A 15326G 16184T 16223T 16256G 16311C 16362C
21	IND_21	M39b1	56del 58A 65.1T 66T 73G 153G 263G 315.1C 463T 485C 489C 750G 1438G 1811G 2706G 3531A 4769G 6257A 7028T 8567C 8679G 8701G 8860G 9374G 9540C 9655A 10398G 10400T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 15938T 16223T
22	IND_22	R6a2	73G 263G 315.1C 750G 1438G 2706G 4769G 5021C 5894G 7028T 7897A 8860G 11075C 11719A 12133T 12285C 14058T 14766T 15067C 15326G 16129A 16158G 16213A 16362C 16519C
23	IND_23	W3a1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1406C 1438G 2706G 3505G 4370C 4769G 5046A 5460A 7028T 8251A 8860G 8994A 11674T 11719A 11947G 12414C 12705T 13263G 14766T 15326G 15784C 15884C 16129A 16223T 16292T 16519C

24	IND_24	M2a1	73G 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 5147C 5252A 6752G 7028T 7961C 8396G 8502G 8701G 8853G 8860G 9540C 9758C 10398G 10400T 10873C 11016A 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16172C 16223T 16224C 16270T 16274A 16319A 16352C 16519C 16524C
25	IND_25	M44a	73G 146C 263G 315.1C 489C 750G 930A 961C 1438G 2706G 4769G 7028T 8179G 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 16223T 16301T
26	IND_26	M65a+@16311	73G 263G 315.1C 489C 511T 750G 1438G 2706G 4769G 4916G 7028T 8047C 8701G 8860G 9254G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 13651G 14766T 14783C 15043A 15301A 15326G 16223T 16289G 16519C
27	IND_27	M2a1	73G 183G 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 5252A 7028T 7961C 8396G 8502G 8701G 8860G 9095C 9540C 9758C 9779T 10398G 10400T 10873C 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16223T 16270T 16274A 16319A 16352C 16519C
28	IND_28	N1a1b1	73G 143A 199C 204C 250C 263G 315.1C 710C 750G 1438G 1719A 2706G 4529T 4769G 4947C 6671C 7028T 8251A 8860G 9804A 10238C 10398G 10790C 10882C 11719A 12501A 12705T 13780G 14766T 15043A 15326G 15924G 15954G 16223T 16311C 16519C
29	IND_29	M2a1a	73G 195C 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 4965G 5252A 7028T 7604A 7961C 8396G 8502G 8701G 8860G 9540C 9758C 9965C 10398G 10400T 10873C 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16223T 16270T 16288C 16319A 16352C 16519C
30	IND_30	M30c1	73G 146C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 7028T 8251A 8701G 8860G 9540C 9797C 10398G 10400T 10873C 11719A 12007A 12234G 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16166del 16223T 16519C

Following the same evaluation procedure outlined earlier, the Indian haplotypes were analyzed, and the results are presented in Table 5.4. This table reflects the unique genetic characteristics observed in the Indian population.

Table 5.4. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for Indian samples.

Set	Ht	Hd	PD	PI
Indians (n=50)	49	0.9992	0.9792	0.0208

5.7.2. Haplogroup Assignments

Haplogroup assignment for this dataset followed the same process of classifying mtDNA sequences into distinct haplogroups based on specific variations. The haplogroup results were successfully recorded for all the samples in the dataset, providing valuable insight of the population (Figures 5.21 and 5.22)

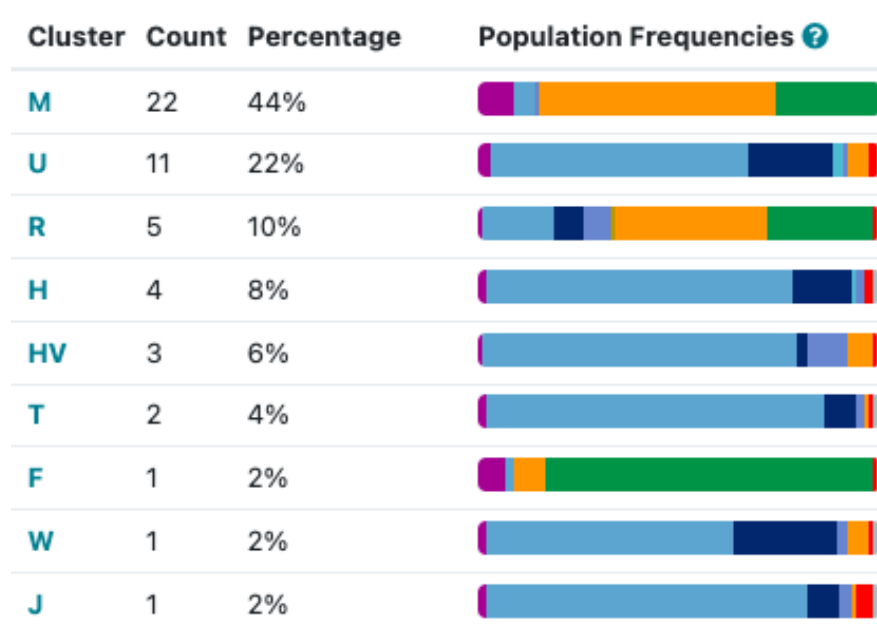


Figure 5.21. An overview of 9 haplogroups in 50 samples of the Indians population, detailing sample counts, frequencies, and percentages.

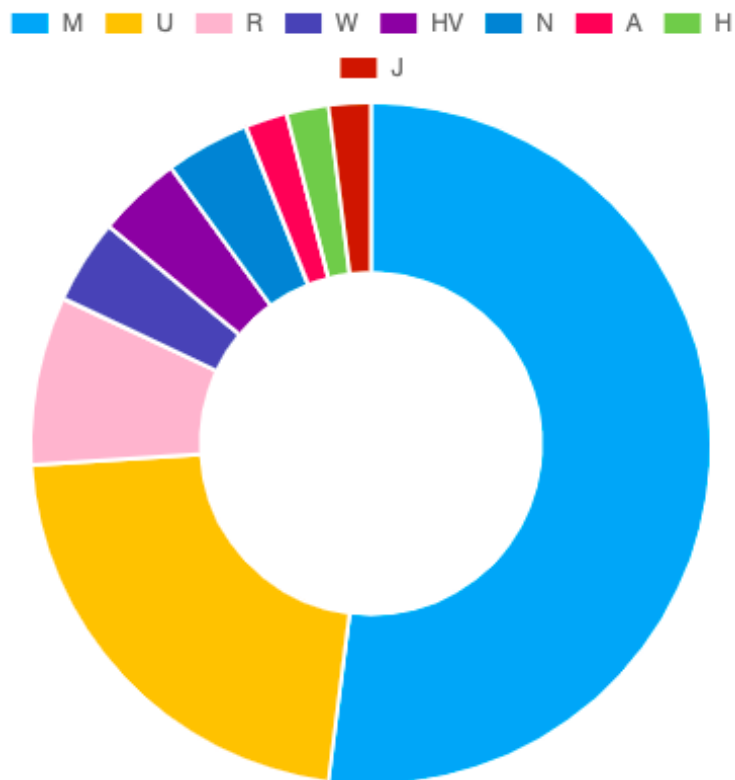


Figure 5.22. Whole mtDNA pie chart representation of the Indians haplogroups frequencies identified by the current study conducted.

5.8. Results: Section 3

5.9. MPS Data Analysis: Pakistanis Samples Set

The haplotype analysis of 50 Pakistani samples using MPS revealed that each sample exhibited a unique combination of SNP variations, including insertions and deletions in heteroplasmic regions, which defined individual haplotypes. While some haplotypes were observed in multiple samples, indicating shared genetic ancestry within the population, others were rare.

Table 5.5 provided a summary of the 30 samples used for comparing Sanger sequencing with MPS, while the remaining samples are included in the Appendices.

Table 5.5. Whole mtDNA genome sequences for 30 Pakistani samples that were initially sequenced using Sanger sequencing. Positions that overlap with the CR Sanger sequencing are indicated in italics, and insertions/deletions (indels) are highlighted in red.

No.	Samples	Haplogroups	Haplotypes
1	PAK_01	U2a2	73G 194T 263G 315.1C 750G 1438G 1811G 2706G 3316A 4769G 4970G 5201C 6116G 7028T 7257G 7859A 8860G 10355A 11299C 11467G 11719A 11932T 12308G 12372A 12477C 12561A 14182C 14766T 14883T 15326G 16051G 16206C 16271C
2	PAK_02	M4	73G 249del 263G 315.1C 489C 750G 1438G 2706G 4769G 6620C 7028T 7673G 7859A 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16145A 16223T 16261T 16311C 16519C
3	PAK_03	M5c1	73G 150T 263G 294C 315.1C 333C 489C 575T 750G 1438G 1888A 2706G 3144G 4769G 4851T 5319G 5417A 6413C 7028T 8701G 8860G 9540C 9632G 10398G 10400T 10873C 11719A 12705T 13708A 14766T 14783C 15043A 15301A 15326G 16129A 16218T 16223T 16519C
4	PAK_04	M18a	73G 93G 194T 246C 315.1C 489C 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12498T 12705T 13135A 14766T 14783C 15043A 15301A 15326G 16223T 16318T 16519C
5	PAK_05	M30	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3338C 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12172G 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16111T 16223T 16320T 16399G 16519C
6	PAK_06	M32'56	73G 146C 263G 315.1C 461T 489C 523del 524del 750G 2706G 3486T 3537G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12507G 12705T 14766T 14783C 15043A 15301A 15326G 16223T 16356C 16362C

7	PAK_07	R30a1c	73G 263G 315.1C 524.ACACAC 750G 1438G 2056A 2706G 3316A 4232C 4769G 5442C 6764A 7028T 8584A 8860G 9142A 9156G 9242G 9869G 11047G 11719A 12714C 13161C 13773G 14766T 150550C 15326G 16172C 16278T 16519C
8	PAK_08	M5a	73G 146C 154C 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10589A 10873C 11719A 12477C 12681C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16223T 16519C
9	PAK_09	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16092C 16519C
10	PAK_10	M3d	73G 146C 263G 315.1C 482C 489C 524.AC 750G 1438G 2706G 4769G 6305A 7028T 8701G 8860G 9266A 9540C 10398G 10400T 10873C 10954T 11150A 11719A 11827C 12705T 12873C 14758T 14766T 14783C 15043A 15301A 15326G 16126C 16172C 16223T 16344T 16519C
11	PAK_11	U2c1a	73G 146C 152C 263G 315.1C 523del 524del 750G 1438G 1811G 2706G 4769G 5790A 6320C 7028T 8023C 8676T 8860G 9101C 9767T 11467G 11719A 12308G 12372A 14766T 14935C 15043A 15061G 15236G 15326G 16051G 16234T 16240C 16242G 16311C 16519C
12	PAK_12	U7a	73G 151T 152C 263G 315.1C 523del 524del 750G 980C 1438G 1811G 2706G 3705A 3741T 4733C 4769G 4947C 5360T 7028T 8137T 8684T 8860G 10142T 11467G 11719A 12308G 12372A 13500C 14569A 14766T 15326G 15448A 16309G 16318T 16519C
13	PAK_13	HV12b1	150T 263G 315.1C 750G 1438G 2706G 4769G 7028T 7852A 8860G 11204C 12618A 13889A 15326G 15682G 16222T 16242T 16273A 16356C
14	PAK_14	U2c1b	73G 146C 152C 263G 315.1C 644G 709A 750G 1438G 1598A 1811G 2706G 4721G 4769G 5790A 7028T 8023C 8676T 8860G 9767T 10810C 11467G 11719A 11890G 12172G 12308G 12372A 14766T 14935C 15061G 15214C 15326G 16051G 16189C 16234T

15	PAK_15	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16519C
16	PAK_16	M30+16234	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11437C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16093C 16223T 16234T 16274A 16519C
17	PAK_17	H13a2a1	263G 315.1C 455.1T 709A 750G 1008G 1438G 2259T 3450A 4769G 8843C 8860G 14872T 15326G 16519C
18	PAK_18	M30d1	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 1598A 2706G 4769G 5557C 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15259T 15301A 15326G 15431A 16179del 16223T 16519C
19	PAK_19	M45	73G 146C 263G 315.1C 489C 750G 1438G 2706G 3083C 4059A 4734G 4769G 5319G 5585A 6827C 7028T 8701G 8860G 9095C 9180G 9509C 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14687G 14766T 14783C 15043A 15301A 15326G 15851G 16145A 16192T 16223T 16300G 16316G 16519C
20	PAK_20	M3c2	73G 93G 263G 315.1C 482C 489C 523del 524del 750G 1438G 2706G 4769G 5178T 7028T 8701G 8860G 9064A 9540C 10398G 10400T 10873C 11719A 12705T 13350G 14766T 14783C 15043A 15301A 15326G 16126C 16154C 16223T 16519C
21	PAK_21	M5a2	73G 195C 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4454C 4769G 5027T 6053T 6293C 6473T 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11275T 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16051G 16093C 16129A 16223T 16519C
22	PAK_22	U1a1c1	73G 263G 285T 315.1C 523del 524del 750G 1438G 2218T 2706G 4769G 4991A 6026A 7028T 7051C 7581C 8860G 11467G 11719A 12308G 12372A 12879C 13104G 14070G 14364A 14514C 14766T 15115C 15148A 15217A 15326G 15954C 16093C 16182C 16183C 16189C 16249C 16311C

23	PAK_23	U5b2	73G 150T 152C 257G 263G 264T 315.1C 750G 1438G 1721T 2706G 3197C 4080C 4769G 7028T 7768G 8860G 9139A 9477A 11467G 11719A 12308G 12372A 13434G 13617C 13637G 14182C 14199C 14409G 14766T 15326G 16270T
24	PAK_24	W3a1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1406C 1438G 2706G 3505G 4370C 4769G 5046A 5460A 7028T 8251A 8860G 8994A 11674T 11719A 11947G 12414C 12705T 13263G 14766T 15326G 15784C 15884C 16093C 16129A 16223T 16292T 16519C
25	PAK_25	U2b2	73G 146C 152C 234G 263G 315.1C 750G 1438G 1811G 1888A 4769G 5186T 7028T 8860G 9094T 9614G 11467G 11719A 12106T 12308G 12372A 12793C 13194A 13305T 13656C 14766T 15049T 15326G 15813C 15930A 16051G 16129A 16209C 16239T 16311C 16352C 16353T
26	PAK_26	M33a2a	73G 150T 263G 315.1C 462T 489C 750G 1438G 2361A 2706G 3543T 5124A 5423G 7028T 8176C 8562T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12705T 13731G 14766T 14783C 15043A 15301A 15326G 15908C 16169T 16172C 16223T 16278T 16519C
27	PAK_27	F1c1a2	73G 234G 249del 263G 315.1C 523del 524del 750G 1438G 1927A 2706G 3970T 4769G 6392C 6599G 6962A 7028T 8860G 9053A 9822T 10310A 10454C 10609C 11719A 12406A 12882T 13759A 13928C 14766T 15326G 16111T 16129A 16304C 16519C
28	PAK_28	M	73G 195C 207A 263G 315.1C 482C 489C 750G 1438G 2706G 3394C 4769G 7028T 8701G 8860G 9341G 9540C 10398G 10400T 10873C 11167G 11719A 11914A 12705T 14766T 14783C 15043A 15301A 15326G 15766G 15951G 16126C 16185T 16223T 16519C
29	PAK_29	M37e2	73G 152C 263G 315.1C 489C 750G 1438G 2706G 4769G 7028T 8410T 8701G 8860G 9540C 10398G 10400T 10556T 10873C 11050C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16111T 16189C 16223T 16224C 16294G 16295T 16519C

30	PAK_30	M5a2a1a1	73G 263G 315.1C 374G 489C 709A 750G 1189C 1438G 1888A 2706G 3921T 4454C 4769G 7028T 8701G 8860G 8886A 9540C 9947A 10398G 10400T 10873C 11719A 12372A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16129A 16223T 16265C 16344T 16519C
----	--------	----------	--

Consistent with the general evaluation process, the Pakistani haplotypes were assessed, and the results are summarized in Table 5.6.

Table 5.6. Number of Haplotypes (Ht), Haplotype Diversity (Hd), Probability of Discrimination (PD) and Probability of Identity (PI) for Pakistani samples.

Set	Ht	Hd	PD	PI
Pakistanis (n=50)	50	1	0.9800	0.0200

5.9.1. Haplogroup Assignment

For this dataset, the same process was followed for mtDNA sequences classification into haplogroups based on specific variations. The haplogroup results were successfully recorded, providing critical insights into the maternal lineages present within the dataset. Figure 5.23 provides an overview of the macro-haplogroups to the sequences obtained. Figure 5.24 represents the proportions of the macro-haplogroups within the dataset.









Cluster	Count	Percentage	Population Frequencies ?
M	25	50%	
U	11	22%	
R	5	10%	
T	3	6%	
HV	3	6%	
F	1	2%	
W	1	2%	
H	1	2%	

Figure 5.23. An overview of 8 macro-haplogroups in 50 samples of the Pakistani population, detailing sample counts, frequencies, and percentages.

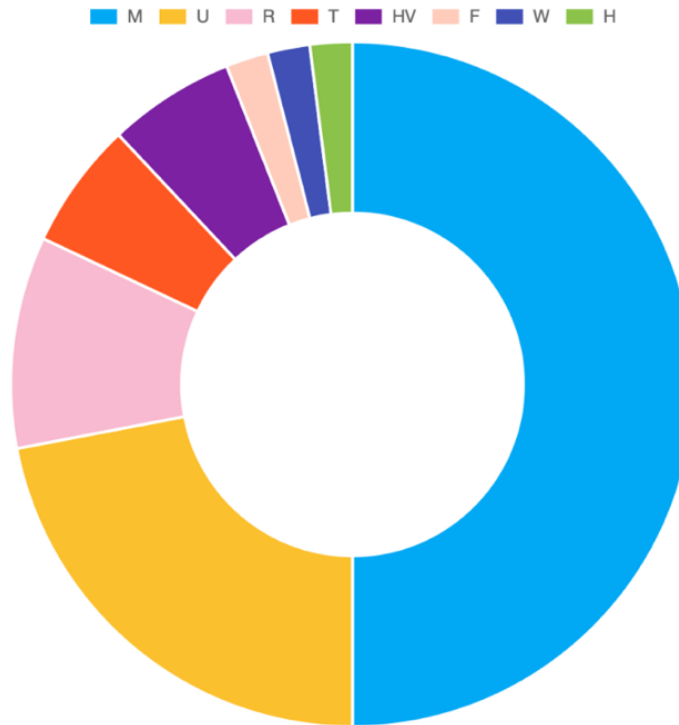


Figure 5.24. Whole mtDNA pie chart representation of the Pakistanis haplogroups frequencies identified by the current study conducted.

5.9.2. Haplogroup Diversity

The haplogroup diversity of each population is show below (Table 5.7)

Table 5.7. Number of Haplogroups (Hg) and calculated Haplogroup Diversity (HgD) for the three populations in this study.

Sets	Hg	HgD
Emiratis (n=510)	241	0.993713
Pakistanis (n=50)	42	0.981932
Indians (n=50)	40	0.991020

5.10. Concordance between Sanger Sequencing and MPS

A comparative analysis of Sanger sequencing and Massively Parallel Sequencing (MPS) data for the Control Region (CR) demonstrated 100% concordance, with no missing or

additional variants observed between the two sequencing approaches. This result confirms the high accuracy and reliability of both methods in mtDNA sequencing, particularly for forensic applications where precision is critical. The results obtained from Sanger sequencing and the Ion Chef/Ion S5 XL System were manually aligned to check for concordance between the two methods: all samples were fully concordant. The Precision ID mtDNA panel reported high uniformity and efficiency for all sequences. Compared to Sanger sequencing, variant calling showed concordance. Coverage shows amplicon were successfully amplified across the mtGenome. The results show that the concordance rates between MPS and conventional Sanger sequencing for the mtDNA control region were high.

5.11. Discussion

Based on the whole mtDNA genome sequences, the samples were further analyzed by dividing them into three distinct sets for additional findings. First, the sequences were converted to include the control region only for all samples. Then, a second analysis was performed focusing on HV1 and HV2 regions, and finally, the sequences were analyzed for the HV1 region alone. For each set, the number of haplotypes and the Power of Discrimination (PD) were calculated and compared to assess the impact of these regions on forensic identification. Table 5.8 summarizes the key findings from these analyses. All heteroplasmies (PH and LH) were included in the haplotypes and all calculations were based on this inclusion.

Table 5.8. Comparison of Sanger Sequencing and MPS.

	Haplotypes	Haplotype Diversity	PD
Whole genome	445	0.998929080	0.996970396
CR	409	0.998412882	0.996455210
HV1 AND HV2	397	0.998096999	0.996139946
HV1	279	0.990431064	0.988489043

In this dataset, the total number of samples analyzed was 510. When the whole mtDNA genome was used for analysis, 445 unique haplotypes were identified. However, when the sequences were restricted to the control region only, the number of haplotypes decreased to 409. This reduction in haplotype diversity indicates that some genetic variation is lost when analyzing only the control region compared to the entire mtDNA genome. The whole mtDNA genome carries more identification variability and thus contributes to the ability to differentiate between individuals. The control region, although informative, can limit the discrimination of individuals specially in a region where the genetic diversity is limited due to the culture of endogamy and consanguineous marriage, therefore it does not capture the full extent of the genetic diversity present in the entire mitochondrial genome. The decrease in haplotype count from 445 to 409 demonstrates that the coding region provides additional distinguishing power that is lost when the analysis is limited to the control region. This has direct implications for forensic applications, as the power of discrimination is reduced when relying solely on the control region.

When comparing the PD of this study to the study by Alshamali et al. (2008) study, the PD for the control region in this dataset is 0.9965 as in table 5.8, corresponding to a Random Match Probability (RMP) of 1:282. In contrast, the study by Alshamali et al.

(2008), which focused on 249 samples from Dubai, reported an RMP of 1 in 141 for the control region. The higher PD observed in this study can be attributed to the inclusion of samples from across the entire UAE, whereas Alshamali et al. (2008) was limited to Dubai. This broader sampling provides a more comprehensive representation of the genetic diversity within the Emirati population, leading to a greater power of discrimination in forensic applications.

To maximize the PD and to ensure the most accurate and comprehensive results, continuing with whole mtDNA analysis is the 'gold standard'. The inclusion of coding regions adds more detail and variation, which is especially important in forensic casework where distinguishing between individuals is critical. While the control region alone still offers substantial haplotype diversity, the whole mtDNA genome provides better resolution. For forensic purposes, where accuracy and individual differentiation are paramount, continuing with whole mtDNA sequencing would be the preferred approach, however, the degree of increased discrimination has to be considered alongside the extra effort and cost involved in sequencing over 16 kb.

Building a database using whole mtDNA is crucial as it captures a more comprehensive set of genetic variations, including those found in the coding regions. This enhanced resolution increases the power of discrimination and helps in more accurately identifying individuals or making population inferences. Such a database is essential for forensic casework and population studies, providing a reliable resource for matching and interpreting mtDNA evidence.

In this study, which consisted of 510 samples, the haplotype diversity was 0.9989 for the whole mtDNA genome, without excluding any variations. This indicates a very high level of genetic variation within the population, similar to findings in the study by Taylor et al. (2020). In this study, 1363 samples were sequenced using the whole mtDNA genome and the overall haplotype diversity was similarly high across all U.S. metapopulations. For example, in the Colorado African American (COAF) and Colorado Caucasian (COCN) datasets, haplotype diversity was 1 when point heteroplasmy (PHP) was included, meaning all haplotypes were unique. Even after excluding PHP, the diversity remained high, with COAF at 0.9997 and COCN still at 1. This is comparable to the results from the present study, where the haplotype diversity was similarly close to 1, reflecting substantial genetic variability.

In the Colorado Hispanic (COHS), the haplotype diversity was slightly lower (0.9983 with PHP included and 0.9958 without PHP), but still comparable to the high diversity observed in our dataset. The NIST African American (NTAF) and NIST Caucasian (NTCN) datasets also showed very high haplotype diversity, with values of 0.9998 for NTAF (0.9997 without PHP) and 0.9998 for NTCN (dropping to 0.9995 without PHP), which closely aligns with the diversity observed in the present analysis. The NIST Hispanic (NTHS) dataset had similarly high diversity (0.9986 with PHP and 0.9976 without), while the Department of Defense Serum Repository (DoDSR) datasets showed some of the highest diversity values, with the Asian American (DSAS) dataset reaching 0.9999

(0.9997 without PHP) and the Native American (DSNA) dataset at 0.9997 (0.9993 without PHP).

Both the current study and the study by Taylor et al. (2020) demonstrate strong haplotype diversity across the populations analyzed, with minimal reduction in diversity when variations such as point heteroplasmy were excluded. These consistently high diversity values, along with high PD ranging from approximately 98.66% to 99.59% in the Taylor et al. (2020) study, demonstrate that whole mtDNA sequencing is effective for forensic identification, distinguishing individuals reliably. Upon examining the data, whole genome sequencing provided a higher resolution for haplogroup classification compared to Sanger sequencing. For instant, since haplogroup H and R are closely defined by similar variations especially in the control region, this overlap makes it difficult to differentiate between the two using control region data alone. Here, whole mtDNA sequence data are necessary to accurately distinguish between the two haplogroup. This demonstrates that whole genome sequencing is particularly effective in distinguishing close haplogroups, which would otherwise be indistinguishable using only the control region, which also enhances individualization capabilities in forensic applications, particularly for haplogroups with complex substructures.

The whole mtDNA genome provides the highest level of resolution for haplotypes and detailed haplogroup assignment, allowing for the identification of even subtle haplogroups and subclades. The control region provides the critical variations for haplogroup assignments. While all haplogroups can still be identified, the uniqueness of haplotypes is

reduced, where some haplotypes with distinctive mutations beyond the control region will become indistinguishable from others. Therefore there may be ambiguity in distinguishing between haplotypes within the same haplogroup. In forensic analysis, the power of discrimination depends on those unique haplotypes that are more specific to individualization rather than identification.

5.12. Conclusion

The 100% concordance observed between Sanger sequencing and MPS in the Control Region (CR) validated the accuracy and forensic applicability of both sequencing approaches. While Sanger sequencing remained a well-established method for forensic mtDNA analysis, the seamless alignment with MPS results supported the continued adoption of high-throughput sequencing for forensic applications (Syndercombe Court, 2021). The absence of discrepancies in the CR indicated that Sanger sequencing was a highly effective method for mtDNA analysis, particularly when focusing on the hypervariable regions (HV1, HV2, and HV3) commonly used in forensic genetics. The MPS results fully aligned with Sanger data, reinforcing the validity of Ion Torrent sequencing for mtDNA analysis while providing additional discriminatory power by including SNPs beyond the CR. This level of concordance aligned with previous studies demonstrating that Sanger sequencing and MPS yielded highly comparable mtDNA profiles in forensic casework. Given that both methods produced identical results, Sanger sequencing could still serve as a quality control measure in forensic validation studies, for laboratories transitioning from traditional sequencing to MPS-based workflows.

The findings from this study reinforce the importance of MPS in analyzing mtDNA variation in the UAE, particularly in comparison to previously published studies on the UAE (MPS-based) as well as Kuwait, Lebanon, Jordan, and Bahrain (Sanger-based).

The Aljasmi et al. (2020) study on UAE mtDNA, which used whole-mtGenome sequencing, provided a strong baseline for evaluating lineage representation. It highlighted haplogroups L, U, and M as key contributors to the UAE's maternal genetic pool, reflecting African, South Asian, and Near Eastern ancestry. The presence of haplogroups U9a for example, which are rare globally, suggests that the UAE harbors underrepresented maternal lineages that may not be well captured in Sanger-based studies.

The comparison with Zimmermann et al. (2019), which analyzed Lebanese, Jordanian, and Bahraini populations using Sanger sequencing, revealed distinct haplogroup frequencies across these populations. Notably, haplogroup U3, found in Jordan, was also detected in the UAE dataset, suggesting a potential maternal ancestry link between the UAE and the Levant. Additionally, the low representation of haplogroup R0 in the UAE compared to Lebanon (36%) and Jordan (28%) suggests that the UAE's genetic structure is distinct, with greater influences from South Asian and African gene flow.

The addition of the Kuwaiti mtDNA dataset (Scheible et al., 2011) provides another regional reference point. The Kuwaiti population exhibited a high genetic diversity (0.9979), with haplogroups of Western Eurasian origin (74%), Asian (16%), and African (10%). The high frequency of haplogroup R0a (12%) in Kuwait aligns with its role as a key maternal lineage in the Arabian Peninsula. The African haplogroup representation in

Kuwait (10%) closely matches that observed in Bahrain and the UAE, supporting the idea of shared maternal ancestry across the Gulf, influenced by historical trade and migration routes.

The presence of underexpressed haplotypes or rare in the UAE dataset could suggest previously undocumented maternal lineages in Middle Eastern populations from previous studies. This could be attributed to MPS's enhanced resolution rather than an actual absence of these lineages in those populations. Future research comparing full mitochondrial genome sequences across Middle Eastern populations could refine the phylogenetic tree and improve forensic reference databases.

This study underscores the advantages of MPS over Sanger sequencing in forensic mtDNA analysis. While Sanger sequencing has historically provided valuable population-level mtDNA data, its limited coverage of the Control Region means that certain haplogroups, sub-lineages, and heteroplasmic variants may be underrepresented in global forensic databases. By integrating findings from UAE, Kuwait, and other Middle Eastern populations, this study strengthens the forensic and anthropological understanding of mtDNA variation in the region. The comparison of UAE, Kuwait, and other Middle Eastern populations confirms that the Arabian Peninsula is a major genetic intersection between Africa, South Asia, and the Levant. While haplogroups R0a and U3 indicate links to Levantine and Gulf populations, the higher representation of haplogroups L and M in the UAE suggests additional South Asian and African influences.

Chapter 6

6. Casework Samples

6.1.Introduction

Traditional forensic laboratories utilize Sanger sequencing for mtDNA analysis, which is costly, labor-intensive, and limited in scope. MPS allows for the sequencing of the entire mitochondrial genome in one run. It offers higher resolution, increased discrimination power, and the potential for mixture interpretation without consuming additional sample (Churchill et al., 2017). The study by Churchill et al. work aimed to provide guidance and criteria to help other laboratories implement MPS technology. Precision ID mtDNA Whole Genome Panel utilizes two primer pools of 81 pairs each, generating amplicons of ≤ 175 bp. This increases the potential for successful analysis of degraded samples. It can be used with the Ion Chef and Ion S5 systems for a largely automated workflow. In a study by Strobl et al. (2018) they evaluated the Precision ID Whole mtDNA Genome Panel for forensic analyses and provided a strong foundation for applying advanced MPS technologies in forensic genetics, to ensure reliable and detailed genetic analysis for law enforcement and legal proceedings.

Degraded samples are common in forensic investigations, making this feature particularly valuable. The panel uses small amplicons (≤ 175 base-pairs), which enhances its ability to analyse degraded DNA effectively. The MPS demonstrated the potential of overcoming

amplicon size, which is another success for implementing this technology in the forensic work schemes. The work also included evaluations of sensitivity, and performance with challenging samples, which will be further detailed in this Chapter.

6.2. Chapter Aim

The aim of this chapter was to evaluate the application of MPS in forensic casework by analyzing real forensic samples. This study aimed to assess the reliability, sensitivity, and overall forensic applicability of MPS, particularly in challenging scenarios involving low-template or degraded DNA samples. By incorporating casework examples, this chapter demonstrates the practical utility of MPS in forensic investigations, highlighting both its strengths and limitations in real-world applications.

6.3. Chapter Objectives

- To sequence forensic samples using the Ion Torrent technology with the Precision ID Whole mtDNA Genome Panel.
- To evaluate the efficacy of the Precision ID Whole mtDNA Genome Panel in forensic samples particularly in processing and analysing challenging DNA samples.

6.4. Methods

Note that detailed methodology can be obtained in Chapter 3 for comparison.

6.5. Casework Samples Sequencing

An evaluation of Precision ID mtDNA Whole Genome Panel using the Ion Torrent technology platform was performed on 56 casework samples (Table 6.1).

Table 6.1. Casework samples obtained from Forensic Biology section – Dubai Police (n=56).

Sample Names	Sample types	Quant. (Quant. Trio)	Degradation Indices	STR Profiles (GlobalFiler™)
CW_01	Blood Stain	1.3762	0.91	F
CW_02	Blood Stain	0.2412	0.67	F
CW_03	Postmortem Blood	0.035	2.21	P (16/24)
CW_04	Saliva	0.0143	0.72	F
CW_05	Touch DNA	0.0003	0.79	ND
CW_06	Touch DNA	0.0027	2.14	ND
CW_07	Touch DNA	0.0013	1.44	ND
CW_08	Touch DNA	0.0001	0.37	ND
CW_09	Touch DNA	0.0004	1.24	ND
CW_10	Touch DNA	0.0004	0.83	ND
CW_11	Touch DNA	0.0025	1.81	ND
CW_12	Touch DNA	0.0002	1.38	ND
CW_13	Touch DNA	0.6130	31.83	P (8/24)
CW_14	Touch DNA	0.0023	0.55	ND
CW_15	Touch DNA	0.0017	1.41	ND
CW_16	Touch DNA	0.0022	2.13	ND
CW_17	Touch DNA	0.0011	0.41	ND
CW_18	Touch DNA	0.0024	1.33	ND
CW_19	Touch DNA	0.0040	0.77	ND
CW_20	Touch DNA	0.0025	1.62	ND
CW_21	Touch DNA	0.0021	1.28	ND
CW_22	Touch DNA	0.3130	25.72	P (6/24)
CW_23	Hair	0.2450	2.233	P (14/24)
CW_24	Hair	1.1032	7.326	P (6/24)
CW_25	Hair	0.3221	3.212	P (9/24)
CW_26	Body Fluid	0.2222	1.233	P (18/24)
CW_27	Body Fluid	1.1034	9.341	P (12/24)
CW_28	Body Fluid	0.0362	1.111	P (6/24)
CW_29	Touch DNA	0.0023	10.46	ND

CW_30	Touch DNA	0.0025	4.11	ND
CW_31	Touch DNA	0.0031	14.33	ND
CW_32	Touch DNA	0.0021	6.01	ND
CW_33	Touch DNA	0.0058	18.42	ND
CW_34	Touch DNA	0.0021	11.23	ND
CW_35	Touch DNA	0.0026	3.12	ND
CW_36	Touch DNA	0.0001	5.22	ND
CW_37	Touch DNA	0.0123	7.32	P (8/24)
CW_38	Body Fluid	0.0134	21.04	ND
CW_39	Body Fluid	0.0827	9.32	P (14/24)
CW_40	Body Fluid	0.0383	2.22	P (17/24)
CW_41	Body Fluid	0.5841	7.659	F
CW_42	Touch DNA	0.0237	4.323	P (11/24)
CW_43	Touch DNA	0.2342	4.321	F
CW_44	Touch DNA	0.0021	6.78	ND
CW_45	Touch DNA	0.0713	26.31	P (9/24)
CW_46	Touch DNA	0.0722	15.01	P (7/24)
CW_47	Touch DNA	0.0021	10.32	ND
CW_48	Touch DNA	0.0037	6.23	ND
CW_49	Touch DNA	0.0201	14.35	ND
CW_50	Blood Stain	0.0128	31.44	P (3/24)
CW_51	Touch DNA	0.0012	24.46	ND
CW_52	Touch DNA	0.0033	3.12	ND
CW_53	Touch DNA	0.0021	3.10	P (9/24)
CW_54	Touch DNA	0.0138	27.01	ND
CW_55	Postmortem Blood	0.0037	13.11	ND
CW_56	Postmortem Blood	0.0292	42.71	P (3/24)

The average fragment size of the samples was 163 bp. General Performance of the samples was 117 bp–153 bp read lengths for casework samples. Average coverage values was 5,075–510,034 for casework samples. Figures 6.1. and 6.2. are the reported quality summary of the chips loaded on the Ion Torrent platform obtained from the TSS. The Figures illustrated the sequencing performance, quality metrics and data for the 2 chips,

including coverage uniformity, and read lengths achieved during the sequencing process.

Table 6.3. provides additional information on the quality controls of the 2 chips.

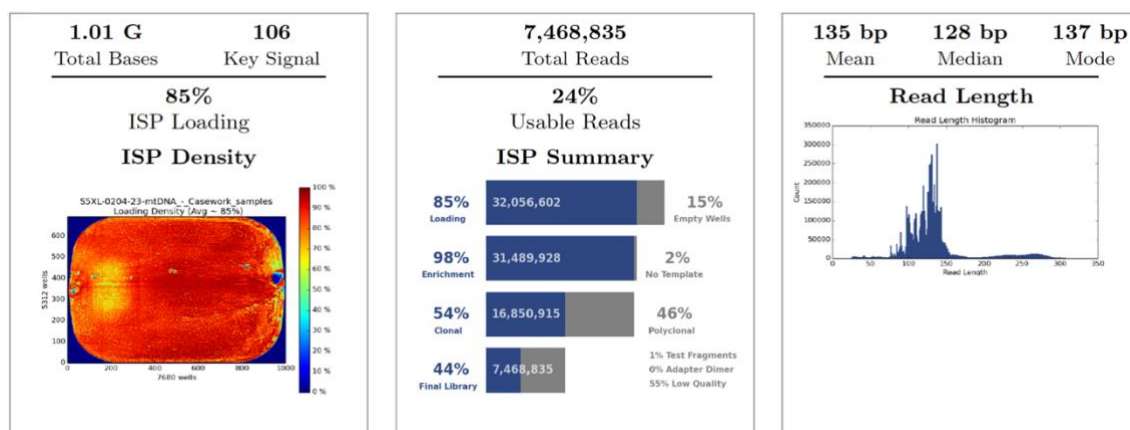


Figure 6.1. Casework samples chip 1 run summary from the TSS (n=28).

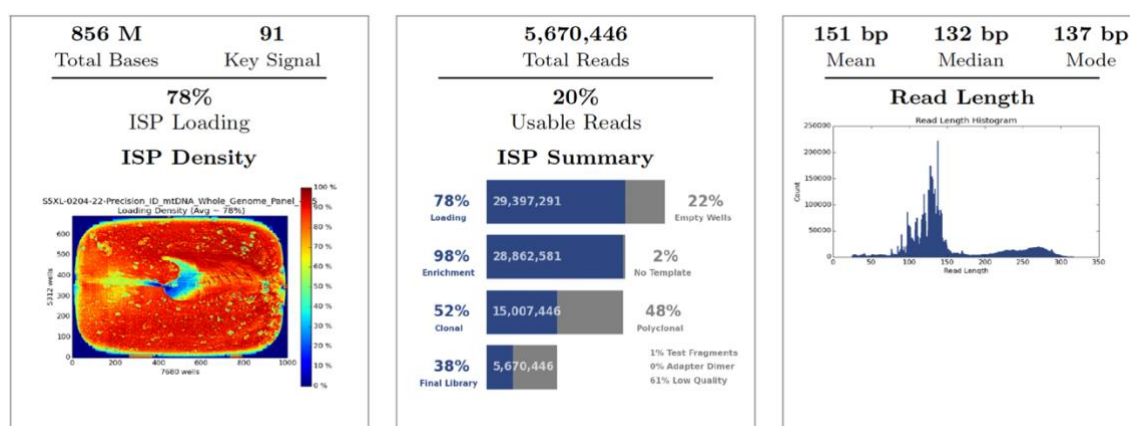


Figure 6.2. Casework samples chip 2 run summary from the TSS (n=28).

Table 6.2. Quality Control Results of casework samples (n=56).

Parameters	Reads	
Samples no.	28	28
Chip number	178	175
Chip type	530v1	530v1
Chip loading (%)	78%	85%
Enrichment (%)	98%	98%
Polyclonal (%)	48%	46%
Low quality (%)	61%	55%
Adapter Dimer (%)	00.1%	00.3%

Usable reads (%)	20%	24%
Mapped usable reads (%)	19.7%	23.8%
Mean Read length (%)	151 bp	135 bp
Total bases	856 M	1.01 G

The results showed that MPS workflow proved to be more sensitive compared to traditional STR profiling. All the samples were successfully sequenced and data were obtained for all casework and bone samples Table 6.3. summarized the resulted sequences from the casework samples. The successful recovery of samples supports the utility of the Precision ID Whole mtDNA Genome Panel for forensic applications, particularly in cases involving degraded or limited DNA samples. The results demonstrate the effectiveness of the panel in producing full mitogenomes with significant advantages in terms of recovery capability, suitable for forensic analyses. The ability to obtain detailed mitochondrial profiles even from highly degraded samples could significantly impact forensic casework involving old or compromised samples. As explained earlier in Chapter 4, the minimum total read coverage per position was set to 20, therefore all positions recovered for both caseworks and bones samples were reported based on this range. Whole coverage of the samples means a full sequence is obtained by forward and reverse amplicons without gaps in between the amplicons. Likewise, partial means that the sequence had gaps in between the amplicons but with good coverage, passing the quality controls. In other words, whole sequence signifies that complete sequence data was obtained.

Table 6.3. Sequence data summary for casework samples in comparison with traditional profiling results (n=56).

Sample Names	STR Profiles (GlobalFiler™)	Sequence Data
CW_01	F	Whole
CW_02	F	Whole
CW_03	P (16/24)	Whole
CW_04	F	Partial
CW_05	ND	Partial
CW_06	ND	Partial
CW_07	ND	Partial
CW_08	ND	Partial
CW_09	ND	Partial
CW_10	ND	Partial
CW_11	ND	Partial
CW_12	ND	Partial
CW_13	P (8/24)	Whole
CW_14	ND	Partial
CW_15	ND	Partial
CW_16	ND	Partial
CW_17	ND	Partial
CW_18	ND	Partial
CW_19	ND	Partial
CW_20	ND	Partial
CW_21	ND	Partial
CW_22	P (6/24)	Whole
CW_23	P (14/24)	Whole
CW_24	P (6/24)	Whole
CW_25	P (9/24)	Whole
CW_26	P (18/24)	Whole
CW_27	P (12/24)	Whole
CW_28	P (6/24)	Whole
CW_29	ND	Partial
CW_30	ND	Partial
CW_31	ND	Partial
CW_32	ND	Partial
CW_33	ND	Partial

CW_34	ND	Partial
CW_35	ND	Partial
CW_36	ND	Partial
CW_37	P (8/24)	Whole
CW_38	ND	Partial
CW_39	P (14/24)	Whole
CW_40	P (17/24)	Whole
CW_41	F	Whole
CW_42	P (11/24)	Whole
CW_43	F	Whole
CW_44	ND	Partial
CW_45	P (9/24)	Whole
CW_46	P (7/24)	Whole
CW_47	ND	Partial
CW_48	ND	Partial
CW_49	ND	Partial
CW_50	P (3/24)	Whole
CW_51	ND	Partial
CW_52	ND	Partial
CW_53	P (9/24)	Whole
CW_54	ND	Partial
CW_55	ND	Partial
CW_56	P (3/24)	Whole

6.6. Results

In forensic casework, partial mtDNA data remain highly valuable, especially when dealing with either degraded, old, or limited samples where obtaining complete sequences can be challenging. The haplotypes derived from partial sequences could still be compared against established databases; EMPOP for instance which can aid in determining the commonality or rarity of a haplotype within a given population. This is an essential factor in narrowing the investigation scope in forensic cases. Although the discriminatory power

is reduced compared to whole mtDNA data, analysis can still be valuable. Matching the partial mtDNA haplotype from a crime scene to a reference sample allowed for the inclusion or exclusion of individuals, thus helping to identify suspects or victims. Additionally, partial mtDNA haplotypes could either confirm or exclude maternal lineage connections, providing critical evidence in forensic casework.

With partial data, haplogroup assignment could still be performed, but the results were often less precise or more generalized. Since many haplogroup-defining mutations are located throughout the whole mtDNA genome, partial data can limit the ability to assign specific haplogroups or sub-haplogroups. For instance, while it was possible to determine that a sample belonged to a broader haplogroup, identifying specific subclades without the full sequence remained challenging.

As discussed in Chapter 5, the power of discrimination (PD) did not decline drastically when shifting from the whole mtDNA genome to the control region, although the whole genome offered additional information on the haplotypes. In the studied samples, the gaps in the data did not significantly impact haplogroup assignments, yet they affected the resolution of haplotype uniqueness. This meant that the data retrieved had a successful recovery of critical variations which was crucial for haplogroups. Yet, in matter of identification, the degree of unique mitotype (haplotypes) is reduced.

Re-analyzing samples to cover additional fragments also improved accuracy, which in the current case was considered to some samples to enhance the resolution of the sequence data. In many forensic cases, samples were either degraded, aged, or available in small

quantities, making whole mtDNA sequencing challenging. Nevertheless, the value of the partial data recovered from the study remained significant. Haplotypes data are not shown in this chapter due to confidentiality, as the samples included real casework material.

6.7. Discussion

A pie chart was generated to visualize the distribution of the casework samples in comparison to analyzed population sets in this study where they resemble the developing mtDNA database. This comparison allowed for an assessment of how the unknown samples aligned with the established haplogroup frequencies of the Emirati, Indian, and Pakistani populations. The analysis aimed to identify the distribution of haplogroups between the samples and the reference populations.

In Figure 6.3. the top pie chart represented the overall distribution of the casework samples, while the three lower charts illustrated the haplogroup distributions for the Emirati, Indian, and Pakistani population sets, respectively. A key observation was that haplogroup L constituted a substantial portion—approximately 33% in total (L0: 4%, L1: 11%, L2: 7%, L3: 9%, L4: 2%). This finding suggested that a significant fraction of the casework samples likely originated from individuals of African or African-descendant ancestry, as haplogroup L is typically prevalent in African lineages (Heinz et al., 2017). The diversity within haplogroup L appeared to be lower in the Emirati population, where it accounted for only 10%. Haplogroup M comprised 25% of the casework samples, which was consistent with its dominance in the Pakistani population

(50%) and the Indian population (48%). Haplogroup U showed a higher presence in the Indian (24%) and Pakistani (20%) population sets, while it represented 15% of the casework samples. Haplogroup H was more prominent in the Emirati population (21%), whereas it constituted only 5% of the casework samples.

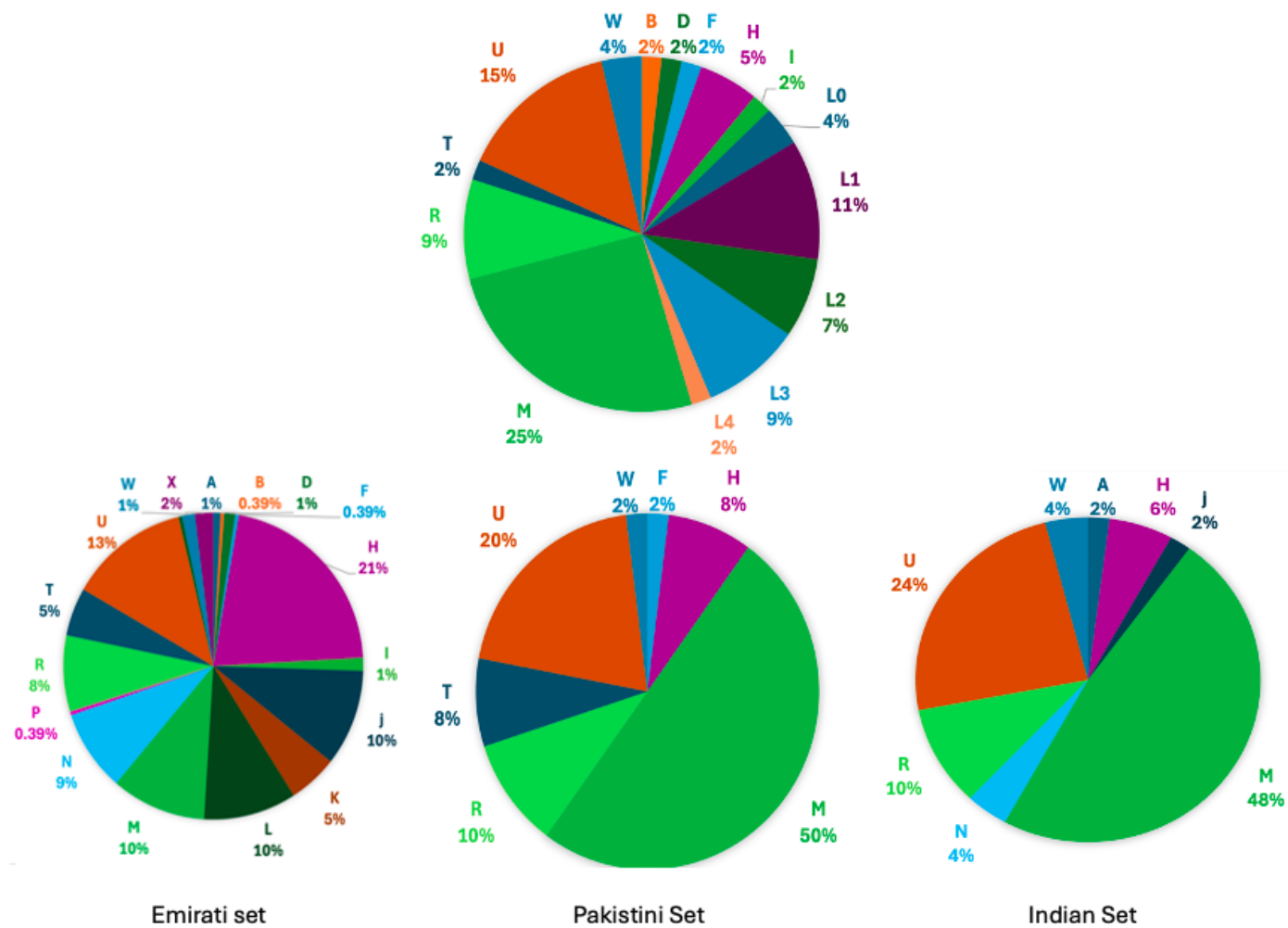


Figure 6.3. Comparison of the casework samples to the three populations data sets.

6.8. Conclusion

This chapter demonstrated the successful application of MPS technology for sequencing forensic samples. The results confirmed the Precision ID Whole mtDNA Genome Panel's capability to generate high-quality mtDNA sequences, supporting its reliability in forensic investigations. Additionally, the processing of challenging forensic samples, including low-template and degraded DNA, provided an assessment of the technology's efficacy in forensic mtDNA analysis. The successful recovery of high-resolution mitochondrial data reinforced confidence in the technology, which may positively influence decisions regarding its acquisition and implementation.

Moreover, the resulting comparisons highlighted the need to expand reference datasets to include a broader range of populations, providing deeper insights into haplogroup distribution.

Chapter 7

7. Bone and High Primate Samples

7.1. Introduction

The sensitivity of mitochondrial DNA (mtDNA) analysis plays a crucial role in forensic casework, particularly when dealing with challenging samples such as bones or high primate specimens. Bone samples, often recovered from forensic or archaeological contexts, tend to be highly degraded due to environmental exposure, making mtDNA an essential tool for identification when nuclear DNA may be compromised. Similarly, high primate samples are of interest not only in evolutionary studies but also for ensuring accurate species identification, given their genetic closeness to humans.

7.2. Chapter Aim

In this chapter of the study, the aim was to assess the sensitivity and effectiveness of MPS for recovering mtDNA from these challenging sample types to understand the limitations and capabilities of MPS in generating reliable mtDNA profiles, which has direct applications in forensic studies. This analysis is critical for enhancing the accuracy of forensic identification in cases where conventional DNA analysis may fail due to either sample degradation, complexity or both.

7.3. Chapter Objectives

- To evaluate the sensitivity of MPS for generating reliable mtDNA profiles.
- To assess the performance of MPS in analyzing mtDNA from high primate samples.
- To sequence solid tissue samples and high primate samples to understand the effectiveness of MPS across different challenging sample types and draw conclusion to differentiate non-human samples from human samples.
- To investigate the quality and accuracy of mtDNA sequencing results of solid tissue samples, with the goal of improving methodologies for difficult forensic cases.

7.4. Methods

Note that detailed methodology can be obtained in chapter 2 and 3 for comparison.

7.5. Solid Tissues Sequencing

An evaluation of Precision ID mtDNA Whole Genome Panel using the Ion Torrent technology platform was performed on 56 solid tissue samples as illustrated in table 7.1..

Table 7.1. Solid Tissue samples obtained from Forensic Biology section – Dubai Police (n=56).

Sample Names	Sample types	Quant. (Quant. Trio)	Degradation Indices	STR Profiles (GF)
BT_01	Femur	0.0102	1.19	P (7/24)
BT_02	Femur	0.0078	1.11	P (5/24)
BT_03	Femur	0.0860	5.85	P (11/24)
BT_04	Femur	0.0043	0.92	P (10/24)
BT_05	Femur	0.0063	1.10	P (10/24)

BT_06	Femur	0.0582	7.73	P (3/24)
BT_07	Femur	0.0013	2.86	P (7/24)
BT_08	Femur	0.0281	1.99	P (5/24)
BT_09	Femur	0.0059	3.91	P (3/24)
BT_10	Femur	0.0086	5.23	ND
BT_11	Femur	0.0182	1.59	P (5/24)
BT_12	Femur	0.0172	1.49	P (5/24)
BT_13	Femur	0.0089	0.99	P (12/24)
BT_14	Femur	0.0896	0.94	F
BT_15	Femur	0.1262	1.03	F
BT_16	Femur	0.0449	3.19	P (15/24)
BT_17	Femur	0.0752	1.51	F
BT_18	Femur	0.0309	1.09	P (13/24)
BT_19	Femur	0.0053	1.13	P (8/24)
BT_20	Femur	0.0018	2.65	ND
BT_21	Femur	0.0018	0.89	ND
BT_22	Femur	0.0112	0.95	F
BT_23	Skull	0.0335	2.98	F
BT_24	Skull	0.0182	3.19	P (6/24)
BT_25	Skull	0.0019	0.89	P (14/24)
BT_26	Skull	0.2045	1.17	ND
BT_27	Skull	0.0053	5.21	ND
BT_28	Skull	0.0006	2.91	ND
BT_29	Skull	0.0184	4.96	P (13/24)
BT_30	Skull	0.0014	0.73	P (7/24)
BT_31	Skull	0.0121	6.11	P (9/24)
BT_32	Skull	0.0147	1.52	P (23/24)
BT_33	Skull	0.8478	0.84	F
BT_34	Skull	1.4324	0.83	F
BT_35	Skull	1.8098	0.77	F
BT_36	Skull	2.1552	1.10	F
BT_37	Skull	0.0290	0.89	P (21/24)
BT_38	Skull	0.0172	15.21	ND
BT_39	Skull	0.0405	7.41	P (19/24)
BT_40	Skull	0.0749	12.57	ND
BT_41	Tooth	0.0111	12	ND
BT_42	Tooth	0.0477	12.55	ND
BT_43	Tooth	0.0022	3.89	P (3/24)

BT_44	Tooth	0.0080	11	ND
BT_45	Tooth	0.0031	10.13	ND
BT_46	Tooth	0.0077	3.45	ND
BT_47	Tooth	0.0421	6.01	ND
BT_48	Tooth	0.0009	0.88	P (6/24)
BT_49	Tooth	0.0121	8.07	ND
BT_50	Tooth	0.0016	8.00	ND
BT_51	Tooth	0.0146	9.13	ND
BT_52	Tooth	0.1327	1.32	F
BT_53	Tooth	0.0022	4.40	ND
BT_54	Tooth	0.0059	0.78	P (8/24)
BT_55	Tooth	0.0356	1.04	F
BT_56	Tooth	0.0327	1.61	F

The average fragment size of the samples was 163 bp. General Performance of the samples was 111 bp–146 bp read lengths for the samples with an average coverage values between 90,137–1,250,961 for the mitogenomes. Figures 7.1. and 7.2. are the quality summary of the chips loaded on the Ion Torrent platform. The Figures illustrated the sequencing performance, quality metrics and data for the 2 chips, including coverage uniformity, and read lengths achieved during the sequencing process. Table 7.2. provides additional information on the quality controls of the 2 chips.

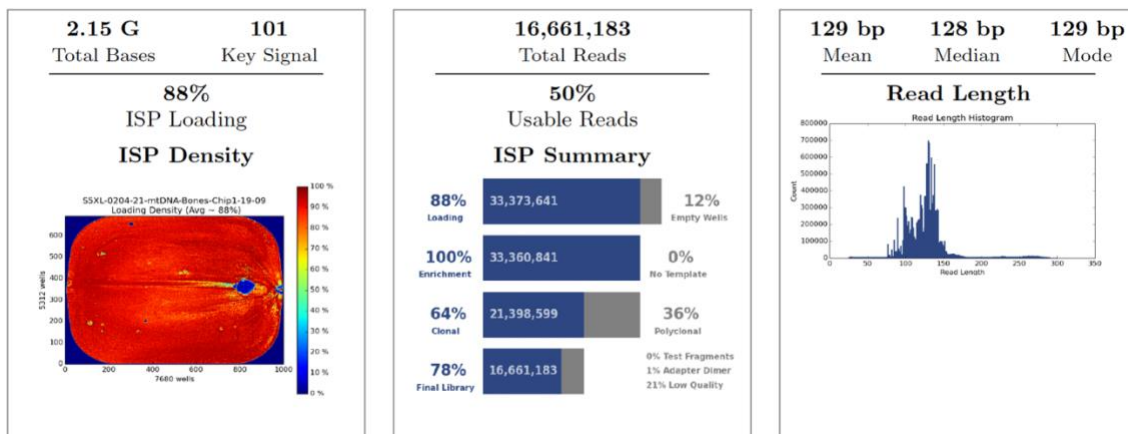


Figure 7.1. Solid tissue samples chip 1 run summary from the TSS (n=28).

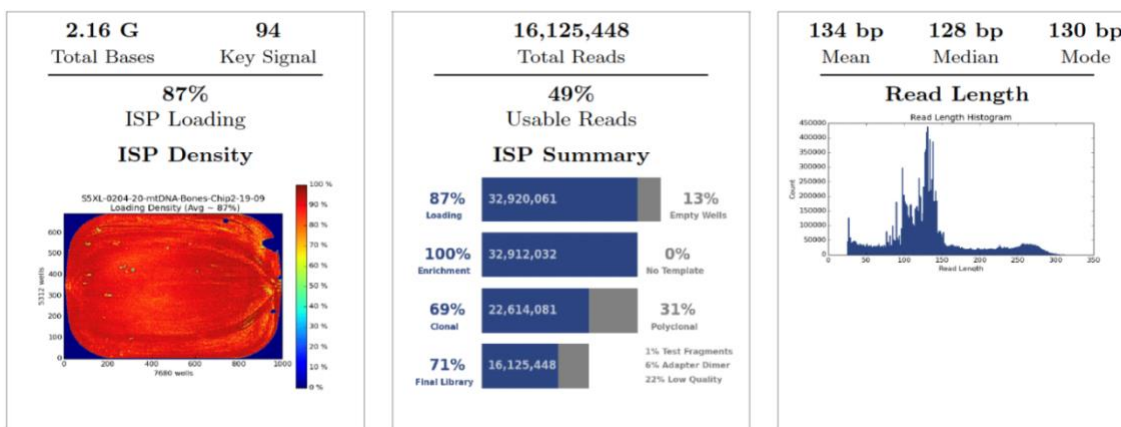


Figure 7.2. Solid tissue samples chip 2 run summary from the TSS (n=31).

Table 7.2. Quality Control Results of solid tissue samples (n=56).

Parameters	Reads	
Samples no.	28	31
Chip number	174	172
Chip type	530v1	530v1
Chip loading (%)	88%	87%
Enrichment (%)	100%	100%
Polyclonal (%)	36%	31%
Low quality (%)	21%	22%
Adapter Dimer (%)	00.6%	04.2%
Usable reads (%)	50%	49%
Mapped usable reads (%)	50.1%	49.5%
Mean Read length (%)	129 bp	134 bp
Total bases	2.15 G	2.16 G

The results show that MPS workflow proved to be more sensitive compared to traditional STR profiling. All the samples were successfully sequenced and data were obtained for all the samples. Table 7.3. summarized the results from the study. The samples were successfully recovered which supported the utility of the Precision ID Whole mtDNA Genome Panel for forensic applications, particularly in degraded or limited DNA samples.

The results demonstrate the effectiveness of the panel in producing full mitogenomes with significant advantages in terms of recovery capability, suitable for forensic analyses. Detailed mtDNA profiles from highly degraded samples were useful for forensic casework involving old or compromised samples. As stated in Chapter 4, positions recovered were reported with a minimum total read coverage of 20 per position. Whole coverage means obtaining a complete sequence with no gaps between forward and reverse amplicons. Partial coverage has gaps but still passes quality controls.

7.6.High Primates Samples

High primates, also known as haplorhines or simians, encompass both Old World and New World monkeys, as well as apes. The study involved sequencing two chimpanzees (*Pan troglodytes*), one female and one male, and a gibbon (*Hylobatidae*). These samples were transferred from the local zoo veterinary clinic for the purpose of this study. Non-human DNA samples were used to test for cross-reactivity and species specificity. The samples were ran on chip 2 of this part of the study.

7.7.Results

The study assessed the sensitivity of the Precision ID mtDNA Whole Genome Panel by determining its ability to analyse DNA samples with low concentrations. The panel was tested for its performance with DNA amounts as low as one picogram (pg) of genomic DNA derived from solid tissues samples. Table 7.3. summarized the returned sequences data resulted from the MPS.

The summary of high primates' findings including full amplification of the Chimpanzees and Gibbon were identifiable through abundant SNPs and point heteroplasmy (Brown et al., 1982). The results are not included due to the signed consent with the clinic, yet a capture of part of the profile is provided in Figure 7.3. for visualization purposes.

Table 7.3. Sequence data summary for the solid tissue samples (bones and teeth) in comparison with traditional profiling results

Sample Names	STR Profiles (GF)	Sequence Data
BT_01	P (7/24)	Whole
BT_02	P (5/24)	Whole
BT_03	P (11/24)	Whole
BT_04	P (10/24)	Whole
BT_05	P (10/24)	Whole
BT_06	P (3/24)	Whole
BT_07	P (7/24)	Whole
BT_08	P (5/24)	Whole
BT_09	P (3/24)	Whole
BT_10	ND	Partial
BT_11	P (5/24)	Whole
BT_12	P (5/24)	Whole
BT_13	P (12/24)	Whole
BT_14	F	Whole
BT_15	F	Whole
BT_16	P (15/24)	Whole
BT_17	F	Whole
BT_18	P (13/24)	Whole
BT_19	P (8/24)	Whole
BT_20	ND	Partial
BT_21	ND	Partial
BT_22	F	Whole
BT_23	F	Whole
BT_24	P (6/24)	Whole
BT_25	P (14/24)	Whole

BT_26	ND	Whole
BT_27	ND	Whole
BT_28	ND	Partial
BT_29	P (13/24)	Whole
BT_30	P (7/24)	Whole
BT_31	P (9/24)	Whole
BT_32	P (23/24)	Whole
BT_33	F	Whole
BT_34	F	Whole
BT_35	F	Whole
BT_36	F	Whole
BT_37	P (21/24)	Whole
BT_38	ND	Whole
BT_39	P (19/24)	Whole
BT_40	ND	Partial
BT_41	ND	Whole
BT_42	ND	Partial
BT_43	P (3/24)	Whole
BT_44	ND	Partial
BT_45	ND	Partial
BT_46	ND	Partial
BT_47	ND	Partial
BT_48	P (6/24)	Whole
BT_49	ND	Partial
BT_50	ND	Whole
BT_51	ND	Whole
BT_52	F	Whole
BT_53	ND	Partial
BT_54	P (8/24)	Whole
BT_55	F	Whole
BT_56	F	Whole

Position	G%	A%	T%	C%	N%	ins%	del%	Polymorphism	State	Frequency	Artefact	Var Strand	Read Stran	EMPOP	Score
29	99.9	0	0	0.1	0	0.1	0	29G	confirmed	99.9	True variant 0.5	0.6		unknown in empop	0.853
40	0	0	4	95.7	0	0.3	0.3	40C	likely	95.7	True variant 0.5	0.6		unknown in empop	0.687
41	0	0	96.6	3.2	0	0	0.1	41T	confirmed	96.6	True variant 0.5	0.6		unchecked	1
73	99.9	0.1	0	0	0	0.3	0	73G	confirmed	99.9	True variant 0.5	0.7		unchecked	1
94	2.2	96.9	0	0	0	2.1	0.8	94A	confirmed	96.9	True variant 0.5	0.7		unchecked	1
102	0	0	0.1	98.3	0	0.6	1.5	102C	confirmed	98.3	True variant 0.5	0.7		unknown in empop	0.853
103	0	0	0.1	99.9	0	0	0	103C	confirmed	99.9	True variant 0.5	0.7		unknown in empop	1
146	0	0	0	100	0	1.2	0	146C	confirmed	100	True variant 0.5	0.6		unchecked	1
151	0	98.8	0	0	0	0	1.2	151A	confirmed	98.8	True variant 0.5	0.6		confirmed	1
152	0	1.2	0	98.8	0	0	0	152C	confirmed	98.8	True variant 0.5	0.6		unchecked	1
153	98.8	0	0	1.2	0	0	0	153G	confirmed	98.8	True variant 0.5	0.6		unchecked	1
155	0	0	11.4	88.6	0	0	0	155C	possible	88.6	True variant 0.5	0.6		unknown in empop	0.52
182	0	0	100	0	0	0	0	182T	confirmed	100	True variant 0.5	0.6		unchecked	1
184	0	0	100	0	0	0	0	184T	confirmed	100	True variant 0.5	0.6		unknown in empop	1
185	0	0	89.7	10.3	0	0	0	185Y	likely	89.7	True variant 0.5	0.6		unknown in empop	0.67
187	0	0	0	100	0	0	0	187C	confirmed	100	True variant 0.5	0.6		unknown in empop	1
189	100	0	0	0	0	0	0	189G	confirmed	100	True variant 0.5	0.6		unchecked	1
194	0	0	0	0	0	100	0	194.1A	confirmed	98.5	True variant 0.5	0.6		unknown in empop	1
197	100	0	0	0	0	0	0	197G	confirmed	100	True variant 0.5	0.6		unchecked	1
198	0	0	89.7	10.3	0	0	0	198T	likely	89.7	True variant 0.5	0.6		unchecked	0.67
204	0	0	0	100	0	0	0	204C	confirmed	100	True variant 0.5	0.6		unchecked	1
236	0	2.7	0	97.3	0	2.7	0	236C	likely	97.3	True variant 0.5	1		unchecked	0.703
239	0	0	0	100	0	0	0	239C	likely	100	True variant 0.5	1		unchecked	0.703
241	100	0	0	0	0	0	0	241G	likely	100	True variant 0.5	1		unchecked	0.703
245	2.7	0	0	97.3	0	0	0	245C	likely	97.3	True variant 0.5	1		unchecked	0.703
246	97.4	0	0	2.6	0	0	0	246G	likely	97.4	True variant 0.5	1		unknown in empop	0.703
247	2.6	97.4	0	0	0	0	0	247A	likely	97.4	True variant 0.5	1		unchecked	0.703
710	0.1	0	1.7	88.4	0	0.3	9.8	710C	likely	88.4	True variant 0.5	0.5		unchecked	0.67
711	0	0	0.1	99.6	0	0	0.3	711C	confirmed	99.6	True variant 0.5	0.5		unchecked	1
714	95.3	4.3	0	0.4	0	0.1	0	714G	likely	95.3	True variant 0.5	0.5		unchecked	0.815

Figure 7.3. A partial screenshot from TSS output file of the male Chimpanzee results.

7.8. Discussion

The results reveal the analysis of high primate samples have demonstrated that the detection of numerous mutations and point heteroplasmy are not only a key factor in identifying non-human DNA but also a reflection of the high sensitivity of MPS. One of the strengths of MPS lies in its ability to detect even the smallest variations across the mitochondrial genome, which is critical for distinguishing non-human samples. The ability to capture such detailed and extensive variation highlights the superior sensitivity of the technology (Canale et al., 2025). The use of small amplicons highlights the capabilities of MPS. These small amplicons offer precise coverage of mtDNA, facilitating the recovery of high-quality data from degraded samples. In forensic contexts, where DNA is often highly fragmented due to either environmental exposure or decay, the ability to amplify and sequence these short fragments is invaluable. Cilhar et al. (2020) demonstrated that PCR-based MPS workflows can recover full mitogenomes from highly degraded samples. The high sensitivity of MPS enables the identification of mtDNA in cases where nuclear DNA may be too degraded for analysis, thus making it an essential tool for forensic investigations (Syndercombe Court, 2021).

Moreover, the capacity to pick up as much reading and identification of variations as possible is a testament to the effectiveness of MPS in analyzing both high primate and human mtDNA. In degraded samples, where DNA exists in very short fragments, traditional methods may struggle to retrieve sufficient information. However, MPS, with its high-throughput sequencing of small amplicons, ensures that even fragmented mtDNA

can be analyzed comprehensively, allowing for accurate species identification and lineage tracing. This sensitivity is particularly important in challenging forensic cases, such as those involving bones or other degraded remains, where identifying the species or individual is critical (Holt et al., 2019). The Precision ID Whole mtDNA Genome Panel's capability for challenging samples is further enhanced by recent protocol refinements (Canale et al., 2025), which address limitations in low-template and degraded DNA analysis. These improvements support its adoption for high-stakes forensic casework.

Figure 7.3. presented a pie chart depicting the distribution of bone samples, compared to the Emirati, Indian, and Pakistani population sets. Differences in haplogroup frequencies between the sample and reference populations were observed. Haplogroup H was the most dominant in the bone samples (36%), with the highest representation in the Emirati population (21%). Haplogroup M, which was most prevalent in the Pakistani (50%) and Indian (48%) populations, constituted only 4% of the overall bone samples. Haplogroup U was present in all sets at varying frequencies.

The results of this study have shown that, the high sensitivity of MPS, combined with the use of small amplicons, enhances the detection of mtDNA variations and allows for the successful analysis of degraded samples. The ability to pick up a wide range of mutations and point heteroplasmy not only serves as a major differentiation factor for non-human samples but also highlights the technological advancements that MPS brings to forensic and evolutionary research. These capabilities make MPS a powerful tool for addressing the challenges of species identification and handling degraded forensic samples.

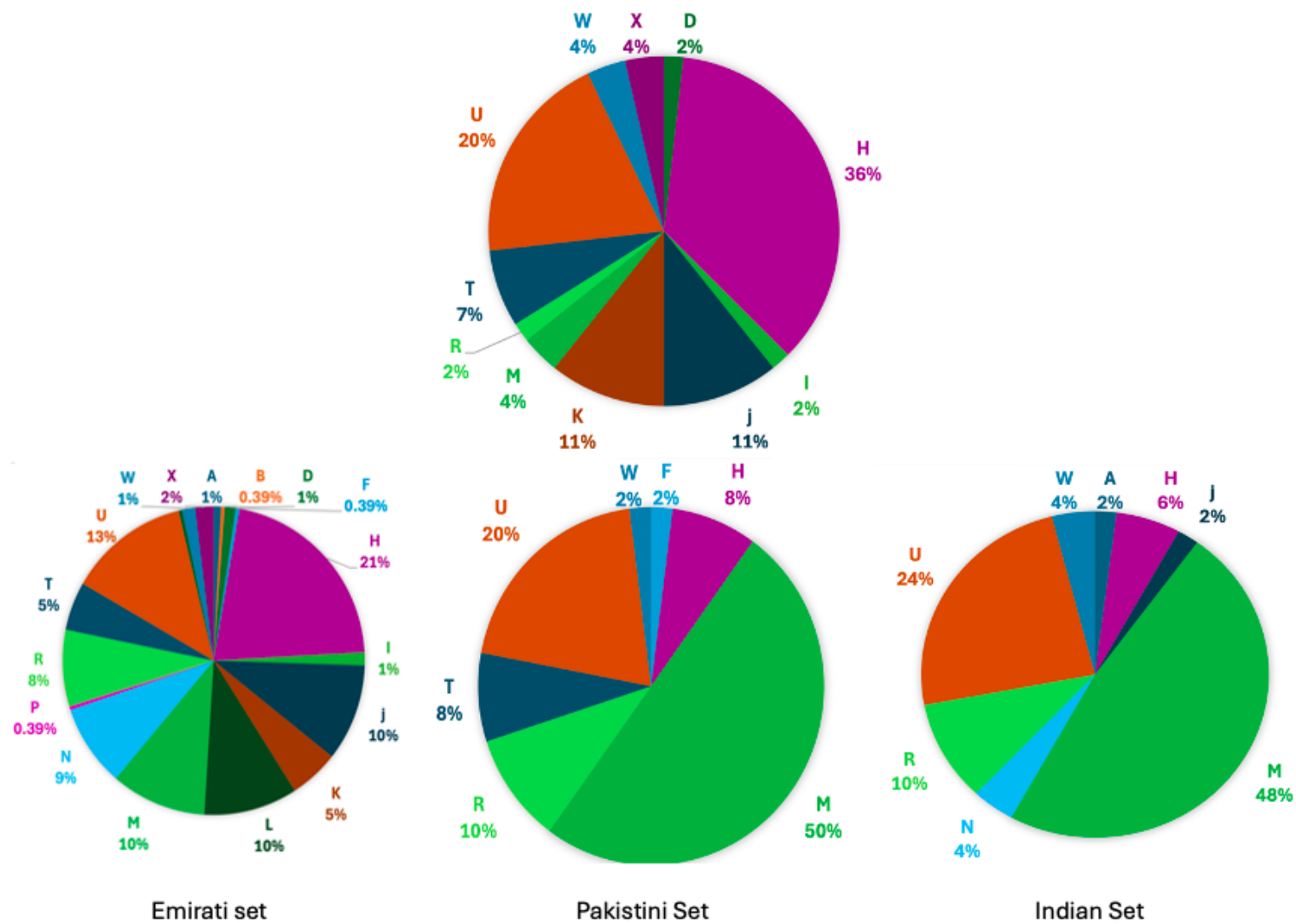


Figure 7.4. Comparison of the samples of solid tissues to the three populations data sets.

7.9. Conclusion

In conclusion, this chapter assessed the applicability of MPS in forensic contexts, particularly its effectiveness in degraded human samples recovery and distinguishing non-human DNA that can be presented in the crime scene. The study aimed to determine whether MPS could reliably analyze mtDNA from degraded or low-template forensic samples, such as bones. Also evaluated the potential for cross-contamination or interference from non-human DNA, which may be encountered at crime scenes.

The results confirmed that MPS effectively recovered mtDNA profiles from degraded bone samples, demonstrating its suitability for forensic casework. Since bones often represent low-quantity or highly fragmented DNA sources, their successful sequencing highlights MPS's capability to analyze forensic samples where nuclear DNA may be too compromised for identification. The use of small amplicons allowed for more efficient sequencing of mtDNA fragments (Zavala et al., 2022). The Precision ID panel's reliability for degraded DNA is further supported by Oliveira et al. (2024), which demonstrated full mitogenome recovery from century-old skeletal remains, with no false positives in haplogroup calls.

Additionally, the study demonstrated that high primate DNA, despite being genetically similar to human DNA, exhibited an abundance of species-specific SNPs and point heteroplasmy, enabling clear differentiation from human samples. This finding is crucial in forensic investigations, as it confirms that MPS can accurately distinguish non-human DNA, reducing the risk of misinterpretation due to potential cross-contamination.

Overall, this study reaffirmed that MPS can be utilized as a powerful and reliable tool for forensic mtDNA analysis that proved its effectiveness in degraded skeletal remains (Marshall et al., 2020). Furthermore, the comparative analysis between human and non-human samples demonstrated that MPS is highly reliable in preventing forensic misidentifications. These findings emphasize the importance of advancing forensic methodologies to enhance and improve the reliability of forensic DNA profiling in complex case scenarios/studies.

Chapter 8

8. General Discussion

8.1. Introduction

Mitochondrial DNA (mtDNA) is maternally inherited genetic material and exists in numerous copies per biological cell, which makes it especially valuable in forensic investigations involving degraded or minimal nuclear DNA recovered from biological samples left by perpetrator at crime scenes. MtDNA analysis is crucial in a variety of forensic contexts, such as identifying human remains, examining ancient DNA, solving missing persons cases, and dealing with highly degraded samples where nuclear DNA may not be viable (Parson and Dur, 2007; Budowle et al., 2003; Holland and Parsons, 1999)

In the context of mitochondrial DNA (mtDNA) analysis, whole-genome (WG) sequencing provides a broader scope compared to traditional control region (CR) sequencing. While CR sequencing focuses exclusively on the hypervariable regions HV1, HV2, and occasionally HV3 within the control region, which are often sufficient for forensic identification due to their high individual variability, WG sequencing encompasses the entire mtDNA, including variations in both coding and non-coding regions (Parson et al., 2014). This approach enhances the resolution for distinguishing maternal lineages and detecting low-level heteroplasmy that might be overlooked in CR-only approaches (Just

et al., 2015). However, despite these advantages, the practical utility of the additional information provided by WG sequencing in forensic casework is often limited, as most forensic databases and comparison standards, such as EMPOP, are based on CR sequences (Budowle et al., 2017). Furthermore, the increased complexity and cost of WG sequencing, combined with the challenges of interpreting novel or rare variants outside the CR, make it less attractive for routine forensic applications (King et al., 2018). Therefore, while WG sequencing offers a more detailed genetic portrait, its added value over CR sequencing in forensic contexts is not always proportional to the increased resource investment required. Sequencing the control region remains essential for comparing forensic samples to reference databases, facilitating the identification of unknown human individuals due to its rich polymorphic sites and established utility in forensic identification.

Sanger sequencing, a widely-used DNA sequencing technique, is renowned for its accuracy and reliability, which are crucial in forensic applications where data integrity is paramount. This method involves the selective incorporation of chain-terminating dideoxynucleotides during DNA synthesis, resulting in DNA fragments of various lengths that can be analyzed to determine the nucleotide sequence. In mitochondrial DNA (mtDNA) analysis, Sanger sequencing is frequently employed to sequence the control region, producing high-quality reference sequences essential for comparing forensic samples. However, the advent of Massively Parallel Sequencing (MPS) has revolutionized forensic genetics by providing a high-resolution tool capable of simultaneously analyzing complex genetic markers,

including mtDNA, Y-chromosomal SNPs, and autosomal STRs (Parson et al., 2016). Compared to Sanger sequencing, MPS enhances the efficiency of sequencing and depth the analysis to aid forensic investigations (Goodwin et al., 2017). In this study, Sanger sequencing was carried out to achieve concordance of the MPS technology to create the basis of the research.

The dynamic landscape of forensic science has evolved significantly in recent years, driven by advancements in research, emerging technologies, and strategic shifts to address societal and legal demands (Forensic Science Environmental Scan 2023). Forensic science is characterized by interconnected scientific, societal, and technological developments, which collectively shape how forensic disciplines, such as mtDNA analysis, contribute to the criminal justice system (Swofford, 2024). These evolving trends highlight the importance of integrating robust, evidence-based methodologies, such as MPS, to address contemporary challenges in forensic investigations. My study aligns with these emerging priorities by implementing mtDNA sequencing and providing practical insights into the application of MPS in degraded or low-template samples. This connection underscores how advancements in mtDNA sequencing fit within the broader strategic priorities of forensic science research and its ongoing efforts to standardize and innovate across disciplines.

8.2. Mitochondrial DNA Analysis Techniques

The study utilized both Sanger sequencing (Chapter 3) and Massive Parallel Sequencing (MPS) (Chapter 5). These methods provided complementary insights into mitochondrial DNA (mtDNA) profiling, which is crucial in forensic genetics.

Sanger Sequencing technique was applied to analyze the mtDNA control region (Chapter 3), particularly focusing on optimizing sample extraction, primer validation, and sequencing conditions. Findings showed successful amplification and sequencing, providing a foundational concordance dataset across Emirati, Indian, and Pakistani populations.

Chapter 5 explored the application of MPS to sequence the entire mtDNA genome for 510 Emirati samples, significantly expanding the amount of genetic data available. This approach revealed higher accuracy and deeper insights into haplogroups and heteroplasmic variations, showing full concordance with Sanger results in selected samples. The ability of MPS to capture extensive SNP data and indels enhances its value for forensic and population studies.

The dual-method approach illustrates how MPS can complement and confirm Sanger sequencing results, with MPS offering a more comprehensive view of genetic diversity and reliability in forensic applications.

8.3. Population-Specific Findings in the UAE

The study provided a novel analysis of mtDNA profiles in the Emirati population, documenting both haplotypes and haplogroups that were previously underrepresented in genetic databases.

A unique genetic signatures were found by the MPS analysis in Chapter 5 where specific haplotypes and haplogroups were generated within Emiratis that reflect distinct maternal lineages, likely shaped by historical endogamy and high consanguinity rates in the region. Chapter 5 highlighted a high haplotype diversity within the Emirati population, with many unique haplotypes, indicating a complex genetic history. For example, haplogroup H emerged as the most common, suggesting an ancient genetic presence potentially influenced by migrations across the Arabian Peninsula.

These findings contribute valuable reference data for forensic purposes and increase understanding of the genetic composition within the Arabian Peninsula, enhancing accuracy in identifying Emirati individuals in forensic cases.

8.4. Comparison Across Populations (Emirati, Indian, Pakistani)

By analyzing mtDNA profiles across three populations residing in the UAE (Emirati, Indian, Pakistani), the study aimed to identify the genetic similarities and differences among these groups (Chapter 5).

Chapter 5 demonstrated that while certain haplotypes were shared across populations, Emiratis showed unique genetic markers that were less prevalent or absent in Indian and Pakistani groups. This indicates distinct maternal lineage patterns.

Chapter 6 reported distinct haplogroup frequencies between populations, with some groups (e.g., L haplogroup) being more common in specific populations, reflecting historical migration and regional ancestry patterns.

The comparative analysis enriches regional databases, helping differentiate individuals of Emirati descent from expatriate populations, a valuable aspect in population-based forensic analysis.

8.5. Forensic Casework Samples

Several casework samples were studied and presented to illustrate the effectiveness of MPS in real forensic scenarios. These examples demonstrated how MPS could resolve complex cases that might have been challenging with traditional autosomal analysis as well as Sanger sequencing alone, such as identifying degraded samples and distinguishing between closely related individuals (Just et al., 2015). The findings emphasized the practical implications of adopting MPS in forensic laboratories. Given its high concordance with traditional methods and its superior capabilities in terms of coverage and heteroplasmy detection, MPS is positioned as a valuable tool for forensic mtDNA analysis. One significant advantage of MPS over Sanger sequencing is its higher coverage and depth of sequencing reads. MPS provides extensive data for each nucleotide position, which enhances the ability to detect heteroplasmy and low-level variants. High coverage also contributes to the robustness of the results, reducing the likelihood of missing variants present at low frequencies. MPS identified heteroplasmic sites with greater sensitivity compared to Sanger sequencing, which is crucial for comprehensive forensic analysis. Despite these advantages, the implementation of MPS is not without challenges. Its high sensitivity can lead to the detection of low-level contaminants, complicating the interpretation of results, especially in cases involving degraded or mixed samples (Tillmar

et al., 2019). Additionally, the cost of MPS equipment and reagents, along with the extended workflow time of up to four days, presents barriers to its widespread adoption in forensic laboratories (Amorim et al., 2019; Butler et al., 2015). Furthermore, MPS requires a higher level of expertise in both wet-lab techniques and bioinformatics, emphasizing the need for specialized training and skilled personnel (Wilson et al., 2016). To optimize its use, batch processing of multiple samples is essential, maximizing the utility of sequencing runs and reducing per-sample costs (Budowle et al., 2014).

The potential effectiveness of using MPS for crime scene data lies in its ability to generate detailed mitotypes from samples where previously little information could be obtained, such as highly degraded or low-quality specimens. Also, in most classical uses of mtDNA in missing persons or maternal kinship analysis, specially in cases of no parents to compare with. Adopting MPS in the context of aiding the law enforcement is a future strategic plan within our local facility to serve as a regional consultancy laboratories to aid humanitarian work in a MOU with ICRC. The plan is to setup readiness of the capability of carrying out MPS for mtDNA for DVI work.

8.6. Challenges in mtDNA Sequencing and Heteroplasmy

Throughout Chapters 3 and 5, challenges related to mtDNA sequencing were addressed, particularly the issues of heteroplasmy and nuclear mitochondrial DNA segments (NUMTs).

The study encountered heteroplasmic variations in some samples, which are known to complicate mtDNA interpretation. MPS, with its deeper sequencing capability, provided

more sensitive detection of these variations, making it possible to identify and distinguish between minor variants.

To address the risk of NUMTs interfering with mtDNA analysis, optimized primers were used in Sanger sequencing (Chapter 3) to minimize the inclusion of NUMTs. NUMTs interference was further addressed and strengthened by MPS, which allowed better resolution and exclusion of NUMT sequences.

By accounting for heteroplasmy and NUMTs, the study improved the reliability of mtDNA sequencing, supporting the accurate interpretation needed for forensic applications.

8.7. Beyond Forensic Significance

Beyond its forensic applications, the study research held broader significance, contributing to the understanding of human migration and population genetics within the Arabian Peninsula.

The established haplotype and haplogroup data provides a valuable reference for forensic labs, aiding in the identification of individuals within the UAE population and supporting casework that involves population-based mtDNA matching.

The study's findings support theories of migration and settlement in the Arabian Peninsula, with specific genetic markers aligning with known historical movements across the region. Haplogroups observed, such as H and L, reflect a mixture of indigenous lineages and external influences, providing insights into the genetic history of the UAE.

Those findings enhanced the study's value, positioning it as both a forensic resource and a contribution to the historical understanding of Arabian Peninsula populations.

The implementation of a high-quality reference database is critical for the effective use of mtDNA in forensic applications, as it allows for accurate haplogroup assignment and comparison against unknown samples (Parson and Dür, 2007). The present project has made significant strides in establishing such a database by incorporating mtDNA haplogroup data from Indian and Pakistani populations residing in the UAE. While this enhances the representation of these communities in the forensic context, a limitation is evident in the diverse haplogroup distribution, as many samples do not reflect native Emirati lineages (Aljasmi et al., 2020). Nonetheless, such a resource is highly beneficial in disaster victim identification (DVI), where a broad reference database can assist in identifying individuals from various backgrounds. For crime scene investigations, the value of a comprehensive and geographically representative database, similar to EMPOP, is underscored, as it ensures accurate interpretation of mtDNA data and strengthens the reliability of forensic conclusions (Parson et al., 2014).

This study represents a novelty to be the first of its kind in the UAE, sequencing whole mtDNA genome for a sample size of 510 individuals to establish a comprehensive local mtDNA database. The high number of samples allowed for more identification of unique haplotypes and haplogroups, providing additional strength from previous studies that were limited by smaller sample sizes and diversity (Aljasmi et al., 2020). This research significantly contributes to the understanding of the genetic landscape of the UAE, providing a valuable resource for future forensic and population genetics studies.

8.8. Human population genetics

Findings reveal a significant level of haplotype diversity within the Emirati population, shaped by both endogamy and consanguinity, as well as extensive interactions among populations across the Arabian Peninsula. Unique haplotypes and specific haplogroups, such as various subclades of haplogroup H, highlight the distinct maternal lineages within this population. The diversity within haplogroup H, alongside other detected haplogroups, offers a window into the ancestral backgrounds influenced by both indigenous lineages and historical migrations.

The prevalence of certain haplogroups suggests ancient connections to population movements across the region. For instance, haplogroups typically linked to Eurasian ancestry, such as H and HV, support historical connections to migrations from Europe and the Near East. African-specific haplogroups, like L, point to gene flow from Africa, reflecting the complex historical interactions likely facilitated by trade routes and regional migrations.

When comparing mtDNA profiles across Emirati, Indian, and Pakistani groups, distinct haplogroup distributions emerge, reflecting varied ancestral histories. The Emirati population displayed a unique haplogroup profile, with a higher frequency of haplogroups H and HV, whereas Indian and Pakistani groups showed greater representation of haplogroups such as M and R, which are more prevalent in South Asian populations. These distributions align with the distinct historical and migratory backgrounds of each group.

Shared genetic markers across these populations suggest some degree of genetic overlap, influenced by the UAE's historical role as a central trade hub. The presence of shared haplotypes between Emirati and South Asian individuals indicates gene flow resulting from migrations and long-standing economic exchanges between the Arabian Peninsula and South Asia. This shared genetic background provides important context for understanding both historic and modern population structures within the UAE.

The detection of heteroplasmic variations in several samples adds an evolutionary perspective on mutation rates within these populations. Observed heteroplasmic sites, varying in frequency among individuals, offer clues about the population's age, genetic stability, and recent mutations. This variability contributes valuable data for both anthropological and forensic analyses.

The prevalence of certain haplogroups, such as H and R, among Emiratis and the relatively lower presence of these haplogroups in South Asian groups within the UAE, support theories of distinct population expansions and adaptations. Haplogroup H, for instance, is associated with prehistoric migration events from Africa to Europe and the Near East, whereas haplogroup R's prevalence in South Asian populations underscores divergent evolutionary paths, shaped by long-term settlement patterns.

By comparing mtDNA sequences to global databases, the study situates Emirati genetics within a broader phylogenetic framework, showing connections to ancient populations in the Arabian Peninsula, as well as influences from migration routes involving Africa, Europe, and Asia. For instance, haplogroups such as U and T, though less common, add

layers of ancestry that link Emiratis with ancient Near Eastern populations, suggesting interactions along migration corridors that date back thousands of years.

Additionally, evidence of substructures within the Emirati population—where certain haplogroups cluster in specific regions—may be attributed to historical tribal affiliations and limited gene flow among communities. Such population substructure is essential in population genetics as it can affect genetic diversity and influence analyses of specific genetic markers across the UAE.

The inclusion of comprehensive mtDNA data for Emirati, Indian, and Pakistani populations addresses a gap in population genetics research for the Arabian Peninsula. The findings expand existing databases by providing unique haplotypes and diverse haplogroups that reflect the complex history and interactions among these populations, enhancing the utility of forensic and anthropological databases.

These findings also contribute to broader evolutionary and migration models, supporting evidence of both historic and recent gene flow within the UAE. The data presented can be leveraged to refine models of human migration in the Arabian Peninsula, highlighting the region's role as a genetic intersection linking Africa, Asia, and Europe. This study's expanded genetic data is instrumental in both forensic identification efforts and the exploration of human population dynamics in this historically rich region.

8.9. Conclusions

This study provides substantial contributions to the fields of forensic genetics and human population genetics, particularly within the Arabian Peninsula. Through a combined

approach utilizing Sanger sequencing and Massive Parallel Sequencing (MPS), this research has generated a detailed genetic profile of the Emirati population, while also offering comparative insights into the Indian and Pakistani groups residing in the UAE. The findings underscore the unique genetic landscape of the UAE, where indigenous lineages are interwoven with traces of ancient migration patterns and recent gene flow from neighboring regions.

One of the most significant outcomes of the study is the identification of unique haplotypes and high haplotype diversity among Emiratis, which reflect a rich genetic history shaped by endogamy, consanguinity, and the UAE's role as a historical trade hub. The prevalence of specific haplogroups, such as H and HV, aligns with known migration routes from Africa, Europe, and Asia, shedding light on the ancestral origins and maternal lineages that are prevalent in the region. In addition, the study's comparative analysis with Indian and Pakistani populations highlights the genetic distinctions and commonalities shaped by centuries of migration, trade, and cultural exchange. Shared genetic markers between Emirati and South Asian groups illustrate the extent of gene flow, contributing to a nuanced understanding of regional population dynamics.

The application of both Sanger sequencing and MPS methods not only validated the reliability of mtDNA analysis but also highlighted the importance of MPS for capturing heteroplasmic sites and subtle variations, which are valuable for forensic applications. By accounting for technical challenges such as heteroplasmy and the presence of nuclear mitochondrial DNA segments (NUMTs), the study enhances the precision of mtDNA

analysis in forensic casework, establishing a robust reference for future forensic investigations within the UAE and beyond. NUMTs are a known source of error in forensic sequencing, requiring stringent bioinformatics filtering to avoid misinterpretation. NUMTs were filtered out using the TSS.

This work contributes valuable genetic data to the growing body of knowledge on Middle Eastern populations, providing a foundation for future research and supporting the development of more comprehensive regional genetic databases. The insights gained have broad implications, not only for improving the accuracy of forensic identifications but also for advancing knowledge on the genetic diversity and historical interactions that define the UAE's population.

Overall, the present study concluded that MPS is a highly reliable and effective method for mtDNA analysis in forensic science. The high concordance rates with Sanger sequencing, combined with the additional benefits of deeper coverage and better heteroplasmy detection, support its adoption in forensic laboratories. The integration of MPS data into reference databases, such as EMPOP, supports robust statistical calculations and enhances the interpretation of forensic evidence, further solidifying its role as a transformative tool in forensic science. Based on the study, further research into optimizing MPS protocols for forensic applications is recommended.

The integration of MPS data into reference databases, such as EMPOP, supports robust statistical calculations and enhances the interpretation of forensic evidence, further solidifying its role as a transformative tool in forensic science.

While the adoption of MPS is widely accepted, the decision to carry out whole mtDNA genome sequencing versus CR sequencing remains debatable until high-quality local databases are fully established. Building such a comprehensive database using whole mtDNA genome sequencing would be highly beneficial, as it provides a more complete and detailed reference. Once a robust database is in place, future work could be flexibly adapted between CR and whole genome sequencing based on the specific sample type or case requirements, ensuring both efficiency and accuracy in forensic analysis. This study also evaluated sequencing accuracy across multiple bioinformatics pipelines, including Torrent Suite™ Software (TSS), and Converge™ Software, ensuring robust variant detection. The high concordance between different analysis pipelines reinforces the reliability of MPS for forensic applications.

8.10. Scope for Future Studies

For future work, studies are ongoing, covering more aspects to include sensitivity, contamination, mixture studies, and inhibition. The panel was evaluated for its capability to handle mixed samples, allowing for the detection and interpretation of mitochondrial DNA from multiple contributors. This is essential for forensic cases involving mixed DNA sources. Mixture interpretation will be further investigated in a separate work. The panel's performance will be tested in the presence of common forensic inhibitors. This assessment ensures that the panel can still function effectively even when substances that inhibit DNA amplification are present and measures its resistance to inhibitors. These studies will further confirm the robustness of the Precision ID mtDNA Whole Genome Panel.

is to, sensitive, and capable of handling a wide range of challenging forensic samples (Holland and Parsons, 2013; Just et al., 2015), thereby making it suitable for routine forensic casework. Additionally, sequencing mother-child samples to explore differences in their mtDNA profiles will be incorporated as a future research direction. This will help to understand mutation rates, inheritance patterns, and low-level heteroplasmy between generations, providing more accurate interpretations of mtDNA profiles. Such work will be instrumental in developing guidelines for the inclusion of related individuals in population databases to avoid skewed haplotype frequency estimations.

The analysis identified various types of errors that could occur during sequencing, such as point heteroplasmy, length heteroplasmy, insertion-deletion mutations (indels), and misreads of homopolymeric tracts. Both homopolymeric tracts and length heteroplasmy can complicate forensic analyses by introducing variability in mtDNA sequences that must be carefully interpreted. The sequencing technology demonstrated how the technologies handle homopolymeric regions. Recent studies, such as Zhang et al. (2024), stress the necessity of validating sequencing results across multiple bioinformatics platforms to improve forensic accuracy.

8.11. Limitations of this Study

MPS technology can reveal informative mtDNA sequences even from trace evidence, providing valuable data that can help identify individuals or establish maternal lineage connections. However, one significant limitation is the difficulty in resolving mixed samples, as MPS is highly sensitive and can detect multiple contributors within a single

sample, complicating the interpretation (Tillmar et al., 2019). In this study, mixtures were not observed, but this issue remains a concern in forensic casework, especially with touch DNA, where determining the origin of the material—who has left it—becomes challenging. This is particularly problematic when there is no suspect in mind, as mtDNA lacks the discrimination power of autosomal STRs and there are currently no comprehensive databases like CODIS for mtDNA comparisons (Parson et al., 2016). Consequently, mtDNA data are more useful when there is a specific suspect to compare against rather than for general database searches. Despite these limitations, there is a growing trend among forensic laboratories to adopt MPS technology for mtDNA analysis due to its ability to provide high-resolution data and improve the investigation of complex cases (Budowle et al., 2018). This adoption indicates a shift towards utilizing MPS to enhance forensic capabilities, particularly in cases where traditional STR analysis may not be sufficient.

Despite its significant findings, this study encountered several limitations that should be acknowledged. These limitations highlight the complexities and challenges inherent in forensic genetic research and population genetics within the Arabian Peninsula.

- a. Lack of Standardized Thresholds for Mixture and Data Interpretation: One of the main challenges was the absence of universally accepted thresholds for interpreting mixed mtDNA samples, particularly in cases with low-level heteroplasmy. This lack of standardization can introduce variability in data

interpretation and limit comparability with other studies, particularly in forensic applications where precise interpretation is critical.

- b. Interpretation of Sequence Artifacts: Sequence artifacts, which can arise from technical errors in sequencing processes, posed challenges during data analysis. Artifacts can mimic genuine mutations, complicating the accurate identification of haplotypes and haplogroups. Efforts were made to distinguish artifacts from real sequence variations, but the risk of misinterpretation remains, especially in highly degraded samples.
- c. Population-Specific Haplogroup Assignments: Although haplogroup assignment provided valuable insights, certain haplogroups are underrepresented in global databases, especially those specific to the Arabian Peninsula. This limitation complicates accurate classification, as the available reference data may not fully capture the diversity within Emirati or other regional populations, impacting the precision of phylogenetic and evolutionary interpretations.
- d. Challenges with Degraded Samples and Bone Analysis: The study encountered limitations with degraded samples, as mtDNA analysis of aged or environmentally compromised samples often yields incomplete or low-quality data. Although MPS is more sensitive than traditional methods, the sequencing of ancient bone or highly degraded samples remains challenging, affecting the study's ability to generate comprehensive data from all sample types.

- e. Haplotype Sharing and Overlap: Shared haplotypes across individuals and populations can complicate individual identification in forensic contexts. The presence of common haplotypes, especially in populations with a history of endogamy, limits the discriminatory power of mtDNA for distinguishing individuals. This limitation underscores the importance of using mtDNA analysis in conjunction with other forensic markers.
- f. Sample Representativeness: While the study aimed to provide a representative sample of the UAE population, the genetic diversity within the Emirates may still be underrepresented. Certain subpopulations and tribal affiliations may have unique genetic characteristics that were not fully captured, potentially impacting the comprehensiveness of the findings. Additionally, the inclusion of expatriate populations, though valuable for comparative purposes, may not fully reflect the diversity within each represented group.

8.12. Recommendations

As forensic genetics continues to evolve, ethical concerns surrounding the use of genetic data for law enforcement purposes have gained prominence. Brown et al. (2024) emphasize how the practice of accessing clinical genetic data without consent raises privacy concerns and undermines public trust in both healthcare and law enforcement institutions. They recommend legal reforms to ensure transparency and proper consent mechanisms, which are essential for maintaining public confidence in forensic applications.

Integrating such reforms into forensic workflows could help address privacy issues while ensuring effective use of genetic databases.

Based on the findings and limitations of this study, several recommendations can be made to advance the field of human population genetics in the UAE and beyond. The study by Taylor et al. (2020) serves as a pivotal reference, showcasing the generation of a robust dataset comprising 1327 platinum-quality mtDNA haplotypes from diverse U.S. populations. This advancement significantly bolsters the forensic community's capacity to conduct precise and reliable mtDNA analyses. The implementation of rigorous QC measures and comprehensive population analyses provides a valuable resource for the forensic community, enhancing the accuracy and reliability of mtDNA analyses in forensic casework. Such recommendation can serve a crucial basis for the expansion of the United Arab Emirates forensic-quality mitogenome haplotypes.

- a. Expand the Population Database: Establishing a comprehensive UAE-specific genetic database would not only improve forensic accuracy within the region but also contribute valuable genetic data to the global population genetics field. By including diverse subpopulations, tribal groups, and expatriate communities, the database would offer a more nuanced representation of the genetic landscape in the UAE. This would enhance forensic identifications and improve understanding of local genetic diversity, which is critical for anthropological studies on migration and ancestry within the Arabian Peninsula.

- b. **Implement Standard Protocols for Mixture Interpretation:** Creating standardized thresholds and protocols for mixture interpretation would bring consistency and reliability to forensic analyses involving mtDNA. By establishing clear guidelines for detecting low-level heteroplasmy and analyzing mixed mtDNA samples, forensic labs could reduce variability in results. This would improve data comparability across studies and jurisdictions, ensuring that forensic conclusions remain robust and that legal outcomes based on mtDNA evidence are more reliable.
- c. **Create a Scalable Bioinformatics Pipeline:** Developing a bioinformatics pipeline that is capable of handling high-throughput sequencing data would streamline the processing and analysis of complex mtDNA datasets. Such a pipeline would automate the identification of sequence artifacts, optimize haplogroup assignment accuracy, and facilitate mixture and heteroplasmy detection. By making the analysis process more efficient and scalable, forensic labs could reduce turnaround times, improve data quality, and adapt to growing data volumes, ultimately enhancing forensic casework and research.
- d. **Establish Quality Control (QC) Measures for Sequence Errors:** Implementing rigorous QC protocols to identify and manage sequence artifacts would significantly increase the accuracy of mtDNA analyses. These QC measures would involve verifying mutations, minimizing false positives, and ensuring consistency in haplotype assignments. In forensic applications, this would strengthen the

reliability of mtDNA results, particularly in cases involving degraded samples, where sequencing errors are more likely. Enhanced QC standards would also support the integrity of genetic databases by reducing error rates and improving data quality.

- e. **Encourage International Collaboration for Regional Databases:** Collaboration with international databases and institutions would allow for broader comparisons and enhance genetic diversity representation. By sharing data with populations from surrounding regions, UAE researchers could contextualize their findings within the Arabian Peninsula's broader genetic history, facilitating insights into gene flow, migration patterns, and ancestral lineages. Such collaborations would also help build robust regional databases, improving the reliability of forensic identifications in multi-national cases. The utility of national and international forensic DNA databases has been critical in resolving forensic casework. As evidenced by the Forensic Information Databases Annual Report (2024), advancements in database management have facilitated cross-referencing of casework samples with existing profiles, enhancing case resolution rates.
- f. **Invest in Training and Resources for Forensic and Genetic Specialists:** Expanding training and resources for forensic and genetic specialists would build local expertise in mtDNA analysis, bioinformatics, and forensic genetics. With increased knowledge and skill, professionals in the UAE could implement advanced technologies like MPS more effectively. This investment in human resources would

support the reliability and accuracy of forensic and population genetic studies, helping ensure that the UAE remains at the forefront of scientific advancements in these fields.

References

- Abed, I. & Hellyer, P. (2001). United Arab Emirates: A New Perspective. Trident Press.
- Abu-Amero, K.K., Cabrera, V.M., Larruga, J.M., Osman, E.A., González, A.M. & Al-Obeidan, S.A. (2011). Eurasian and Sub-Saharan African mitochondrial DNA haplogroup influences pseudoexfoliation glaucoma development in Saudi patients. *Molecular Vision*, 17: 543-547.
- Abu-Amero, K.K., Larruga, J.M., Cabrera, V.M. & González, A.M. (2008). Mitochondrial DNA structure in the Arabian Peninsula. *BMC Evolutionary Biology*, 8: Article 45.
- Al Rawi, S., Louvet-Vallée, S., Djeddi, A., Sachse, M., Culetto, E., Hajjar, C., Boyd, L., Legouis, R. & Galy, V. (2011). Postfertilization autophagy of sperm organelles prevents paternal mitochondrial DNA transmission. *Science*, 334(6059): 1144–1147.
- Ali Alhmoudi, O., Jones, R.J., Tay, G.K., Alsafar, H. & Hadi, S. (2015). Population genetics data for 21 autosomal STR loci for United Arab Emirates (UAE) population using next generation multiplex STR kit. *Forensic Science International: Genetics*, 19: 190-191.
- Aljasmi, F.A., Vijayan, R., Sudalaimuthuasari, N., Souid, A.-K., Karuvantevida, N., Almaskari, R., Abdul Kader, H.M., Kundu, B., Hazzouri, K.M. & Amiri, K.M.A. (2020). Genomic landscape of the mitochondrial genome in the United Arab Emirates native population. *Genes*, 11(8): Article 876.
- Almeida, M., Betancor, E., Fregel, R., Suárez, N.M. & Pestano, J. (2011). Efficient DNA extraction from hair shafts. *Forensic Science International: Genetics Supplement Series*, 3(1): e319-e320.
- AlSafar, H.S., Al-Ali, M., Daw Elbait, G., Al-Maini, M.H., Ruta, D., Peramo, B., Henschel, A. & Tay, G.K. (2019). Introducing the first whole genomes of nationals from the United Arab Emirates. *Scientific Reports*, 9: Article 14725.
- Alshamali, F., Brandstätter, A., Zimmermann, B. & Parson, W. (2008). Mitochondrial DNA control region variation in Dubai, United Arab Emirates. *Forensic Science International: Genetics*, 2(1): e9-e10.
- al-Tabari, I. J. (1987). The history of al-Tabari, Volume IV: The ancient kingdoms (F. Rosenthal, Trans.). Albany, NY: State University of New York Press. (Original work published ca. 915 CE)
- Alvarez-Iglesias, V., Jaime, J.C., Carracedo, Á. & Salas, A. (2007). Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Science International: Genetics*, 1(1): 44-55.

- Amorim, A., Fernandes, T., and Taveira, N. (2019). Mitochondrial DNA in human identification: a review. *PeerJ*, 7: Article e7314.
- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R. & Young, I.G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290: 457-465.
- Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C., Podowski, R.M. & El-Sayed, N.M. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 396(6707): 133-140.
- Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23: Article 147.
- Arnheim, N. & Cortopassi, G. (1992). Deleterious mitochondrial DNA mutations accumulate in aging human tissues. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 275(3-6): 157-167.
- Asari, M., Azumi, J., Shimizu, K., and Shiono, H. (2008). Differences in tissue distribution of HV2 length heteroplasmy in mitochondrial DNA between mothers and children. *Forensic Science International*, 175(2-3): 155-159.
- Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., and Saunders, N.C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, 18: 489-522.
- Ballard, D., Winkler-Galicki, J., and Wesoły, J. (2020). Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *International Journal of Legal Medicine*, 134(4): 1291-1303.
- Bandelt, H. J., Lahermo, P., Richards, M., and Macaulay, V. (2001). Detecting errors in mtDNA data by phylogenetic analysis. *International Journal of Legal Medicine*, 115(2): 64-69.
- Bandelt, H.J., Kong, Q.P., Parson, W. & Salas, A. (2005). More evidence for non-maternal inheritance of mitochondrial DNA? *Journal of Medical Genetics*, 42(12): 957-960.
- Bandelt, H.J., and Parson, W. (2008). Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *International Journal of Legal Medicine*, 122(1): 11-21.

- Bandelt, H.J., Salas, A., and Lutz-Bonengel, S. (2004). Artificial recombination in forensic mtDNA population databases. *International Journal of Legal Medicine*, 118(5): 267-273.
- Bandelt, H.J., van Oven, M., and Salas, A. (2012). Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *International Journal of Legal Medicine*, 126(6): 901-916.
- Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Holland, M., Lincoln, P.J., Mayr, W., Morling, N., Olaisen, B., Schneider, P.M., Tully, G., and Wilson, M. (2000). Guidelines for mitochondrial DNA typing. DNA Commission of the International Society for Forensic Genetics. *Vox Sanguinis*, 79(2): 121-125.
- Barnabas, S., Shouche, Y., and Suresh, C.G. (2006). High-resolution mtDNA studies of the Indian population: implications for Palaeolithic settlement of the Indian subcontinent. *Annals of Human Genetics*, 70(Pt 1): 42-58.
- Bauer, C.M., Bodner, M., Niederstätter, H., Niederwieser, D., Huber, G., Hatzner-Grubwieser, P., Holubar, K., and Parson, W. (2013). Molecular genetic investigations on Austria's patron saint Leopold III. *Forensic Science International: Genetics*, 7(2): 313-315.
- Bauer, C. M., Niederstätter, H., McGlynn, G., Stadler, H. & Parson, W. (2013). Comparison of morphological and molecular genetic sex-typing on mediaeval human skeletal remains. *Forensic Science International: Genetics*, 7(6): 581-586.
- Bauer, T. (2010). The relevance of early Arabic poetry for Qur'anic studies including observations on kull and on Q 22:27, 26:225, and 52:31. In A. Neuwirth, N. Sinai, & M. Marx (Eds.), *The Qur'an in Context*. Brill, Leiden, The Netherlands: 699-732.
- Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N.M., Kivisild, T., Torroni, A. & Villems, R. (2012). A Copernican Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am J Hum Genet*, 90, 675-684.
- Bendall, K.E. & Sykes, B.C., (1995). Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *American Journal of Human Genetics*, 57, 248-256.
- Berger, C. & Parson, W. (2009). Mini-midi-mito: adapting the amplification and sequencing strategy of mtDNA to the degradation state of crime scene samples. *Forensic Science International: Genetics*, 3(3): 149-153.
- Bhacker, M.R. and Bhacker, B. (1997). *Digging in the Land of Magan*. *Archaeology*, 50(3).
- Bodner, M., Iuvare, A., Strobl, C., Nagl, S., Huber, G., Pelotti, S., Pettener, D., Luiselli, D. & Parson, W. (2015). Helena, the hidden beauty: Resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample. *Forensic Science International: Genetics*, 15, 21-26.

- Bogenhagen, D. & Clayton, D. A. (1974). The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. *The Journal of Biological Chemistry*, 249(24), 7991–7995.
- Borst, P. & Grivell, L.A. (1978). *The mitochondrial genome of yeast*. *Cell*, 15(3), 705-723.
- Boxma, B., et al., 2005. An anaerobic mitochondrion that produces hydrogen. *Nature*, 434, 74-79.
- Børsting, C. & Morling, N., 2015. Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics*, 18, 78-89.
- Brandhagen, M.D., Loreille, O. and Irwin, J.A., 2018. Fragmented Nuclear DNA is the Predominant Genetic Material in Human Hair Shafts. *Genes (Basel)*, 9(12), Article 640.
- Brandstatter, A., Klein, R., Duftner, N., Wiegand, P. & Parson, W., 2006. Application of a quasi-median network analysis for the visualization of character conflicts to a population sample of mitochondrial DNA control region sequences from southern Germany (Ulm). *International Journal of Legal Medicine*, 120(5), 310-314.
- Brandstätter, A., Niederstätter, H., Pavlic, M., Grubwieser, P. & Parson, W. (2007). Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. *Forensic Science International*, 166(2-3): 164–175.
- Brandstatter, A., Peterson, C. T., Irwin, J. A., & Parsons, T. J. (2004b). DNA sequence diversity of the first hypervariable segment of the mitochondrial control region in 105 individuals from Humboldt County, Nevada. *Forensic Science International*, 139(2-3), 173-179.
- Boxma, B., et al. (2005). An anaerobic mitochondrion that produces hydrogen. *Nature*, 434: 74-79.
- Børsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics*, 18: 78-89.
- Brandhagen, M.D., Loreille, O., and Irwin, J.A. (2018). Fragmented nuclear DNA is the predominant genetic material in human hair shafts. *Genes (Basel)*, 9(12): Article 640.
- Brandstätter, A., Klein, R., Duftner, N., Wiegand, P., and Parson, W. (2006). Application of a quasi-median network analysis for the visualization of character conflicts to a population sample of mitochondrial DNA control region sequences from southern Germany (Ulm). *International Journal of Legal Medicine*, 120(5): 310-314.
- Brandstätter, A., Niederstätter, H., Pavlic, M., Grubwieser, P., and Parson, W. (2007). Generating population data for the EMPOP database—an overview of the mtDNA

- sequencing and data evaluation processes considering 273 Austrian control region sequences as an example. *Forensic Science International*, 166(2-3): 164-175.
- Brandstätter, A., Peterson, C.T., Irwin, J.A., and Parsons, T.J. (2004). DNA sequence diversity of the first hypervariable segment of the mitochondrial control region in 105 individuals from Humboldt County, Nevada. *Forensic Science International*, 139(2-3): 173-179.
 - Brandstätter, A., Peterson, C. T., Irwin, J. A., Mpoke, S., Koech, D. K., Parson, W. & Parsons, T. J. (2004). Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *International Journal of Legal Medicine*, 118(5): 294–306.
 - Brandstätter, A., Sängler, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., Kong, Q.P., Bravi, C.M. & Bandelt, H.J. (2005). Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis*, 26(18): 3414–3429.
 - Breiting, F., Hilgert, J.-N., Hargreaves, C., Sheppard, J., Overdorf, R., and Scanlon, M. (2024). DFRWS EU 10-year review and future directions in digital forensic research. *Forensic Science International: Digital Investigation*, 48: Article 301685.
 - Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U. & Paabo, S. (2009). Primer extension capture: targeted sequence retrieval from heavily degraded DNA sources. *Journal of Visualized Experiments*, (31): Article 1573.
 - Bright, J.-A. & Coble, M. (2019). *Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios*. CRC Press.
 - Brown, J.R., Beckenbach, A.T., and Smith, M.J. (1993). Intraspecific DNA sequence variation of the mitochondrial control region of white sturgeon (*Acipenser transmontanus*). *Molecular Biology and Evolution*, 10: 326-341.
 - Brown, T. A. (2010). *Gene cloning and DNA analysis: an introduction*. Wiley-Blackwell.
 - Brown, T., Duensing, S., and Wong, B. (2024). Forensic genetics in the shadows. *Journal of Law and the Biosciences*, Article Isae028.
 - Brown, W. M. (1980). Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proceedings of the National Academy of Sciences*, 77(6), 3605-3609.
 - Brown, W.M., Prager, E.M., Wang, A., and Wilson, A.C. (1982). Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *Journal of Molecular Evolution*, 18(4): 225-239.

- Budowle, B., Allard, M.W., Wilson, M.R., and Chakraborty, R. (2003). Forensics and mitochondrial DNA: applications, debates, and foundations. *Annual Review of Genomics and Human Genetics*, 4: 119-141.
- Budowle, B., Allard, M.W., Wilson, M.R., et al. (1995). Mitochondrial DNA typing: a population study. *Journal of Forensic Sciences*, 40(3): 478-485.
- Budowle, B., van Daal, A., Stray, J. E., Garofano, P., & Power, M. W. (2014). Forensic genomics: Massively parallel sequencing and beyond. *Forensic Science International: Genetics*, 12, 47-57.
- Butler, J.M. (2012). *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press.
- Butler, J.M. (2015) *Advanced Topics in Forensic DNA Typing: Methodology*. San Diego: Academic Press.
- Butler, J. (2011). *Advanced Topics in Forensic DNA Typing: Interpretation*. 2nd ed. Academic Press, 407-408.
- Butler, J.M. (2010). Chapter 16 - Lineage markers: Y chromosome and mtDNA testing. In J.M. Butler (ed.), *Fundamentals of Forensic DNA Typing*. Academic Press, 363-396.
- Butler, J.M. (2012). Chapter 14 - Mitochondrial DNA analysis. In J.M. Butler (ed.), *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, 405-456.
- Butler, J.M., Hill, C.R., Coble, M.D., Kline, M.C., and Duewer, D.L. (2015). New autosomal STR loci for forensic DNA profiling. *Croatian Medical Journal*, 56(3): 178-192.
- Calabrese, C., Gomez-Duran, A., Reyes, A., and Attimonelli, M. (2020). Methods for the identification of mitochondrial DNA variants. *The Human Mitochondrial Genome*, Academic Press: 243-275.
- Canale, L.C., Date-Chong, M., Wallin, J., Sheehan, S., Battaglia, J., Halsing, M., and Cuenca, D. (2025). Enhancement of the Precision ID mitochondrial DNA whole genome system for challenging unidentified human remains. *Genes*, 16(2): Article 119.
- Cann, R.L., Stoneking, M., and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099): 31-36.
- Cardena, M.M.S.G., Ribeiro-dos-Santos, Â., Santos, S., Mansur, A.J., Pereira, A.C., et al. (2013). Assessment of the relationship between self-declared ethnicity, mitochondrial haplogroups, and genomic ancestry in Brazilian individuals. *PLOS ONE*, 8(4): Article e62005.
- Carracedo, A. (ed.) (2005). *Forensic DNA Typing Protocols*. Totowa, NJ: Humana Press, 3-20.

- Carracedo, A., Bär, W., Lincoln, P., Mayr, W., Morling, N., Olaisen, B., Schneider, P., Budowle, B., Brinkmann, B., Gill, P., Holland, M., Tully, G., and Wilson, M. (2000). DNA commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International*, 110(2): 79-85.
- Carracedo, A., Butler, J. M., Gusmão, L., Linacre, A., Parson, W., Roewer, L., and Schneider, P. M. (2013). New guidelines for the publication of genetic population data. *Forensic Science International: Genetics*, 7(2), 217-220.
- Cavelier, L., Jazin, E., Jalonen, P., Gyllensten, U. (2000). MtDNA substitution rate and segregation of heteroplasmy in coding and noncoding regions. *Human Genetics*, 107, 45–50.
- Chatre, L., and Ricchetti, M. (2013). Large heterogeneity of mitochondrial DNA transcription and initiation of replication exposed by single-cell imaging. *Journal of Cell Science*, 126(4): 914-926.
- Chen, D., and Clark, A. (2018). Mitochondrial DNA selection in human germ cells. *Nature Cell Biology*, 20(2): 118-120.
- Chen, X.J., and Clark-Walker, G.D. (2018). Unveiling the mystery of mitochondrial DNA replication in yeasts. *Mitochondrion*, 38: 17-22.
- Chen, Y.S., Olckers, A., Schurr, T.G., Kogelnik, A.M., Huoponen, K., and Wallace, D.C. (2000). mtDNA variation in the South African Kung and Khwe—and their genetic relationships to other African populations. *American Journal of Human Genetics*, 66(4): 1362-1383.
- Chomyn, A., and Attardi, G. (2003). mtDNA mutations in aging and apoptosis. *Biochemical and Biophysical Research Communications*, 304(3): 519-529.
- Churchill, J.D., Peters, D., Capt, C., Strobl, C., Parson, W., and Budowle, B. (2017). Working towards implementation of whole genome mitochondrial DNA sequencing into routine casework. *Forensic Science International: Genetics Supplement Series*, 6: e388-e389.
- Churchill, J.D., Stoljarova, M., King, J.L., and Budowle, B. (2018). Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples. *International Journal of Legal Medicine*, 132: 1263-1272.
- Cilhar, J.C., Amory, C., Lagacé, R., Roth, C., Parson, W. and Budowle, B. (2020). Developmental validation of a MPS workflow with a PCR-based short amplicon whole mitochondrial genome panel. *Genes*, 11(11): Article 1345.
- Coble, M.D., Loreille, O.M., Wadhams, M.J., Edson, S.M., Maynard, K., Meyer, C.E., Niederstätter, H., Berger, C., Berger, B., Falsetti, A.B., Gill, P., Parson, W., and Finelli, L.N.

- (2009). Mystery solved: the identification of the two missing Romanov children using DNA analysis. *PLOS ONE*, 4(3): Article e4838.
- Connell, J.R., Benton, M.C., Lea, R.A., et al. (2022). Pedigree-derived mutation rate across the entire mitochondrial genome of the Norfolk Island population. *Scientific Reports*, 12: Article 6827.
 - Dayama, G., Emery, S.B., Kidd, J.M., and Mills, R.E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, 42(20): 12640-12649.
 - Den Hartog, B.K., and Elling, J.W. (2009). Clustering for forensic mitotype quality analysis. *Forensic Science International: Genetics Supplement Series*, 2(1): 317-319.
 - Denaro, M., Blanc, H., and Wallace, D.C. (1981). Mitochondrial DNA polymorphism in the American Indian. *Genetics*, 98(3): 491-501.
 - Dolezal, P., Likic, V., Tachezy, J., and Lithgow, T. (2006). Evolution of the molecular machines for protein import into mitochondria. *Science (New York, N.Y.)*, 313(5785): 314-318.
 - Dream in Arabic (2016). *Adventures in the United Arab Emirates & Beyond*. Available at: <https://dreaminginarabic.wordpress.com/interesting-snippets/consanguineous-marriages/> [Accessed 30 March 2018].
 - Dur, A., Huber, N., and Parson, W. (2021). Fine-tuning phylogenetic alignment and haplogrouping of mtDNA sequences. *International Journal of Molecular Sciences*, 22(11): Article 5747.
 - Dyll, S.D., et al. (2004). Ancient invasions: From endosymbionts to organelles. *Science*, 304: 253-257.
 - Finehout, E., Bonanni, P., Duthie, S., Griffin, W., Khan, Z., Shoemaker, P., Wang, X., Pi, E.F., and Shoemaker, P. (2013). Automated processing of FTA samples. U.S. Department of Justice, National Institute of Justice. Report No. 241347. Available at: NCJRS website.
 - Eduardoff, M., Xavier, C., Strobl, C., Casas-Vargas, A., and Parson, W. (2017). Optimized mtDNA control region primer extension capture analysis for forensically relevant samples and highly compromised mtDNA of different age and origin. *Genes*, 8(10): Article 237.
 - Eichmann, C., and Parson, W. (2008). 'Mitominis': multiplex PCR analysis of reduced size amplicons for compound sequence analysis of the entire mtDNA control region in highly degraded samples. *International Journal of Legal Medicine*, 122(5): 385-388.
 - Elliott, K.S., Haber, M., Daggag, H., Busby, G.B., Sarwar, R., Kennet, D., Petraglia, M., Petherbridge, L.J., Yavari, P., Heard-Bey, F.U., Shobi, B., Ghulam, T., Haj, D., Al Tikriti, A.,

- Mohammad, A., Antony, S., Alyileili, M., Alaydaroos, S., Lau, E., Butler, M., Yavari, A., Knight, J.C., Ashrafian, H., and Barakat, M.T. (2022). Fine-scale genetic structure in the United Arab Emirates reflects endogamous and consanguineous culture, population history, and geography. *Molecular Biology and Evolution*, 39(3): Article msac039.
- Embley, T.M., and Martin, W. (2006). Eukaryotic evolution, changes and challenges. *Nature*, 440: 623-630.
 - EMPOP (2015). *mtDNA Database V3/R11*. Available at: <https://empop.online> [Accessed 20 March 2018].
 - FBI (2024). CODIS—NDIS Statistics. Available at: <https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints/codis/codis-ndis-statistics> [Accessed 11 May 2024].
 - FBI (n.d.). CODIS and NDIS Fact Sheet. Available at: <https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet> [Accessed 17 December 2019].
 - Federico, A., Cardaioli, E., Da Pozzo, P., Formichi, P., and Gallus, G.N. (2012). Genetic bases and clinical manifestations of mitochondrial disorders with a focus on the skeletal muscle. *Handbook of Experimental Pharmacology*, 214: 155-183.
 - Fendt, L., Zimmermann, B., Daniaux, M., and Parson, W. (2009). Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. *BMC Genomics*, 10: Article 139.
 - Fernandez-Vizarra, E., and Zeviani, M. (2021). Mitochondrial disorders of the OXPHOS system. *FEBS Letters*, 595(8): 1062-1106.
 - Finehout, E., Bonanni, P., Duthie, S., Griffin, W., Khan, Z., Shoemaker, P., and Wang, X. (2013). Automated processing of FTA samples. Document No. 241347. Prepared for the U.S. Department of Justice under Award Number 2009-DN-BX-K187. This report has not been published by the U.S. Department of Justice. Available electronically through NCJRS.
 - Firestone, R. (1990). *Journeys in Holy Lands: The Evolution of the Abraham-Ishmael Legends in Islamic Exegesis*. Albany, NY: State University of New York Press.
 - Foley, R.A., and Lahr, M.M. (1997). Mode 3 technologies and the evolution of modern humans. *Cambridge Archaeological Journal*, 7(1): 3-36.
 - Forensic Information Databases Annual Report (2024). *Annual Report on the National DNA Database (NDNAD)*. UK Government Report: April 2023 to March 2024.

- Forensic Science Environmental Scan (2023). *Trends and Future Directions in Forensic Science*. National Institute of Standards and Technology (NIST) IR 8515: 1-60.
- Galton, F. (1892) *Finger Prints*. London: Macmillan & Co.
- Garcia-Bertrand, R., Simms, T., Cadenas, A., and Herrera, R. (2014). United Arab Emirates: Phylogenetic relationships and ancestral populations. *Gene*, 533(1): 411-419.
- Gettings, K.B., Aponte, R.A., Vallone, P.M., and Butler, J.M. (2015). STR allele sequence variation: current knowledge and future issues. *Forensic Science International: Genetics*, 18: 118-130.
- Ghiselli, F., Milani, L., Scali, V., and Passamonti, M. (2007). The *Leptynia hispanica* species complex (Insecta Phasmida): polyploidy, parthenogenesis, hybridization, and more. *Molecular Ecology*, 16(20): 4256-4268.
- Ginther, C., Issel-Tarver, L., and King, M. (1992). Identifying individuals by sequencing mitochondrial DNA from teeth. *Nature Genetics*, 2: 135-138.
- Goodwin, W., Linacre, A., and Hadi, S. (2017). *An Introduction to Forensic Genetics* (2nd ed.). Wiley.
- Goodwin, W., Linacre, A., and Vanezis, P. (1999). The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. *Electrophoresis*, 20(8): 1707-1711.
- Gouveia, N., Brito, P., Bogas, V., Bento, A., Balsa, F., Serra, A., Lopes, V., Sampaio, L., São Bento, M., Cunha, P., and Porto, M. (2017). Massively parallel sequencing of forensic samples using Precision ID mtDNA Whole Genome Panel on the Ion S5™ system. *Forensic Science International: Genetics Supplement Series*, 6: E167-E168.
- Gray, M.W. (1992). The endosymbiont hypothesis revisited. *International Review of Cytology*, 141: 233-357.
- Grubb (2007). *Pakistan India Locator*, image, Wikimedia Commons, original upload date 21 July 2007 [Accessed 08 July 2024].
- Gunnarsdóttir, E.D., Li, M., Bauchet, M., Finstermeier, K., and Stoneking, M. (2011). High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Research*, 21(1): 1-11.
- Gusmão, L., Butler, J., Linacre, A., Parson, W., Roewer, L., Schneider, P.M., and Carracedo, A. (2017). Revised guidelines for the publication of genetic population data. *Forensic Science International: Genetics*, 30: 160 – 163.
- Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41: 95 – 98.

- Hazkani-Covo, E., et al. (2010). Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLOS Genetics*, 6: Article e1000834.
- Heinz, T., Pala, M., Gómez-Carballa, A., Richards, M.B. and Salas, A (2017). Updating the African human mitochondrial DNA tree: Relevance to forensic and population genetics. *Forensic Science International: Genetics*, 27: 156–159.
- Hemanth, K., Tharmavaram, M., and Pandey, G. (2020). *History of Forensic Science*. In *Technology in Forensic Science: Sampling, Analysis, Data and Regulations*.
- Higuchi, R., von Beroldingen, C., Sensabaugh, G., and Erlich, H. (1988). DNA typing from single hairs. *Nature*, 332: 543 – 546.
- Holland, M.M., and Parsons, T.J. (1999). Mitochondrial DNA sequence analysis—validation and use for forensic casework. *Forensic Science Review*, 11(1): 21-50.
- Holland, M.M., Fisher, D.L., Mitchell, L.G., Rodriquez, W.C., Canik, J.J., Merrill, C.R., and Weedn, V.W. (1993). Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War. *Journal of Forensic Sciences*, 38(3): 542-553.
- Holt, C., Walichiewicz, P., Eagles, J., Dauilo, A., Didier, M., Edwards, C., Fleming, K., Han, Y., Hill, T., Li, S., Rensfield, A., Sa, D., and Stephens, K. (2019). Mitochondrial DNA data analysis strategies that inform MPS-based forensic casework implementation. *Forensic Science International: Genetics Supplement Series*, 7(1): 389-391.
- Hoyland, R.G. (2001). *Arabia and the Arabs: From the Bronze Age to the Coming of Islam*. London: Routledge.
- Hoyland, R.G. (2015). *In God's Path: The Arab Conquests and the Creation of an Islamic Empire*. Oxford: Oxford University Press.
- Holland, T. (2012). *In the Shadow of the Sword: The Birth of Islam and the Rise of the Global Arab Empire*. New York: Doubleday.
- Huber, N., Parson, W., and Dur, A. (2018). Next generation database search algorithm for forensic mitogenome analyses. *Forensic Science International: Genetics*, 37: 204-214.
- Hublin, J.-J., Ben-Ncer, A., Bailey, S.E., Freidline, S.E., Neubauer, S., Skinner, M.M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature*, 546(7657): 289-292.
- HumanMitoSeq, n.d. Revised Cambridge Reference Sequence (RCRS) of the human mitochondrial DNA. MITOMAP. [online] Available at: <https://www.mitomap.org/MITOMAP/HumanMitoSeq> [Accessed 9 July 2024].

- Hühne, J., Pfeiffer, H., Waterkamp, K., and Brinkmann, K. (1999). Mitochondrial DNA in human hair shafts—existence of intra-individual differences? *International Journal of Legal Medicine*, 112(3): 172-175.
- Irwin, J.A., Just, R.S., and Parson, W. (2015). Massively parallel mitochondrial DNA sequencing in forensic genetics: principles and opportunities. In *Handbook of Forensic Genetics, Volume 2. Security Science and Technology*. World Scientific (Europe), 293-335.
- Irwin, J.A., Parson, W., Coble, M.D., and Parsons, T.J. (2011). Case studies of the use of mitochondrial DNA sequencing in forensic anthropology. *Croatian Medical Journal*, 52(3): 290-297.
- Ivanov, P.L., Wadhams, M.J., Roby, R.K., Holland, M.M., Weedn, V.W., and Parsons, T.J. (1996). Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nature Genetics*, 12(4): 417-420.
- Jagadeesan, A., Ebenesersdóttir, S.S., Guðmundsdóttir, V.B., Thordardóttir, E.L., Moore, K.H.S., and Helgason, A. (2021). HaploGrouper: a generalized approach to haplogroup classification. *Bioinformatics*, 37(4): 570-572.
- Taanman, J.-W. (1999). The mitochondrial genome: structure, transcription, translation, and replication. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 141(2): 103-123.
- Jeffreys, A.J., Wilson, V., and Thein, S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006): 67-73.
- Jehaes, E., Gilissen, A., Cassiman, J.J., and Decorte, R. (1998). Evaluation of a decontamination protocol for hair shafts before mtDNA sequencing. *Forensic Science International*, 94(1-2): 65-71.
- Jobling, M.A., and Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10): 739-751.
- Just, R.S., Scheible, M.K., Fast, S.A., Sturk-Andreaggi, K., Röck, A.W., Bush, J.M., Higginbotham, J.L., Peck, M.A., Ring, J.D., Huber, G.E., Xavier, C., Strobl, C., Lyons, E.A., Diegoli, T.M., Bodner, M., Fendt, L., Kralj, P., Nagl, S., Niederwieser, D., Zimmermann, B., ... Irwin, J.A. (2015). Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. *Forensic Science International: Genetics*, 14: 141-155.
- Just, R.S., Irwin, J.A., and Parson, W. (2015). Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Science International: Genetics*, 18: 131-139.

- Just, R.S., Scheible, M.K., Fast, S.A., Sturk-Andreaggi, K., Higginbotham, J.L., Lyons, E.A., Bush, J.M., Peck, M.A., Ring, J.D., Diegoli, T.M., Röck, A.W., Huber, G.E., Nagl, S., Strobl, C., Zimmermann, B., Parson, W., and Irwin, J.A. (2014). Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens. *Forensic Science International: Genetics*, 10: 73-79.
- King, J. L., Churchill, J. D., Novroski, N. M. M., Seah, L. H., Budowle, B., & Kaye, D. H. (2018). Understanding and interpreting DNA evidence in a legal context: Forensic DNA technologies beyond short tandem repeats. *Journal of Forensic Sciences*, 63(5), 1234-1245.
- King, J.L., LaRue, B.L., Novroski, N.M., Stoljarova, M., Seo, S.B., Zeng, X., Warshauer, D.H., Davis, C.P., Parson, W., Sajantila, A., and Budowle, B. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Science International: Genetics*, 12: 128-135.
- King, T.E., Fortes, G.G., Balaesque, P., Thomas, M.G., Balding, D., Delser, P.M., Neumann, R., Parson, W., Knapp, M., Walsh, S., Tonasso, L., Holt, J., Kayser, M., Appleby, J., Forster, P., Ekserdjian, D., Hofreiter, M., and Schurer, K. (2014). Identification of the remains of King Richard III. *Nature Communications*, 5: Article 5631.
- Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., ... and Underhill, P.A. (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 172(1): 373-387.
- Kloss-Brandstätter, A., Klein, R., Duftner, N., Wiegand, P., and Parson, W. (2006). Application of a quasi-median network analysis for the visualization of character conflicts to a population sample of mitochondrial DNA control region sequences from southern Germany (Ulm). *International Journal of Legal Medicine*, 120: 310-314.
- Kloss-Brandstätter, A., Weissensteiner, H., Erhart, G., Schäfer, G., Forer, L., Schönherr, S., Pacher, D., Seifarth, C., Stöckl, A., Fendt, L., Sottas, I., Klocker, H., Huck, C.W., Rasse, M., Kronenberg, F., and Kloss, F.R. (2015). Validation of next-generation sequencing of entire mitochondrial genomes and the diversity of mitochondrial DNA mutations in oral squamous cell carcinoma. *PLOS ONE*, 10(8): Article e0135643.
- Kumar, S., Padmanabham, P.B.S.V., Ravuri, R.R., Uttaravalli, K., Koneru, P., Mukherjee, P.A., Das, B., Kotal, M., Xaviour, D., Saheb, S.Y., and Rao, V.R. (2008). The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. *BMC Evolutionary Biology*, 8: Article 230.

- Lapidus, I. M. (2002). A history of Islamic societies (2nd ed.). Cambridge: Cambridge University Press.
- Larsson, G. (2003). Ibn Garcia's Shu'ubiyya letter: Ethnic and theological tensions in medieval al-Andalus. Leiden, Netherlands: Brill.
- Lee, H.Y., Chung, U., Yoo, J.E., Park, M.J., and Shin, K.J. (2004). Quantitative and qualitative profiling of mitochondrial DNA length heteroplasmy. *Electrophoresis*, 25(1): 28-34.
- Lee, H.Y., Kim, N.Y., Park, M.J., Sim, J.E., Yang, W.I., and Shin, K. (2010). DNA typing for the identification of old skeletal remains from Korean War victims. *Journal of Forensic Sciences*, 55: 1422-1429.
- Lee, J.C.I., Tsai, L.C., Yu, Y.J., Lin, C.Y., Linacre, A., and Hsieh, H.M. (2016). Investigation into length heteroplasmy in the mitochondrial DNA control region after treatment with bisulfite. *Journal of the Formosan Medical Association*, 115: 284-287.
- Lee, S.E., Kim, G.E., Kim, H., Chung, D.H., Lee, S.D., and Kim, M.Y. (2023). Comparison of two variant analysis programs for next-generation sequencing data of whole mitochondrial genome. *Journal of Korean Medical Science*, 38(36): Article e297.
- Lodeiro, M.F., Uchida, A., Bestwick, M., Moustafa, I.M., Arnold, J.J., Shadel, G.S., and Cameron, C.E. (2012). Transcription from the second heavy-strand promoter of human mtDNA is repressed by transcription factor A in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17): 6513-6518.
- Luo, S., Valencia, C.A., Zhang, J., Lee, N.C., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., Chen, S.M., Chien, Y.H., Hwu, W.L., Fan, P.C., Wong, L.J., Atwal, P.S., and Huang, T. (2018). Biparental inheritance of mitochondrial DNA in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 115(51): 13039-13044.
- Lutz-Bonengel, S., and Parson, W. (2019). No further evidence for paternal leakage of mitochondrial DNA in humans yet. *Proceedings of the National Academy of Sciences of the United States of America*, 116(6): 1821-1822.
- Lutz, S., Weisser, H.J., Heizmann, J., Pollak, S., and Holland, M.M. (1997). Investigation of the heterogeneity of mitochondrial DNA in a sample with poor PCR amplification. *Forensic Science International*, 87(1): 49-58.
- Lutz, S., Weisser, H.J., Heizmann, J., and Pollak, S. (1997). A third hypervariable region in the human mitochondrial D-loop. *Human Genetics*, 101(3): Article 384.
- Lutz, S., Weisser, H.J., Heizmann, J., and Pollak, S. (1998). Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. *International Journal of Legal Medicine*, 111: 67-77.

- Lutz, S., Wittig, H., Weisser, H.-J., Heizmann, J., Junge, A., Dimo-Simonin, N., Parson, W., Edelmann, J., Anslinger, K., Jung, S., and Augustin, C. (2000). 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*, 113(1-3): 97-101.
- Mabuchi, T., Susukida, R., Kido, A., and Oya, M. (2007). Typing the 1.1 kb control region of human mitochondrial DNA in Japanese individuals. *Journal of Forensic Sciences*, 52(2): 355-363.
- Maca-Meyer, N., González, A.M., Larruga, J.M., Flores, C., and Cabrera, V.M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics*, 2(1): Article 13.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonnét-Tamir, B., Sykes, B., and Torroni, A. (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *American Journal of Human Genetics*, 64(1): 232-249.
- Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., ... and Oppenheimer, S. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724): 1034-1036.
- Malik, S., Sudoyo, H., Pramoonjago, P., Suryadi, H., Sukarna, T., Njunting, M., Sahiratmadja, E., and Marzuki, S. (2002). Nuclear mitochondrial interplay in the modulation of the homopolymeric tract length heteroplasmy in the control (D-loop) region of the mitochondrial DNA. *Human Genetics*, 110(5): 402-411.
- Manfredi, G., Thyagarajan, D., Papadopoulou, L.C., Pallotti, F., Schon, E.A., and Moraes, C.T. (1997). Increased reactive oxygen species production in cells with limited mitochondrial DNA replication. *Journal of Biological Chemistry*, 272(28): 4319-4324.
- Manfredi, G., Thyagarajan, D., Papadopoulou, L.C., Pallotti, F., and Schon, E.A. (1997). The fate of human sperm-derived mtDNA in somatic cells. *American Journal of Human Genetics*, 61(4): 953-960.
- Maras, M.H., and Miranda, M.D. (2014). Encyclopedia of Law and Economics. *Springer*: 1-6.
- Marshall, C., and Parson, W. (2021). Interpreting NUMTs in forensic genetics: seeing the forest for the trees. *Forensic Science International: Genetics*, 53: Article 102497.
- Marshall, C., Sturk-Andreaggi, K., Gorden, E.M., Daniels-Higginbotham, J., Sanchez, S.G., Bašić, Ž., Kružić, I., Anđelinović, Š., Bosnar, A., Čoklo, M., Petaros, A., McMahon, T.P.,

- Primorac, D., and Holland, M.M. (2020). A forensic genomics approach for the identification of Sister Marija Crucifiksa Kozulić. *Genes*, 11(8): Article 938.
- Martin, W., and Mentel, M. (2010). The origin of mitochondria. *Nature Education*, 3(9): Article 58.
 - McCrery, N. (2013) Silent Witnesses: The Often Gruesome but Always Fascinating History of Forensic Science. Chicago: Chicago Review Press.
 - McElhoe, J.A., Holland, M.M., Makova, K.D., Su, M.S.-W., Paul, I.M., Baker, C.H., Faith, S.A., Young, B., and Long, J.R. (2024). A new tool for probabilistic assessment of MPS data associated with mtDNA mixtures. *Genes*, 15(1): Article 194.
 - McEwen, J.E. (1995). Forensic DNA data banking by state crime laboratories. *American Journal of Human Genetics*, 56: 1487-1492.
 - Melton, T., Dimick, G., Higgins, B., Lindstrom, L., and Nelson, K. (2005). Forensic mitochondrial DNA analysis of 691 casework hairs. *Journal of Forensic Sciences*, 50(1): 73-80.
 - Melton, T., Holland, C., and Holland, M. (2012). Forensic mitochondrial DNA analysis: current practice and future potential. *Forensic Science Review*, 24(2): 101-122.
 - Mercier, N., Valladas, H., Bar-Yosef, O., Vandermeersch, B., Stringer, C., and Joron, J.L. (1993). Thermoluminescence date for the Mousterian burial site of Es Skhul, Mt. Carmel. *Journal of Archaeological Science*, 20(2): 169-174.
 - Milani, L., and Ghiselli, F. (2017). The inheritance of viable mitochondria. *PeerJ Preprints*, 10.7287/peerj.preprints.3122v1.
 - MITOMAP, 2018. <https://www.mitomap.org/foswiki/bin/view/MITOMAP/WebHome> visited: 20/03/2018.
 - Mohammed, A.A., Linacre, A.M., Vanezis, P., and Goodwin, W. (2001). STR data for the GenePrint PowerPlex 1.2 system loci from three United Arab Emirates populations. *Forensic Science International*, 119(3): 328-329.
 - Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51 Pt 1: 263-273.
 - Nakahara, H., Sekiguchi, K., Imaizumi, K., Mizuno, N., and Kasai, K. (2008). Heteroplasmies detected in an amplified mitochondrial DNA control region from a small amount of template. *Journal of Forensic Sciences*, 53(2): 306-311.

- Naue, J., Hörer, S., Sängler, T., Strobl, C., Hatzler-Grubwieser, P., Parson, W., and Lutz-Bonengel, S. (2015). Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. *Mitochondrion*, 20: 82-94.
- Naue, J., Xavier, C., Hörer, S., Parson, W., and Lutz-Bonengel, S. (2024). Assessment of mitochondrial DNA copy number variation relative to nuclear DNA quantity between different tissues. *Mitochondrion*, 74: Article 101823.
- Nebel, A. (2002). Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. *American Journal of Human Genetics*, 70(6): 1594-1596.
- Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97(1): 145-163.
- Nei, M., and Roychoudhury, A.K. (1993). Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution*, 10(5): 927-943.
- Nevo, Y.D., and Koren, J. (1999). *The Origins of the Arab Religion and the Arab State: Crossroads to Islam*. Prometheus Books.
- Niederstätter, H., Köchl, S., Grubwieser, P., Pavlic, M., Steinlechner, M., and Parson, W. (2007). A modular real-time PCR concept for determining the quantity and quality of human nuclear and mitochondrial DNA. *Forensic Science International: Genetics*, 1(1): 29-34.
- NIST (2024). Forensic DNA Interpretation and Human Factors: Challenges and Solutions. *National Institute of Standards and Technology (NIST) IR 8503*: 1-50.
- Nogueira, T.L.S., Oliveira, T.P., Braz, E.B.V., Santos, O.C.L., Silva, D.A., Amaral, C.R.L., and Carvalho, E.F. (2015). Mitochondrial DNA direct PCR sequencing of blood FTA paper. *Forensic Science International: Genetics Supplement Series*, 5: e611-e613.
- Nwawuba, U.S., Mohammed, A.K., Bukola, A.T., Omusi, I.P., and Ayeubomwan, E.D. (2020). Forensic DNA profiling: autosomal short tandem repeat as a prominent marker in crime investigation. *Malaysian Journal of Medical Sciences*, 27: Article 22.
- Oliveira, M.L., Santos, C., Lima, G.F., Almeida, J.R., Fernandes, A.P., Ribeiro-dos-Santos, Â., Paiva, L.S., Silva, R. and Pinheiro, M.F. (2024). Evaluation of the Precision ID mtDNA Whole Genome Panel for forensic analysis of degraded samples from human skeletal remains. *Anais da Academia Brasileira de Ciências*, 96(2): Article e20231179.
- Parakatselaki, M.-E., and Ladoukakis, E.D. (2021). mtDNA heteroplasmy: origin, detection, significance, and evolutionary consequences. *Life*, 11(7): Article 633.
- Parolin, G.P. (2009). *Citizenship in the Arab World: Kin, Religion and Nation-State*. Amsterdam: Amsterdam University Press.

- Parson, W., and Bandelt, H.J. (2007). Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Science International: Genetics*, 1(1): 13-19.
- Parson, W., and Dür, A. (2007). EMPOP—a forensic mtDNA database. *Forensic Science International: Genetics*, 1(2): 88-92.
- Parson, W., Ballard, D., Budowle, B., Butler, J.M., Gettings, K.B., Gill, P., Gusmão, L., Hares, D.R., Irwin, J.A., King, J.L., de Knijff, P., Morling, N., Prinz, M., Schneider, P.M., Van Neste, C., Willuweit, S., and Phillips, C. (2016). Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics*, 22: 54-63.
- Parson, W., Brandstätter, A., Alonso, A., Brandt, N., Brinkmann, B., Carracedo, A., Corach, D., Froment, O., Furac, I., Grzybowski, T., Hedberg, K., Keyser-Tracqui, C., Kupiec, T., Lutz-Bonengel, S., Mevag, B., Ploski, R., Schmitter, H., Schneider, P., Syndercombe-Court, D., Sørensen, E., Thew, H., Tully, G., and Scheithauer, R. (2004). The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results, and perspectives. *Forensic Science International*, 139(2-3): 215-226.
- Parson, W., Gusmão, L., Hares, D.R., Irwin, J.A., Mayr, W.R., Morling, N., Pokorak, E., Prinz, M., Salas, A., Schneider, P.M., and Parsons, T.J.(2014). DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics*, 13: 134-142.
- Parson, W., Huber, G., Moreno, L., Madel, M.B., Brandhagen, M.D., Nagl, S., Xavier, C., Eduardoff, M., Callaghan, T.C., and Irwin, J.A. (2015). Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. *Forensic Science International: Genetics*, 15: 8-15.
- Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S.M., Souto, L., Fendt, L., Delport, R., Langit, R., Wootton, S., Lagacé, R., and Irwin, J. (2013). Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Science International: Genetics*, 7(5): 543-549.
- Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S.M., Souto, L., Fendt, L., Delport, R., Langit, R., Wootton, S., Lagacé, R., and Irwin, J. (2013). Reprint of: Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Science International: Genetics*, 7(6): 632-639.

- Parsons, T.J., Holland, M.M., and Weedn, V.W. (1997). Mitochondrial DNA typing in forensic casework: validation of a one-step PCR protocol for HV1 and HV2. *Forensic Science International*, 92(1): 1-16.
- Passos, J.F., Saretzki, G., Ahmed, S., Nelson, G., Richter, T., Peters, H., Wappler, I., Birket, M.J., Harold, G., Schaeuble, K., Birch-Machin, M.A., Kirkwood, T.B.L., and von Zglinicki, T. (2007). Mitochondrial dysfunction accounts for the stochastic heterogeneity in telomere-dependent senescence. *PLOS Biology*, 5(5): Article e110.
- Poulton, J., Brown, M.S., Cooper, A., Marchington, D.R., and Phillips, D.I. (1998). A common mitochondrial DNA variant is associated with insulin resistance in adult life. *Diabetologia*, 41(1): 54-58.
- Quintana-Murci, L., Chaix, R., Wells, R.S., Behar, D.M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A.S., Coppa, A., Ayub, Q., Mohyuddin, A., Tyler-Smith, C., Qasim Mehdi, S., Torroni, A., and McElreavey, K. (2004). Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *American Journal of Human Genetics*, 74(5): 827-845.
- Ramos, A., Santos, C., Mateiu, L., Gonzalez, M.M., Alvarez, L., Azevedo, L., Amorim, A., and Aluja, M.P. (2013). Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. *PLOS ONE*, 8(10): Article e74636.
- Rathbun, M.M., McElhoe, J.A., Parson, W., and Holland, M.M. (2017). Considering DNA damage when interpreting mtDNA heteroplasmy in deep sequencing data. *Forensic Science International: Genetics*, 26: 1-11.
- Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H.V., ... and Villems, R. (2003). Origin and diffusion of mtDNA haplogroup X. *American Journal of Human Genetics*, 73(5): 1178-1190.
- Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., ... and Oppenheimer, S. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics*, 67(5): 1251-1276.
- Röck, A., Irwin, J., Dür, A., Parsons, T., and Parson, W. (2011). SAM: string-based sequence search algorithm for mitochondrial DNA database queries. *Forensic Science International: Genetics*, 5(2): 126-132.
- Rogan, E. (2009). *The Arabs: A History*. New York: Basic Books.
- Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., Badii, R., Al-Nabet Al-Marri, A., Abi Khalil, C., Zirie, M., Jayyousi, A., Salit, J., Keinan, A., Clark, A.G., Crystal, R.G., and Mezey,

- J.G. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Research*, 26(2): 151-162.
- Roewer, L. (2013). DNA fingerprinting in forensics: past, present, future. *Investigative Genetics*, 4(1): Article 22.
 - Rowold, D.J., and Herrera, R.J. (2010). Mitochondrial DNA variation in the Jordanian population. *American Journal of Physical Anthropology*, 143(2): 241-248.
 - Russell, P.J. (n.d.). *iGenetics*, Chapter 15, Non-Mendelian Inheritance, edited by Wang, Y.-W., Dept. of Agronomy, NTU.
 - Salas, A., Bandelt, H.J., Macaulay, V., and Richards, M.B. (2007). Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Science International*, 168(1): 1-13.
 - Salas, A., Carracedo, A., Macaulay, V., Richards, M., and Bandelt, H.J. (2005). A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochemical and Biophysical Research Communications*, 335(3): 891-899.
 - Salzano, F.M., and Sans, M. (2014). Interethnic admixture and the evolution of Latin American populations. *Genetics and Molecular Biology*, 37(1 Suppl 1): 151-170.
 - Sanchez, N., Paneto, G.G., Hadi, S., Salas, A., and Lareu, M.V. (2014). Classification of Latin American admixed populations based on mtDNA HVRI polymorphism using the Eaton–Kidd criteria. *Legal Medicine*, 16(3): 114-119.
 - Santos, T.L.S., Cavalcanti, P., de Carvalho, E.F., and da Silva, D.A. (2024). Forensic use of human mitochondrial DNA: A review. *Anais da Academia Brasileira de Ciências*, 96(4): Article e20231179.
 - Santos, C., Sierra, B., Alvarez, L., Ramos, A., Fernández, E., Nogués, R., and Aluja, M.P. (2008). Frequency and pattern of heteroplasmy in the control region of human mitochondrial DNA. *Journal of Molecular Evolution*, 67(2): 191-200.
 - Sato, M., and Sato, K. (2012). Maternal inheritance of mitochondrial DNA: degradation of paternal mitochondria by allogeneic organelle autophagy, allophagy. *Autophagy*, 8(3): 424-425.
 - Saunier, J.L., Irwin, J.A., Strouss, K.M., Ragab, H., Sturk, K.A., and Parsons, T.J. (2009). Mitochondrial control region sequences from an Egyptian population sample. *Forensic Science International: Genetics*, 3(3): e97-e103.
 - Scheible, M., Alenizi, M., Sturk-Andreaggi, K., Coble, M.D., Ismael, S., and Irwin, J.A. (2011). Mitochondrial DNA control region variation in a Kuwaiti population sample. *Forensic Science International: Genetics*, 5(4): e112-e113.

- Schönherr, S., Weissensteiner, H., Kronenberg, F., and Forer, L. (2023). Haplogrep 3 - an interactive haplogroup classification and analysis platform. *Nucleic Acids Research*.
- Scientific Working Group on DNA Analysis Methods (2013). *Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories*.
- Seo, S.B., Zeng, X., King, J.L., Larue, B.L., Assidi, M., Al-Qahtani, M.H., Sajantila, A., and Budowle, B. (2015). Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform Ion Torrent™ PGM™. *BMC Genomics*, 16(Suppl 1): Article S4.
- Shin, M.G., Kajigaya, S., McCoy, J.P., Levin, B.C., and Young, N.S. (2004). Marked mitochondrial DNA sequence heterogeneity in single CD34+ cell clones from normal adult bone marrow. *Blood*, 103(2): 553-561.
- Siddiqi, M.H., Akhtar, T., Rakha, A., Abbas, G., Ali, A., Haider, N., Ali, A., Hayat, S., Masooma, S., Ahmad, J., Tariq, M.A., van Oven, M., and Khan, F.M. (2015). Genetic characterization of the Makrani people of Pakistan from mitochondrial DNA control-region data. *Legal Medicine*, 17(2): 134-139.
- Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23): 6380–6387.
- Smith, D.R. (2016). The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? *Briefings in Functional Genomics*, 15(1): 47-54.
- Smith, S.C. (2004). *Britain's Revival and Fall in the Gulf: Kuwait, Bahrain, Qatar, and the Trucial States, 1950-71* (1st ed.). London: Routledge.
- Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., ... and Richards, M.B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *The American Journal of Human Genetics*, 84(6): 740-759.
- Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilová, E., Macaulay, V., Richards, M.B., Cerny, V., and Pereira, L. (2012). The expansion of mtDNA haplogroup L3 within and out of Africa. *Molecular Biology and Evolution*, 29(3): 915-927.
- Sosa, M.X., Sivakumar, I.K., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K.M., Harkins, T.T., Lin, J., Feldman, A.B., Tata, P., Ehret, G.B., and Chakravarti, A. (2012). Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. *PLOS Computational Biology*, 8(10): Article e1002737.

- Starr, S.F. (2013). *Lost Enlightenment: Central Asia's Golden Age from the Arab Conquest to Tamerlane*. Princeton, NJ: Princeton University Press.
- Stone, A.C., Starrs, J.E., and Stoneking, M. (1994). Mitochondrial DNA analysis of the presumptive remains of Jesse James. *Proceedings of the National Academy of Sciences of the United States of America*, 95(2): 1427-1432.
- Strobl, C., Churchill Cihlar, J., Lagacé, R., Wootton, S., Roth, C., Huber, N., Schnaller, L., Zimmermann, B., Huber, G., Lay Hong, S., Moura-Neto, R., Silva, R., Alshamali, F., Souto, L., Anslinger, K., Egyed, B., Jankova-Ajanovska, R., Casas-Vargas, A., Usaquén, W., Silva, D., ... and Parson, W. (2019). Evaluation of mitogenome sequence concordance, heteroplasmy detection, and haplogrouping in a worldwide lineage study using the Precision ID mtDNA Whole Genome Panel. *Forensic Science International: Genetics*, 42: 244-251.
- Strobl, C., Eduardoff, M., Bus, M.M., Allen, M., and Parson, W. (2018). Evaluation of the Precision ID Whole mtDNA Genome Panel for forensic analyses. *Forensic Science International: Genetics*, 35: 21-25.
- Sullivan, K.M., Hopgood, R., and Gill, P. (1992). Identification of human remains by amplification and automated sequencing of mitochondrial DNA. *International Journal of Legal Medicine*, 105(2): 83-86.
- Suomalainen, A., Ciafaloni, E., Koga, Y., Peltonen, L., DiMauro, S., and Schon, E.A. (1992). Use of single strand conformation polymorphism analysis to detect point mutations in human mitochondrial DNA. *Journal of the Neurological Sciences*, 111(2): 222-226.
- Swallow, D. (2004). *Human Evolutionary Genetics: Origins, Peoples & Disease*. *Journal of Medical Genetics*, 41: 958-959.
- Syndercombe Court, D. (2021). Mitochondrial DNA in forensic use. *Emerging Topics in Life Sciences*, 5(3): 415-426.
- Szabo, S., Jaeger, K., Fischer, H., Tschachler, E., Parson, W., and Eckhart, L. (2012). In situ labeling of DNA reveals interindividual variation in nuclear DNA breakdown in hair and may be useful to predict success of forensic genotyping of hair. *International Journal of Legal Medicine*, 126(1): 63-70.
- Szibor, R., Michael, M., Zhang, X., Eriksson, S., and Lüdecke, H.J. (1997). Sequence heterogeneity of mitochondrial DNA polymorphisms in genomic DNA preparations from individual cells. *BioTechniques*, 23(2): 342-350.
- Tadmouri, G.O., Nair, P., Obeid, T., Al Ali, M.T., Al Khaja, N., and Hamamy, H.A. (2009). Consanguinity and reproductive health among Arabs. *Reproductive Health*, 6: Article 17.

- Tagliabracci, A., and Turchi, C. (2020). mtDNA exploitation in forensics. *The Human Mitochondrial Genome*, Academic Press: 145-169.
- Taylor, C.R., Kiesler, K.M., Sturk-Andreaggi, K., Ring, J.D., Parson, W., Schanfield, M., Vallone, P.M., and Marshall, C. (2020). Platinum-quality mitogenome haplotypes from United States populations. *Genes*, 11(11): Article 1290.
- Taylor, D., Bright, J.A., and McGovern, C.E. (2022). Evaluating the suitability of current mitochondrial DNA interpretation guidelines in the massively parallel sequencing era: A case study involving the Norfolk Island population. *Forensic Science International: Genetics*, 58: Article 102674.
- Teebi, A., and Teebi, S. (2005). Genetic diversity among the Arabs. *Public Health Genomics*, 1(8): 21-26.
- Templeton, J.E.L., Brotherton, P.M., Llamas, B., et al. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative Genetics*, 4: Article 26.
- Thomas, A.W., Morgan, R., Sweeney, M., et al. (1994). The detection of mitochondrial DNA mutations using single stranded conformation polymorphism (SSCP) analysis and heteroduplex analysis. *Human Genetics*, 94(6): 621-623.
- Tikochinski, Y., Carreras, C., Tikochinski, G., and Vilaça, S.T. (2020). Population-specific signatures of intra-individual mitochondrial DNA heteroplasmy and their potential evolutionary advantages. *Scientific Reports*, 10(1): Article 211.
- Tillmar, A., Sjöholm, M.I., Hagelberg, E., and Kalling, M. (2019). Massively parallel sequencing for forensic genetic casework: challenges, benefits, and potential solutions. *Forensic Science International: Genetics*, 43: Article 102153.
- Timmis, J.N., et al. (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5: 123-135.
- Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*, 22(6): 339-345.
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.L., and Wallace, D.C. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics*, 144(4): 1835-1850.
- Tully, G., Bär, W., Brinkmann, B., Carracedo, A., Gill, P., Morling, N., Parson, W., and Schneider, P. (2001). Considerations by the European DNA Profiling (EDNAP) group on the working practices, nomenclature, and interpretation of mitochondrial DNA profiles. *Forensic Science International*, 124(1): 83-91.

- Tully, G., Barritt, S.M., Bender, K., Brignon, E., Capelli, C., Dima-Simonin, N., Eichmann, C., Ernst, C.M., Lambert, C., Lareu, M.V., Ludes, B., Mevag, B., Parson, W., Pfeiffer, H., Salas, A., Schneider, P.M., and Staalstrom, E. (2004). Results of a collaborative study of the EDNAP group regarding mitochondrial DNA heteroplasmy and segregation in hair shafts. *Forensic Science International*, 140(1): 1-11.
- Turvey, B.E. (2022). *Criminal Profiling: An Introduction to Behavioral Evidence Analysis* 5th edn. San Diego: Academic Press.
- Valladas, H., Reyss, J.-L., Joron, J.-L., Valladas, G., Bar-Yosef, O., and Vandermeersch, B. (1988). Thermoluminescence dates of Mousterian “Proto-Cro-Magnon” remains from Israel and the origin of modern man. *Nature*, 331: 614-616.
- van Dijk, E.L., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research*, 322(1): 12-20.
- van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2): e386-e394.
- van Oven, M. (2015). PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series*, 5: e392-e394.
- Wallace, D.C. (1994). Mitochondrial DNA sequence variation in human evolution and disease. *Proceedings of the National Academy of Sciences of the United States of America*, 91(19): 8739-8746.
- Wallace, D.C., and Chalkia, D. (2013). Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor Perspectives in Biology*, 5(11): Article 021220.
- Wallace, D.C. (2010). Mitochondrial DNA mutations in disease and aging. *Environmental and Molecular Mutagenesis*, 51(5): 440-450.
- Watson, E., Forster, P., Richards, M., and Bandelt, H.J. (1997). Mitochondrial footprints of human expansions in Africa. *American Journal of Human Genetics*, 61(3): 691-704.
- Wei, W., and Chinnery, P.F. (2020). Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. *Journal of Internal Medicine*, 287(6): 634-648.
- Weissensteiner, H., Forer, L., Fendt, L., Kheirkhah, A., Salas, A., Kronenberg, F., and Schoenherr, S. (2021). Contamination detection in sequencing studies using the mitochondrial phylogeny. *Genome Research*, 31(2): 309-316.

- Wilson, M.R., Allard, M.W., Monson, K., Miller, K.W., and Budowle, B. (2002). Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Science International*, 129(1): 35-42.
- Wilson, M.R., Polanskey, D., Butler, J., DiZinno, J.A., Replogle, J., and Budowle, B. (2002). Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. *BioTechniques*, 33(2): 290-291.
- Wilson, M.R., Polanskey, D., Butler, J.M., DiZinno, J.A., Replogle, J., and Budowle, B. (2016). The future of forensic DNA analysis. *Forensic Science Review*, 28(2): 89-102.
- Woerner, A.E., Cihlar, J.C., Smart, U., and Budowle, B. (2020). Numt identification and removal with RtN!. *Bioinformatics (Oxford, England)*, 36(20): 5115-5116.
- Xavier, C., Eduardoff, M., Strobl, C., and Parson, W. (2019). SD quant—sensitive detection tetraplex-system for nuclear and mitochondrial DNA quantification and degradation inference. *Forensic Science International: Genetics*, 42: 39-44.
- Yang, Y., Xie, B., and Yan, J. (2014). Application of next-generation sequencing technology in forensic science. *Genomics, Proteomics & Bioinformatics*, 12(5): 190-197.
- Yao, Y.G., Bravi, C.M., and Bandelt, H.J. (2004). A call for mtDNA data quality control in forensic science. *Forensic Science International*, 141(1): 1-6.
- Yasmin, M., Rakha, A., Noreen, S., and Salahuddin, Z. (2017). Mitochondrial control region diversity in Sindhi ethnic group of Pakistan. *Legal Medicine*, 26: 11-13.
- Yonsei University, n.d. mtDNA-CR Protocol. Available at: <http://forensic.yonsei.ac.kr/protocol/mtDNA-CR.pdf>.
- Zavala, E. I., Thomas, J. T., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Meyers, K. K., Barrit-Ross, S., Aximu-Petri, A., Richter, J., Nickel, B., Berg, G. E., McMahon, T. P., Meyer, M., & Marshall, C. (2022). Ancient DNA methods improve forensic DNA profiling of Korean War and World War II unknowns. *Genes*, 13(1): Article 129.

Appendices

Appendix I: Ethical approval letter from the University of Central Lancashire's STEM Ethics Committee.



8 November 2018

Will Goodwin / Reem Almheiri
School of Forensic and Applied Sciences
University of Central Lancashire

Dear Will / Reem

Re: STEMH Ethics Committee Application
Unique reference Number: STEMH 937

The STEMH ethics committee has granted approval of your proposal application 'Development and Analysis of Mitochondria DNA Database in UAE Populations for Forensic Applications'. Approval is granted up to the end of project date*. It is your responsibility to ensure that

- ☐ the project is carried out in line with the information provided in the forms you have submitted
- ☐ you regularly re-consider the ethical issues that may be raised in generating and analysing your data
- ☐ any proposed amendments/changes to the project are raised with, and approved, by Committee
- ☐ you notify EthicsInfo@uclan.ac.uk if the end date changes or the project does not start
- ☐ serious adverse events that occur from the project are reported to Committee
- ☐ a closure report is submitted to complete the ethics governance procedures (Existing paperwork can be used for this purposes e.g. funder's end of grant report; abstract for student award or NRES final report. If none of these are available use [e-Ethics Closure Report Proforma](#)).

Please also note that it is the responsibility of the applicant to ensure that the ethics committee that has already approved this application is either run under the auspices of the National Research Ethics Service or is a fully constituted ethics committee, including at least one member independent of the organisation or professional group.

Yours sincerely

Karen Rouse
Chair
STEMH Ethics Committee

* for research degree students this will be the final lapse date

NB - Ethical approval is contingent on any health and safety checklists having been completed, and necessary approvals as a result of gained.

Appendix II: Ethical approval letter from the Dubai Health Authority's Ethics Committee.



**DUBAI SCIENTIFIC RESEARCH ETHICS
COMMITTEE
APPROVAL LETTER**



From :	Dubai Scientific Research Ethics Committee (DSREC) Dubai Health Authority	Date :	20 MARCH 2018
To :	Ms. Reem Matar Khalifa Almazaina Almheiri, Student of doctor of philosophy, University of Central Lancashire	Ref :	DSREC-SR-03/2018_03
Study Site	Molecular Genetic Department or Dubai Blood Donors Centre at Dubai Health Authority, DHA		

Subject: Approval for the research proposal, **"Development and Analysis of Mitochondria DNA Database in UAE Populations for Forensic Applications"**

Dear Ms. Reem Matar Khalifa Almazaina Almheiri,

Thank you for submitting the above mentioned research proposal to Dubai Scientific Research Ethics Committee, DHA. The Dubai Scientific Research Ethics Committee has been organized and operates in accordance with the ICH/GCP guidelines and the committee is registered with the Office for Human Research Protection (OHRP).

Your request was discussed with Dubai Scientific Research Ethics Committee. We are pleased to advice you that the committee has granted ethical approval for the above mentioned study involving the collection of the blood samples from above mentioned centers in Dubai Health Authority, analyzing the same in the General Department of Forensic and Criminology, Dubai Police and finally submitting the results to University of Central Lancashire. However, you will have make sure all the administrative approval are obtained from the sites involves, prior to the study initiation.

Please note that it is DSREC's policy that the principal investigator should report to the committee of the following:

1. Anything which might warrant review of ethical approval of the project in the specified format, including:
 - any serious or unexpected adverse events and
 - unforeseen events that might affect continued ethical acceptability of the project
2. Any proposed changes to the research protocol or to the conduct of research
3. Any new information that may affect adversely the safety of the subjects
4. If the project is discontinued before the expected date of completion (reason to be specified)
5. Annual report to DSREC about the progress of the study
6. A final report of the finding on completion of the study

The approval for the study expires on **20 MARCH 2019**. Should you wish to continue the study after this date, please submit an application for renewal together with the Annual Study site progress report no later than 30 days prior to the expiry date.



The DSREC wishes you every success in your research.

Yours faithfully,

Dr. Suhail Abdulla Mohd Arif
Chairman
Dubai Scientific Research Ethics Committee
Dubai Health Authority

Dubai Scientific Research Ethics Committee
Dubai Health Authority
Dubai, UAE.

Appendix III: Poster displayed in ISFG 2019



Whole Mitochondrial Genome Analysis In The United Arab Emirates Populations

R. Almheiri^{1,2*}, M. Tracy¹, R. Alghafri^{1,3}, W. Goodwin²

¹General Department of Forensic Sciences and Criminology, Dubai Police G.H.Q, Dubai, UAE

²Forensic and Applied Sciences, University of Central Lancashire, United Kingdom

³Biology Department, United Arab Emirates University, Al-Ain, UAE

Introduction

Mitochondrial genome analysis has earned its place in the field of genetics to provide a new window into understanding population ancestry. Its ability to produce results from minimal or decayed samples where nucleus DNA profiling proves are difficult, it serves an additional benefit to forensic genetics. As a result, it is essential to begin building a precise regional study of populations. In this study, 350 whole blood samples were collected on FTA paper, from the United Arab Emirates populations, including inhabitants from rural regions. The three main ethnic groups studied were 100 Indians, 100 Pakistanis, and 150 Emiratis. Both extraction of whole blood samples using PrepFiler[®] as well as direct amplification of mtDNA control region from purified 1.2mm disc of FTA card stained with blood were attempted in this study [1]. Quantity of the amplified control region was estimated using gel electrophoresis. Three sets of primers were used afterward to sequence purified products of control region of mtDNA using Sanger sequencing method for subset of samples. 150 Whole mtDNA sequence was obtained for UAE Arabs population, generated using MPS technology – Ion[™] 5S. Concordance with control region sequences generated using Sanger Sequencing approach was investigated. Resulted haplotypes were compared against worldwide mtDNA database (EMPOP) and estimated haplotypes frequency is shown in this study. As a forensic lab, non-probative challenging bone samples were tested to highlight the potentials value of using MPS technology – Ion[™] 5S. This study shows the first mtGenome data generated for UAE Arabs population which helps extending the current available mtDNA control region database. As a result this study shows great value of implementing MPS in the routine work in forensic genetics at Dubai Police.

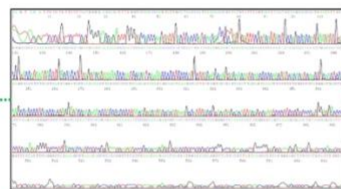


Figure 1a + 1b: Electropherogram of one of the samples analyzed using sanger sequencing method. Three primers were used to cover the control region.

One forward (left) and two reverse (right) (only one reverse is showing here). 33 samples showed concordance with MPS approach using Ion 5S[™] technology.

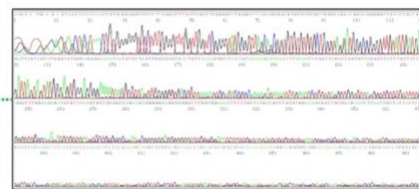


Figure 3: Quasi-median network analysis generated for 150 whole mtDNA sequences for Emirati samples using Network software along with the facilitating tool in EMPOP website online.

The analysis showed t^* value of 1 which is reflected in the star-shaped network, which is in agreement with the generally expected evolutionary pattern in mtGenome.

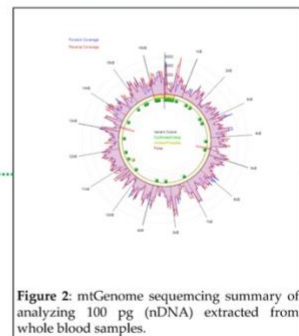
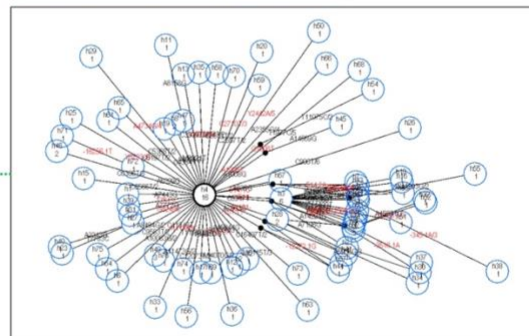


Figure 4: An example of online alignment and haplogrouping assignment using EMPOP platform showing the haplogroup I (I0a2c haplotype).

Whole mtGenome in Emirati population

No. of Samples: 150
No. of Polymorphisms: 132
No. of Haplotypes: 149
Ten Macro-haplogroups were observed in this study including H, I, J, K, L, M, N, R, T and U.

References

- [1] T.L.S. Nogueira, T.P. Oliveira, E.B.V. Braz, O.C.L. Santos, D.A. Silva, C.R.L. Amaral, E.F. Carvalho, 2015. Mitochondrial DNA direct PCR sequencing of blood FTA paper. Forensic Science International: Genetics Supplement Series, 5, e611-e613.
- [2] Nomenclature: ISFG recommendations. Parson et al. 2014
- [3] Quality Control (QC): Parson and Ditt 2007, Zimmermann et al. 2011
- [4] Database Search: SAM2 Huber et al. 2018

EMPOP

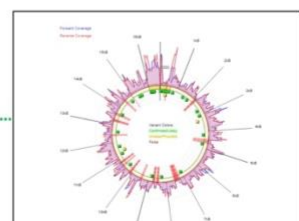


Figure 5: mtGenome summary of analyzing 45 pg (nDNA) extracted from severely burnt bone sample. An Example of Forensic Application.



Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss



Whole mtGenome analysis in United Arab Emirates populations

R. Almheiri^{a,b,*}, M. Bakri^a, R. Alghafri^{a,c,d}, W. Goodwin^d

^a General Department of Forensic Sciences and Criminology, Dubai Police G.H.Q, Dubai, United Arab Emirates

^b Forensic and Applied Sciences, University of Central Lancashire, United Kingdom

^c Biology Department, United Arab Emirates University, Al-Ain, United Arab Emirates

^d Faculty of Health Science and Medicine, Bond University, Queensland, Australia



ARTICLE INFO

Keywords:
mtGenome
UAE
Haplogroups
Control region
MPS

ABSTRACT

Mitochondrial DNA analysis has earned its place in the field of genetics for providing a new window into understanding population ancestry. Its ability to produce results from minimal or decayed samples where nucleus DNA profiling proves difficult is an additional benefit to forensic genetics. A total of 150 whole blood samples were collected on FTA paper, from Arabs population in United Arab Emirates, including inhabitants from rural regions. Both extraction of whole blood samples using PrepFiler® as well as direct amplification of mtDNA control region from purified 1.2 mm disc of FTA card stained with blood were attempted in this study. Quantity of the amplified control region was estimated using gel electrophoresis. Three sets of primers were used afterward to sequence purified products of amplified control region of mtDNA using Sanger sequencing method. 150 mtGenome sequences were obtained for UAE Arabs population, generated using MPS technology – Ion™ 5S. Concordance with control region sequences obtained using Sanger Sequencing approach was investigated. Resulted haplotypes were compared against worldwide mtDNA database (EMPOP) and estimated haplotypes frequency is shown in this study. As a forensic lab, non-probative challenging bone samples were tested to highlight the potentials value of using MPS technology – Ion™ 5S. This study shows the first mtGenome data generated for UAE Arabs population which helps extending the current available mtDNA control region database. As a result, this study shows great value of implementing MPS in the routine work in forensic genetics at Dubai Police.

1. Introduction

Mitochondrial DNA analysis application in forensic case work emerged in the late 1980's and early 1992, where it has shown a great advantages over the restriction fragment length polymorphisms (RFLP) especially in analyzing hair shafts, bones and teeth samples which were considered challenging for STR analysis. Despite of the advantages of mtDNA, there are very few studies been done in Arabs populations and more specifically in United Arab Emirates (UAE) Arabs population to evaluate mtDNA for forensic applications, such reluctance in studying mtDNA most likely because of the intensive work involved in the traditional workflow to sequences mtDNA which makes it hardly applicable in forensic laboratories. With the emerging of the massive parallel sequencing (MPS) there have been a remarkable increase in mtDNA studies and application due to the simplified workflow to sequence the mtDNA [1] It allows not only sequence of the control region, which shows high polymorphisms across individuals, but it can easily sequence the whole mtDNA. Therefore, the great advantages of mtDNA

analysis will find its way soon to the forensic laboratories especially in United Arab Emirates. This study, evaluates the technology of MPS and investigate the value of mtDNA in the United Arab Emirates population.

2. Materials and methods

2.1. Samples collection and extraction

A total of 150 consented whole blood samples were collected and extracted in addition to bone samples using PrepFiler® as per the manufacture protocol for the purpose of this study.

2.2. Sanger sequencing analysis

For the purpose of concordance study, 100 samples were sequenced using Sanger Sequencing as per the published protocol [2]. Initially L15879 and H727 Primers were used to amplify the control region of the mtDNA. The PCR conditions were: initial denaturation at 94 °C for

* Corresponding author at: General Department of Forensic Sciences and Criminology, Dubai Police G.H.Q, Dubai, United Arab Emirates.
E-mail address: r.almezaina@gmail.com (R. Almheiri).

<https://doi.org/10.1016/j.fsigs.2019.10.031>

Received 17 September 2019; Accepted 6 October 2019

Available online 15 October 2019

1875-1768/ © 2019 Published by Elsevier B.V.

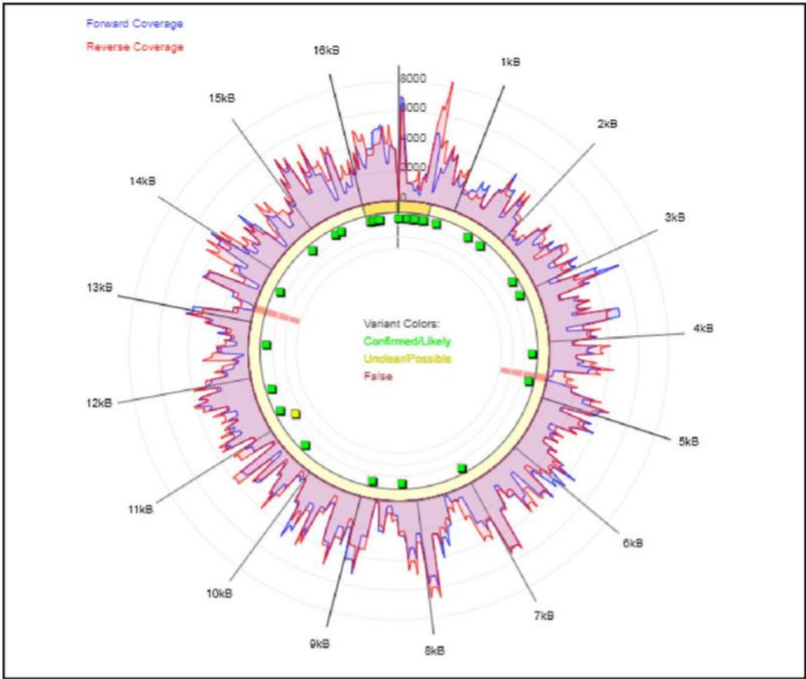


Fig. 1. mtGenome sequencing summary of analyzing 100 pg (nDNA) extracted from whole blood samples developed by Converge™ Software.

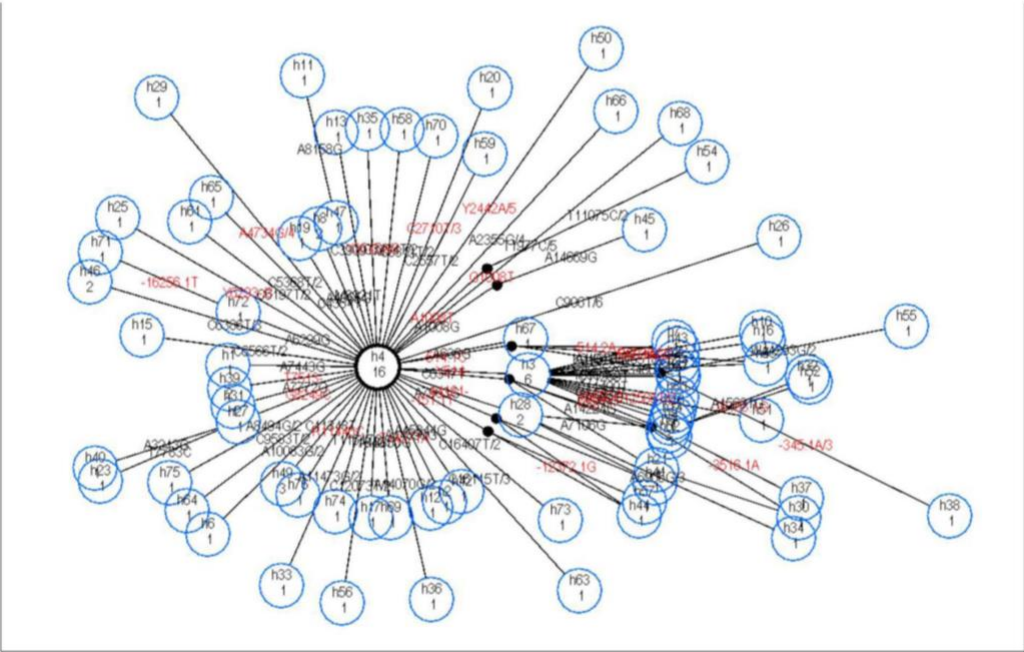


Fig. 2. network analysis generated for 150 whole mtDNA sequences for Emirati samples using Network software along with the facilitating tool in EMPOP website online. The analysis showed t^* value of 1 which is reflected in the star-shaped network, which is in agreement with the generally expected evolutionary pattern in mtGenome.

30 s, 60 °C for 90 s, 72 °C for 60 s, and final extension at 72 °C for 10 min. Then amplified products were purified using MicroClean as per the manufacturer's protocol. The sequencing was performed using the BigDye™ Terminator Kit v3.1 Cycle Sequencing Kit (ThermoFisher) using F15975, R635 and R240 primers. Unincorporated dye terminators were removed by the addition of 1 µl 100 mM EDTA, 1 µl cold 3 M NaOAc, 1 µl Glycogen, and 30 µl cold ethanol and kept at room temperature overnight for precipitation. Followed by the next day spin down at 13 K at 4 °C for 30 min. Remove the supernatant and add an adequate amount of cold ethanol, and spin down at 13 K at 4 °C for 30 min. Remove the supernatant and dry in PCR machine at 50 °C for 10 min. The pellet was finally resuspended in 11 µl HiDi Formamide, mixed and centrifuged briefly and transferred for 96 MicroAmp® Plate. Finally, sequencing was carried out using 3500 Genetic Analyser 50 cm capillary array with POP-7.

2.3. Massive parallel sequencing

Library preparation was performed using the Precision ID mtDNA Whole Genome Panel and the Ion AmpliSeq™ Library kit 2.0, according to the user's guide. The optimum amount of mtDNA was estimated by using as reference the input of approximately 0.1 ng of nuclear DNA. FuPa® reagent was added into each PCR product to partially digest the primers and the libraries were barcoded using the Ion Xpress™ Barcode Adapters kit, following the manufacturer's protocol. Each unamplified library was purified using Agencourt® AMPure® XP reagent and then quantified with the Ion Library TaqMan™ Quantitation kit on the 7500 Real-Time PCR System, following the manufacturer's protocol. According to the quantification results, all barcoded libraries were diluted in equimolar volumes of 30 and 7.5 pM to ensure equal contribution in the sequencing run. Ion Chef™ incorporated pooled libraries into Ion 520™ chips, according to the manufacturer's protocol. Finally, chips were sequenced on the Ion S5™ System using the specified reagents, as described in the user's guide.

3. Results and discussion

A total of 33 sequences were generated by 3500 Genetic Analyser and analysed using SeqScape®, whereas the alignment of sequences

were done using BioEdit software [3]. A total of 150 Whole mtGenome Emirati population were successfully obtained. Raw data were analysed with Converge™ Software, using the revised Cambridge Reference Sequence (NC_012920.1). Shown in Fig. 1, the coverage of one UAE population individual representing a minimum of 1000 coverage. All 33 samples showed concordance with the sequences obtained by Ion S5™.

All samples were searched in EMPPOP [4] for haplotypes and haplogroups. As a result, 150 Samples expressed 132 number of polymorphisms, with 149 haplotypes, and Ten Macro-haplogroups were observed in this study including H, I, J, K, L, M, N, R, T and U haplogroups. Due to the history of the migration movement in the Arabian Peninsula, such haplogroups are expected to be present in the Emirati population. Fig. 2 shows Network analysis of all samples generated using Network software facilitated by EMPPOP database website. The star shaped network represent the quality of the samples representing one population ($t'' = 1$).

4. Conclusion

As a conclusion, by implementing MPS for mtDNA analysis, it shows remarkable results as a robust promising tool for the forensics lab routine. Not only it expresses an easier workflow but also, it gives more information beyond the control region when it comes to mtDNA analysis therefore, higher discrimination capacity expected for forensic purposes.

References

- [1] C. Strobl, M. Eduardoff, M. Bus, M. Allen, W. Parson, Evaluation of the precision ID whole mtDNA genome panel for forensic analyses, *Forensic Sci. Int. Genet.* 35 (2018) 21–25.
- [2] Jana Naue, Steffen Hörer, Timo Sängler, Christina Strobl, Petra Hatzler-Grubwieser, Walther Parson, Sabine Lutz-Bonengel, Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA, *Mitochondrion* 20 (2015) 82–94, <https://doi.org/10.1016/j.mito.2014.12.002> ISSN 1567-7249.
- [3] BioEdit 7.2 software program. Hall TA, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, *Nucl. Acids. Symp. Ser.* 41 (1999) 95–98.
- [4] W. Parson, et al., EMPPOP-a forensic mtDNA database, *Forensic Sci. Int. Genet.* 1 (2) (2007) 88–92.

Appendix V: Control region sequences for 30 Pakistanis samples with the haplotypes nomenclatures, including the SNPs types counted for transition and transversion.

Control region sequences for 30 Pakistanis samples

No.	Sample	Haplotypes
1	PAK_01	16051G 16206C 73G 194T 263G 315.1C Transition: 4 Transversion: -
2	PAK_02	16145A 16223T 16261T 16311C 16519C 73G 249d 263G 315.1C 489C Transition: 8 Transversion: -
3	PAK_03	16129A 16218T 16223T 16519C 73G 150T 263G 294C 309d 315.1C 333C 489C Transition: 10 Transversion: -
4	PAK_04	16166d 16223T 16318T 16519C 73G 93G 194T 246C 263G 294C 309.1C 315.1C 489C Transition: 9 Transversion: 1
5	PAK_05	16111T 16223T 16320T 16399G 16519C 73G 195A 263G 315.1C 489C 523d 524d Transition: 8 Transversion: 1
6	PAK_06	16223T 16356C 16362C 73G 146C 263G 315.1C 461T 489C 523del 524del Transition: 10 Transversion: -
7	PAK_07	16172C 16278T 16519C 73G 263G 315.1C 524.ACACAC Transition: 5 Transversion: -
8	PAK_08	16223T 16519C 73G 146C 154C 263G 315.1C 489C Transition: 8 Transversion: -
9	PAK_09	16071T 16092C 16519C 73G 152C 263G 315.1C Transition: 6 Transversion: -
10	PAK_10	16126C 16172C 16223T 16344T 16519C 73G 146C 263G 315.1C 482C 489C 524.AC Transition: 10 Transversion: -
11	PAK_11	16051G 16234T 16240C 16242G 16311C 16519C 73G 146C 152C 263G 315.1C 523d 524d Transition: 9 Transversion: 2
12	PAK_12	16309G 16318T 16519C 73G 151T 152C 263G 315.1C 523d 524d Transition: 6 Transversion: 1
13	PAK_13	16222T 16242T 16273A 16356C 150T 263G 315.1C Transition: 6 Transversion: -
14	PAK_14	16051G 16189C 16234T 73G 146C 152C 263G 315.1C Transition: 7 Transversion: -

15	PAK_15	16071T 16519C 73G 152C 263G 315.1C Transition: 5 Transversion: 0
16	PAK_16	16093C 16223T 16234T 16274A 16519C 73G 195A 263G 315.1C 489C 523d 524d Transition: 8 Transversion: 1
17	PAK_17	16519C 263G 315.1C 455.1T Transition: 14 Transversion: -
18	PAK_18	16179d 16223T 16519C 73G 195A 263G 309d 315.1C 489C 523d 524d Transition: 5 Transversion: 1
19	PAK_19	16145A 16192T 16223T 16300G 16316G 16519C 73G 146C 263G 309.1C 315.1C 489C Transition: 10 Transversion: -
20	PAK_20	16126C 16154C 16223T 16519C 73G 93G 263G 315.1C 482C 489C 523d 524d Transition: 9 Transversion: -
21	PAK_21	16051G 16093C 16129A 16223T 16519C 73G 195C 263G 315.1C 489C Transition: 9 Transversion: -
22	PAK_22	16093C 16182C 16183C 16189C 16249C 16311C 73G 263G 285T 315.1C 523d 524d Transition: 7 Transversion: 2
23	PAK_23	16270T 73G 150T 152C 257G 263G 264T 315.1C Transition: 7 Transversion -
24	PAK_24	16093C 16129A 16223T 16292T 16519C 73G 189G 194T 195C 204C 207A 263G 315.1C Transition: 12 Transversion: -
25	PAK_25	16051G 16129A 16209C 16239T 16311C 16352C 16353T 73G 146C 152C 234G 263G 315.1C Transition: 12 Transversion: -
26	PAK_26	16169T 16172C 16223T 16278T 16519C 73G 150T 263G 315.1C 462T 489C Transition: 10 Transversion: -
27	PAK_27	16111T 16129A 16304C 16519C 73G 234G 249d 263G 315.1C 523d 524d Transition: 7 Transversion: -
28	PAK_28	16126C 16185T 16223T 16519C 73G 195C 207A 263G 315.1C 482C 489C Transition: 10 Transversion: 0
29	PAK_29	16111T 16189C 16223T 16224C 16294G 16295T 16519C 73G 152C 263G 315.1C 489C Transition: 10 Transversion: 1
30	PAK_30	16129A 16223T 16265C 16344T 16519C 73G 263G 315.1C 374G 489C Transition: 8 Transversion: 1

Appendix VI: Control region sequences for 30 Indians samples with the haplotypes nomenclatures, including the SNPs types counted for transition and transversion.

Control region sequences for 30 Indians samples.

No.	Sample	Haplotypes
1	IND_01	16223T 16519C 73G 236C 263G 315.1C 489C Transition: 6 Transversion: -
2	IND_02	16172C 16390A 16519C 73G 150T 154C 195C 263G 315.1C 455.1T 456T Transition: 9 Transversion: -
3	IND_03	16223T 16311C 16519C 73G 143A 199C 204C 250C 263G 297G 315.1C 573.1C 573.2C Transition: 11 Transversion: -
4	IND_04	16217C 72C 73G 152C 195C 263G 315.1C 523del 524del Transition: 6 Transversion: -
5	IND_05	16497G 16519C 16524G 73G 152C 263G 315.1C 373G Transition: 7 Transversion: -
6	IND_06	16309G 16318T 16519C 73G 152C 263G 315.1C 523del 524del Transition: 5 Transversion: 1
7	IND_07	16051G 16093G 16154C 16206C 16230G 16311C 73G 263G 315.1C Transition: 7 Transversion: 1
8	IND_08	16051G 16129A 16179T 16234T 16247G 16519C 73G 152C 239C 247A 263G 315.1C Transition: 11 Transversion: -
9	IND_09	16051G 16193T 16234T 16278T 16357C 73G 200G 263G 315.1C 499A Transition: 9 Transversion: -
10	IND_10	16051G 16126C 16223T 16278T 16382.1C 16519C 73G 263G 315.1C 482C 489C Transition: 9 Transversion: -
11	IND_11	16129A 16223T 16519C 73G 263G 315.1C 489C Transition: 6 Transversion: -
12	IND_12	16069T 16274A 16318T 16519C 73G 151T 152C 263G 315.1C 523del 524del Transition: 7 Transversion: 1
13	IND_13	16172C 16173T 16223T 16235G 16290T 16311C 16319A 16362C 16519C 73G 152C 234G 235G 263G 315.1C 523del 524del Transition: 14 Transversion: -
14	IND_14	16223T 16274A 16319A 16320T 16362C 16518T 16519C 73G 143A 195C 263G 315.1C 337G 447G 489C Transition: 12

Transversion: 2
15 IND_15 16129A 16223T 16519C 73G 263G 315.CC 489C
Transition: 7
Transversion: -
16 IND_16 16051G 16206C 73G 263G 315.1C
Transition: 3
Transversion: 1
17 IND_17 16051G 16126C 16519C 73G 263G 315.1C
Transition: 5
Transversion: -
18 IND_18 16193T 16223T 16519C 73G 152C 239C 263G 315.1C 489C 523del 524del
Transition: 8
Transversion: -
19 IND_19 16051G 16093G 16154C 16206C 16230G 16311C 73G 151T 263G 315.1C
Transition: 7
Transversion: 2
20 IND_20 16184T 16223T 16256G 16311C 16362C 73G 146C 152C 263G 315.1C 461T 489C 523del
524del
Transition: 10
Transversion: 1
21 IND_21 16223T 56del 58A 65.1T 66T 73G 153G 263G 315.1C 463T 485C 489C
Transition: 7
Transversion: 2
22 IND_22 16129A 16158G 16213A 16362C 16519C 73G 263G 315.1C
Transition: 7
Transversion: -
23 IND_23 16129A 16223T 16292T 16519C 73G 189G 194T 195C 204C 207A 263G 315.1C
Transition: 11
Transversion: -
24 IND_24 16172C 16223T 16224C 16270T 16274A 16319A 16352C 16519C 16524C 73G 204C 263G
315.1C 447G 489C
Transition: 12
Transversion: 2
25 IND_25 16189C 16223T 16301T 73G 146C 263G 315.1C 489C
Transition: 7
Transversion: -
26 IND_26 16223T 16289G 16519C 73G 263G 315.1C 489C 511T
Transition: 7
Transversion: -
27 IND_27 16223T 16270T 16274A 16319A 16352C 16519C 73G 183G 204C 263G 315.1C 447G 489C
Transition: 11
Transversion: 1
28 IND_28 16223T 16311C 16519C 73G 143A 199C 204C 250C 263G 315.1C
Transition: 9
Transversion: -
29 IND_29 16223T 16270T 16288C 16319A 16352C 16519C 73G 195C 204C 263G 315.1C 447G 489C
Transition: 11
Transversion: 1

30	IND_30	16166del 16223T 16519C 73G 146C 195A 263G 315.1C 489C 523del 524del
		Transition: 6
		Transversion: 1

Appendix VII: Whole mitochondrial DNA sequences using MPS for 50 Pakistanis samples with the haplotypes nomenclatures

Whole mitochondrial DNA sequences using MPS for 50 Pakistanis samples

No.	Sample	Haplogroup	Haplotypes
1	PAK_01	U2a2	73G 194T 263G 315.1C 750G 1438G 1811G 2706G 3316A 4769G 4970G 5201C 6116G 7028T 7257G 7859A 8860G 10355A 11299C 11467G 11719A 11932T 12308G 12372A 12477C 12561A 14182C 14766T 14883T 15326G 16051G 16206C 16271C
2	PAK_02	M4	73G 249del 263G 315.1C 489C 750G 1438G 2706G 4769G 6620C 7028T 7673G 7859A 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16145A 16223T 16261T 16311C 16519C
3	PAK_03	M5c1	73G 150T 263G 294C 315.1C 333C 489C 575T 750G 1438G 1888A 2706G 3144G 4769G 4851T 5319G 5417A 6413C 7028T 8701G 8860G 9540C 9632G 10398G 10400T 10873C 11719A 12705T 13708A 14766T 14783C 15043A 15301A 15326G 16129A 16218T 16223T 16519C
4	PAK_04	M18a	73G 93G 194T 246C 315.1C 489C 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12498T 12705T 13135A 14766T 14783C 15043A 15301A 15326G 16223T 16318T 16519C
5	PAK_05	M30	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3338C 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12172G 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16111T 16223T 16320T 16399G 16519C
6	PAK_06	M32'56	73G 146C 263G 315.1C 461T 489C 523del 524del 750G 2706G 3486T 3537G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12507G 12705T 14766T 14783C 15043A 15301A 15326G 16223T 16356C 16362C
7	PAK_07	R30a1c	73G 263G 315.1C 524.ACACAC 750G 1438G 2056A 2706G 3316A 4232C 4769G 5442C 6764A 7028T 8584A 8860G 9142A 9156G 9242G 9869G 11047G 11719A 12714C 13161C 13773G 14766T 150550C 15326G 16172C 16278T 16519C
8	PAK_08	M5a	73G 146C 154C 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10589A 10873C 11719A 12477C 12681C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16223T 16519C
9	PAK_09	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16092C 16519C

10	PAK_10	M3d	73G 146C 263G 315.1C 482C 489C 524.AC 750G 1438G 2706G 4769G 6305A 7028T 8701G 8860G 9266A 9540C 10398G 10400T 10873C 10954T 11150A 11719A 11827C 12705T 12873C 14758T 14766T 14783C 15043A 15301A 15326G 16126C 16172C 16223T 16344T 16519C
11	PAK_11	U2c1a	73G 146C 152C 263G 315.1C 523del 524del 750G 1438G 1811G 2706G 4769G 5790A 6320C 7028T 8023C 8676T 8860G 9101C 9767T 11467G 11719A 12308G 12372A 14766T 14935C 15043A 15061G 15236G 15326G 16051G 16234T 16240C 16242G 16311C 16519C
12	PAK_12	U7a	73G 151T 152C 263G 315.1C 523del 524del 750G 980C 1438G 1811G 2706G 3705A 3741T 4733C 4769G 4947C 5360T 7028T 8137T 8684T 8860G 10142T 11467G 11719A 12308G 12372A 13500C 14569A 14766T 15326G 15448A 16309G 16318T 16519C
13	PAK_13	HV12b1	150T 263G 315.1C 750G 1438G 2706G 4769G 7028T 7852A 8860G 11204C 12618A 13889A 15326G 15682G 16222T 16242T 16273A 16356C
14	PAK_14	U2c1b	73G 146C 152C 263G 315.1C 644G 709A 750G 1438G 1598A 1811G 2706G 4721G 4769G 5790A 7028T 8023C 8676T 8860G 9767T 10810C 11467G 11719A 11890G 12172G 12308G 12372A 14766T 14935C 15061G 15214C 15326G 16051G 16189C 16234T
15	PAK_15	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16519C
16	PAK_16	M30+16234	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11437C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16093C 16223T 16234T 16274A 16519C
17	PAK_17	H13a2a1	263G 315.1C 455.1T 709A 750G 1008G 1438G 2259T 3450A 4769G 8843C 8860G 14872T 15326G 16519C
18	PAK_18	M30d1	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 1598A 2706G 4769G 5557C 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15259T 15301A 15326G 15431A 16179del 16223T 16519C
19	PAK_19	M45	73G 146C 263G 315.1C 489C 750G 1438G 2706G 3083C 4059A 4734G 4769G 5319G 5585A 6827C 7028T 8701G 8860G 9095C 9180G 9509C 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14687G 14766T 14783C 15043A 15301A 15326G 15851G 16145A 16192T 16223T 16300G 16316G 16519C
20	PAK_20	M3c2	73G 93G 263G 315.1C 482C 489C 523del 524del 750G 1438G 2706G 4769G 5178T 7028T 8701G 8860G 9064A 9540C 10398G 10400T 10873C 11719A 12705T 13350G 14766T 14783C 15043A 15301A 15326G 16126C 16154C 16223T 16519C
21	PAK_21	M5a2	73G 195C 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4454C 4769G 5027T 6053T 6293C 6473T 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11275T 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16051G 16093C 16129A 16223T 16519C

22	PAK_22	U1a1c1	73G 263G 285T 315.1C 523del 524del 750G 1438G 2218T 2706G 4769G 4991A 6026A 7028T 7051C 7581C 8860G 11467G 11719A 12308G 12372A 12879C 13104G 14070G 14364A 14514C 14766T 15115C 15148A 15217A 15326G 15954C 16093C 16182C 16183C 16189C 16249C 16311C
23	PAK_23	U5b2	73G 150T 152C 257G 263G 264T 315.1C 750G 1438G 1721T 2706G 3197C 4080C 4769G 7028T 7768G 8860G 9139A 9477A 11467G 11719A 12308G 12372A 13434G 13617C 13637G 14182C 14199C 14409G 14766T 15326G 16270T
24	PAK_24	W3a1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1406C 1438G 2706G 3505G 4370C 4769G 5046A 5460A 7028T 8251A 8860G 8994A 11674T 11719A 11947G 12414C 12705T 13263G 14766T 15326G 15784C 15884C 16093C 16129A 16223T 16292T 16519C
25	PAK_25	U2b2	73G 146C 152C 234G 263G 315.1C 750G 1438G 1811G 1888A 4769G 5186T 7028T 8860G 9094T 9614G 11467G 11719A 12106T 12308G 12372A 12793C 13194A 13305T 13656C 14766T 15049T 15326G 15813C 15930A 16051G 16129A 16209C 16239T 16311C 16352C 16353T
26	PAK_26	M33a2a	73G 150T 263G 315.1C 462T 489C 750G 1438G 2361A 2706G 3543T 5124A 5423G 7028T 8176C 8562T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12705T 13731G 14766T 14783C 15043A 15301A 15326G 15908C 16169T 16172C 16223T 16278T 16519C
27	PAK_27	F1c1a2	73G 234G 249del 263G 315.1C 523del 524del 750G 1438G 1927A 2706G 3970T 4769G 6392C 6599G 6962A 7028T 8860G 9053A 9822T 10310A 10454C 10609C 11719A 12406A 12882T 13759A 13928C 14766T 15326G 16111T 16129A 16304C 16519C
28	PAK_28	M	73G 195C 207A 263G 315.1C 482C 489C 750G 1438G 2706G 3394C 4769G 7028T 8701G 8860G 9341G 9540C 10398G 10400T 10873C 11167G 11719A 11914A 12705T 14766T 14783C 15043A 15301A 15326G 15766G 15951G 16126C 16185T 16223T 16519C
29	PAK_29	M37e2	73G 152C 263G 315.1C 489C 750G 1438G 2706G 4769G 7028T 8410T 8701G 8860G 9540C 10398G 10400T 10556T 10873C 11050C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16111T 16189C 16223T 16224C 16294G 16295T 16519C
30	PAK_30	M5a2a1a1	73G 263G 315.1C 374G 489C 709A 750G 1189C 1438G 1888A 2706G 3921T 4454C 4769G 7028T 8701G 8860G 8886A 9540C 9947A 10398G 10400T 10873C 11719A 12372A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16129A 16223T 16265C 16344T 16519C
31	PAK_31	M33d	73G 152C 204C 207A 263G 315.1C 489C 513A 750G 1438G 2361A 2706G 3116T 4024G 4769G 6563T 7028T 8668C 8701G 8860G 9540C 9966A 10398G 10400T 10873C 10969T 11719A 12705T 13788T 14766T 14783C 14950T 15043A 15301A 15326G 15924G 16178C 16223T 16288C 16519C
32	PAK_32	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16092C 16519C

33	PAK_33	R2d	73G 152C 263G 315.1C 750G 1438G 2706G 4216C 7028T 7657C 8143C 8473C 8860G 9932A 10685A 11719A 12654G 13434G 13500C 13914A 14305A 14766T 15326G 16071T 16519C
34	PAK_34	M30+16234	73G 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11437C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16223T 16234T 16274A 16519C
35	PAK_35	M5a2a1a1	73G 263G 315.1C 374G 489C 709A 750G 1189C 1438G 1888A 2706G 3921T 4454C 4769G 7028T 8701G 8860G 8886A 9540C 9947A 10398G 10400T 10873C 11719A 12372A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16129A 16223T 16265C 16344T 16519C
36	PAK_36	M4a	73G 263G 315.1C 489C 750G 1438G 2706G 3866C 4769G 5899.CCC 6620C 7028T 7859A 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16145A 16176T 16223T 16261T 16262T 16266T 16291T 16311C 16325C 16519C
37	PAK_37	U2c1b	73G 146C 152C 263G 315.1C 644G 709A 750G 1438G 1598A 1811G 2706G 3705R 4721G 4769G 5790A 7028T 8023C 8676T 8860G 9767T 10810C 11467G 11719A 11890G 12172G 12308G 12372A 14766T 14935C 15061G 15214C 15326G 16051G 16189C 16234T 16519C
38	PAK_38	U2b2	73G 146C 152C 234G 315.1C 750G 1438G 1811G 1888A 2852G 4113A 4541A 4769G 5186T 7028T 8860G 9094T 9614G 11467G 11719A 12106T 12308G 12372A 12793C 13194A 13656C 14766T 15049T 15326G 15930A 16051G 16209C 16239T 16352C 16353T
39	PAK_39	T2c	73G 263G 315.1C 709A 750G 1438G 1888A 2706G 3606G 4216C 4769G 4917G 7028T 7903G 8697A 8860G 10463C 10822T 11215T 11251G 11719A 11812G 13368A 13818C 14233G 14560A 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16294T 16296T 16400T 16519C 16527T
40	PAK_40	U2e1h	73G 150T 217C 228A 263G 315.1C 340T 508G 750G 1438G 1811G 2706G 3202C 3720G 4769G 5390G 5426C 6045T 6152C 7028T 8860G 10876G 11467G 11719A 12308G 12372A 13020C 13734C 14766T 15326G 15907G 16051G 16129C 16183C 16362C 16519C
41	PAK_41	M3b	73G 263G 315.1C 482C 489C 750G 1438G 2706G 4769G 6353G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12031A 12705T 14766T 14783C 15043A 15301A 15326G 16126C 16223T 16344T 16519C
42	PAK_42	T1a	73G 263G 315.1C 709A 750G 1438G 1888A 2706G 3537G 4216C 4769G 4917G 5123G 7028T 8697A 8860G 10463C 11251G 11719A 12633A 13368A 14766T 14905A 15326G 15452A 15607G 15928A 16126C 16163G 16186T 16189C 16294T 16445C 16519C
43	PAK_43	M30c1	73G 146C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4014T 4769G 5249C 7028T 7269A 8251A 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12234G 12236A 12705T 14569A 14766T 14783C 15043A 15301A 15326G 15431A 16166del 16223T 16519C

44	PAK_44	T1a1	73G 152C 195C 263G 315.1C 709A 750G 1438G 1888A 2706G 4216C 4769G 4917G 7028T 8697A 8860G 9899C 10463C 11251G 11719A 12633A 13368A 14766T 14905A 15326G 15452A 15607G 15928A 16093C 16126C 16163G 16186T 16189C 16294T 16519C
45	PAK_45	M33a2	73G 263G 315.1C 462T 489C 750G 1438G 2361A 2706G 4769G 5042G 5423G 6677G 7028T 7624C 8562T 8860G 9045G 9540C 10398G 10400T 10873C 11025C 11719A 12705T 13731G 14766T 14783C 15043A 15301A 15326G 15908C 16124C 16169T 16172C 16223T 16519C
46	PAK_46	HV2a2	72C 73G 152C 182T 195C 263G 315.1C 523del 524del 750G 1438G 2706G 4769G 5153G 7028T 7193C 7861C 8860G 9336G 11935C 12061T 15326G 16217C 16325C
47	PAK_47	M30g	73G 195A 204C 207A 263G 309T 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 6119T 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16209C 16223T 16519C
48	PAK_48	HV2a2	72C 73G 152C 195C 263G 315.1C 523del 524del 750G 1438G 2706G 4769G 5153G 5747G 7028T 7193C 7861C 8860G 9336G 11935C 12061T 15326G 15459T 16172C 16217C
49	PAK_49	M3c2	73G 263G 315.1C 482C 489C 523del 524del 750G 1438G 2706G 4601G 4769G 6467A 7028T 8701G 8860G 9064A 9540C 10398G 10400T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 16126C 16154C 16223T 16519C
50	PAK_50	U7a	73G 151T 152C 263G 315.1C 523del 524del 750G 980C 1438G 1811G 2706G 3705A 3741T 4733C 4769G 4947C 5360T 7028T 8137T 8684T 8860G 10142T 11467G 11719A 12308G 12372A 13500C 14569A 14766T 15326G 15448A 16309G 16318T 16519C

Appendix VIII: Whole mitochondrial DNA sequences using MPS for 50 Indian samples with the haplotypes nomenclatures

Whole mitochondrial DNA sequences using MPS for 50 Indian samples.

No.	Sample	Haplogroup	Haplotypes
1	IND_01	M5a2a	73G 236C 263G 315.1C 489C 709A 742C 750G 1438G 1888A 2706G 2833G 3921T 4454C 4769G 4907C 6062T 6293C 6378C 7028T 8158G 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16223T 16519C
2	IND_02	R8b1	73G 150T 154C 195C 263G 315.1C 455.1T 456T 750G 1438G 1709A 2706G 2746C 2755G 3384G 4769G 5096C 6485G 7028T 7759C 8860G 9449T 9758C 11719A 12007A 13194A 13215C 14766T 15250T 15326G 16172C 16390A 16519C
3	IND_03	N1a1b	73G 143A 199C 204C 250C 263G 297G 315.1C 573.CC 710C 750G 1438G 1719A 2706G 4529T 4769G 4947C 6671C 7028T 8251A 8860G 9804A 10238C 10398G 10790C 10882C 11719A 12501A 12705T 13780G 14766T 15043A 15326G 15924G 15954G 16223T 16311C 16519C
4	IND_04	HV2a2	72C 73G 152C 195C 263G 315.1C 523del 524del 709A 750G 1438G 2706G 4769G 5153G 6366A 7028T 7193C 7861C 8860G 9336G 10306G 11935C 12061T 15326G 16217C
5	IND_05	R30b2a	73G 152C 263G 315.1C 373G 750G 1438G 2706G 2831A 4769G 6290T 7028T 7280T 7843G 8584A 8860G 11719A 11916A 12028C 13539G 14000A 14766T 15148A 15326G 16497G 16519C 16524G
6	IND_06	U7b	73G 152C 263G 315.1C 523del 524del 750G 980C 1438G 1811G 2706G 3741T 4769G 4851T 5360T 7028T 8137T 8684T 8860G 10053T 10084C 10142T 11467G 11719A 12308G 12372A 13500C 14569A 14766T 14971C 15326G 16309G 16318T 16519C
7	IND_07	U2a1a	73G 263G 315.1C 750G 1438G 1811G 2706G 4769G 7028T 8572A 11368C 11467G 11719A 12308G 12372A 13708A 14766T 15326G 16051G 16093G 16154C 16206C 16230G 16311C
8	IND_08	U2c1	73G 152C 239C 247A 263G 315.1C 750G 1438G 1811G 2706G 3915A 4769G 5790A 7028T 8023C 8676T 8860G 9692G 9767T 11467G 11719A 12308G 12361G 12372A 13368A 14766T 14935C 15061G 15326G 16051G 16129A 16179T 16234T 16247G 16519C
9	IND_09	U9a1	73G 200G 263G 315.1C 499A 750G 1438G 1811G 2706G 3290C 3531A 3834A 4769G 5351G 5999C 6386T 7028T 8860G 11467G 11719A 12308G 12372A 14094C 14766T 15077A 15326G 16051G 16193T 16234T 16278T 16357C
10	IND_10	M3a2	73G 263G 315.1C 482C 489C 750G 1438G 2706G 4580A 4769G 5783A 6359G 7028T 8701G 8860G 8950A 9540C 10398G 10400T 10727T 10873C 11719A 12705T 14766T 14783C 15043A 15169G 15301A 15326G 16051G 16126C 16223T 16278T 16519C

11	IND_11	M5a	73G 263G 315.1C 489C 709A 750G 1438G 1888A 2706G 3921T 4721G 4769G 7028T 8701G 8860G 9540C 9773T 9947A 10398G 10400T 10873C 11719A 12477C 12705T 13708A 14323A 14766T 14783C 15043A 15301A 15326G 15927A 16129A 16223T 16519C
12	IND_12	U7a3a	73G 151T 152C 263G 315.1C 523del 524del 750G 824C 980C 1438G 1811G 2706G 2863C 3741T 4769G 5360T 6620C 7028T 8137T 8684T 8860G 9852G 10142T 11467G 11719A 12308G 12372A 12618A 13500C 14569A 14766T 15326G 16069T 16274A 16318T 16519C
13	IND_13	A17	73G 152C 234G 235G 263G 315.1C 523del 524del 663G 750G 1438G 1736G 2706G 4113A 4248C 4592C 4769G 4824G 5147A 5514G 7028T 8794T 8860G 9126C 11719A 12705T 14766T 15217A 15326G 16172C 16173T 16223T 16235G 16290T 16311C 16319A 16362C 16519C
14	IND_14	M2a'b	73G 143A 195C 263G 315.1C 337G 447G 489C 750G 1007A 1438G 1780C 2706G 3432T 4769G 6647G 7028T 7337A 8212T 8502G 8701G 8860G 9540C 9899C 10398G 10400T 10873C 11083G 11518A 11719A 12705T 14766T 14783C 14861A 15043A 15253G 15301A 15326G 15670C 15721C 16223T 16274A 16319A 16320T 16518T 16519C
15	IND_15	M5a	73G 263G 315.CC 489C 709A 750G 1438G 1888A 2706G 3921T 4314A 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15301A 15326G 16129A 16223T 16519C
16	IND_16	U2a2	73G 263G 315.1C 750G 1095C 1438G 1760A 1811G 2393T 2706G 3316A 4769G 4970G 5021C 5201C 6116G 7028T 7859A 8860G 11299C 11467G 11719A 12308G 12372A 12477C 12561A 14182C 14766T 14883T 15326G 16051G 16206C
17	IND_17	H	73G 263G 315.1C 750G 1438G 4769G 5120G 5216T 8860G 15326G 15439T 16051G 16126C 16519C
18	IND_18	M36a	73G 152C 239C 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3783T 4769G 6320C 6917A 7028T 7271G 8701G 8860G 9540C 10398G 10400T 10873C 11608G 11719A 12346T 12705T 13831T 14203G 14302C 14766T 14783C 15043A 15110A 15301A 15326G 16193T 16223T 16519C
19	IND_19	U2a1a	73G 151T 263G 315.1C 750G 1438G 1811G 2706G 4769G 7028T 8572A 11368C 11467G 11719A 12308G 12372A 13708A 14766T 15326G 16051G 16093G 16154C 16206C 16230G 16311C
20	IND_20	M6b	73G 146C 152C 263G 315.1C 461T 489C 523del 524del 750G 1438G 2706G 3254A 3444T 4216C 4417G 4769G 5301G 5558G 7028T 8281del 8282del 8283del 8284del 8285del 8286del 8287del 8288del 8289del 8701G 8860G 9540C 10321C 10398G 10400T 10640C 10667C 10873C 11719A 12634G 12705T 13161C 14128G 14696G 14766T 14783C 15043A 15301A 15326G 16184T 16223T 16256G 16311C 16362C
21	IND_21	M39b1	56del 58A 65.1T 66T 73G 153G 263G 315.1C 463T 485C 489C 750G 1438G 1811G 2706G 3531A 4769G 6257A 7028T 8567C 8679G 8701G 8860G 9374G 9540C 9655A 10398G 10400T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 15938T 16223T
22	IND_22	R6a2	73G 263G 315.1C 750G 1438G 2706G 4769G 5021C 5894G 7028T 7897A 8860G 11075C 11719A 12133T 12285C 14058T 14766T 15067C 15326G 16129A 16158G 16213A 16362C 16519C

23	IND_23	W3a1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1406C 1438G 2706G 3505G 4370C 4769G 5046A 5460A 7028T 8251A 8860G 8994A 11674T 11719A 11947G 12414C 12705T 13263G 14766T 15326G 15784C 15884C 16129A 16223T 16292T 16519C
24	IND_24	M2a1	73G 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 5147C 5252A 6752G 7028T 7961C 8396G 8502G 8701G 8853G 8860G 9540C 9758C 10398G 10400T 10873C 11016A 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16172C 16223T 16224C 16270T 16274A 16319A 16352C 16519C 16524C
25	IND_25	M44a	73G 146C 263G 315.1C 489C 750G 930A 961C 1438G 2706G 4769G 7028T 8179G 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 16223T 16301T
26	IND_26	M65a+@16311	73G 263G 315.1C 489C 511T 750G 1438G 2706G 4769G 4916G 7028T 8047C 8701G 8860G 9254G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 13651G 14766T 14783C 15043A 15301A 15326G 16223T 16289G 16519C
27	IND_27	M2a1	73G 183G 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 5252A 7028T 7961C 8396G 8502G 8701G 8860G 9095C 9540C 9758C 9779T 10398G 10400T 10873C 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16223T 16270T 16274A 16319A 16352C 16519C
28	IND_28	N1a1b1	73G 143A 199C 204C 250C 263G 315.1C 710C 750G 1438G 1719A 2706G 4529T 4769G 4947C 6671C 7028T 8251A 8860G 9804A 10238C 10398G 10790C 10882C 11719A 12501A 12705T 13780G 14766T 15043A 15326G 15924G 15954G 16223T 16311C 16519C
29	IND_29	M2a1a	73G 195C 204C 263G 315.1C 447G 489C 750G 1438G 1780C 2706G 4769G 4965G 5252A 7028T 7604A 7961C 8396G 8502G 8701G 8860G 9540C 9758C 9965C 10398G 10400T 10873C 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15301A 15326G 15670C 16223T 16270T 16288C 16319A 16352C 16519C
30	IND_30	M30c1	73G 146C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4769G 7028T 8251A 8701G 8860G 9540C 9797C 10398G 10400T 10873C 11719A 12007A 12234G 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16166del 16223T 16519C
31	IND_31	M6	73G 152C 263G 315.1C 461T 489C 523del 524del 593C 750G 1438G 2706G 3745A 4096T 4380T 4418C 4592C 4769G 5301G 5558G 6746T 7028T 7804G 7975G 8860G 9540C 10398G 10400T 10640C 10751T 10873C 11719A 12705T 13830C 13998T 14128G 14766T 14783C 15043A 15301A 15326G 16223T 16362C
32	IND_32	M36a	73G 152C 239C 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3783T 4769G 6320C 6917A 7028T 7271G 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12346T 12705T 14203G 14302C 14766T 14783C 15043A 15110A 15301A 15326G 16193T 16223T 16324C
33	IND_33	M30	73G 152C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3866C 4769G 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 15431A 16145A 16223T 16519C

34	IND_34	M5a1b	73G 263G 315.1C 489C 709A 750G 1303A 1438G 1888A 2706G 3921T 3954T 4769G 4916G 6461G 7028T 7678C 8215T 8701G 8860G 9540C 9833C 10398G 10400T 10873C 11719A 11809C 12477C 12705T 13130A 14323A 14766T 14783C 15043A 15287C 15301A 15326G 16129A 16223T 16291T 16519C
35	IND_35	R30a1c	73G 263G 315.1C 524.ACAC 750G 1438G 2056A 2706G 3316A 4232C 4769G 5442C 6764A 7028T 8584A 8860G 9142A 9156G 9242G 9869G 10199T 11047A 11719A 12714C 13161C 13773G 14766T 15055C 15326G 15970C 16172C 16519C
36	IND_36	M30c1	73G 146C 152C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 3372C 4769G 5121T 7028T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12007A 12234G 12705T 14766T 14783C 15043A 15208G 15301A 15326G 15431A 16166d 16223T 16257T 16519C
37	IND_37	M2a1	73G 183G 204C 263G 315.1C 447G 750G 1438G 1780C 2706G 4769G 5252A 7028T 7961C 8396G 8502G 8701G 8860G 9095C 9540C 9758C 9779T 10398G 10400T 10873C 11083G 11719A 12705T 12810G 14766T 14783C 15043A 15227A 15301A 15326G 15670C 16223T 16270T 16274A 16319A 16352C 16519C
38	IND_38	M3a1b	73G 152C 204C 217C 263G 315.1C 482C 489C 750G 1438G 2706G 4580A 4703C 4769G 7028T 8701G 8860G 9540C 9554A 10398G 10400T 10845T 10873C 11719A 12705T 14766T 14783C 15043A 15301A 15326G 15697C 16126C 16223T 16224C 16311C 16519C
39	IND_39	M65b	73G 241G 263G 315.1C 489C 511T 750G 1438G 2706G 3398C 4769G 7028T 7598A 8701G 8860G 8865A 9540C 10398G 10400T 10873C 11719A 12007A 12705T 14766T 14783C 15043A 15301A 15326G 16172C 16181G 16223T 16311C 16519C
40	IND_40	U1a1c1d	73G 152C 263G 285T 307d 308d 309d 315.1C 494G 750G 1438G 2218T 2706G 4769G 4991A 6026A 7028T 7581C 8860G 10253C 11467G 11719A 12308G 12372A 12879C 13104G 14070G 14364A 14766T 15115C 15148A 15217A 15326G 15954C 16182C 16183C 16189C 16249C 16519C 16527T
41	IND_41	U2c1	73G 152C 263G 315.1C 750G 1438G 1811G 2706G 4769G 4991A 5790A 7028T 7684C 8023C 8676T 8860G 9767T 11467G 11719A 12308G 12372A 14766T 14935C 15061G 15326G 16051G 16234T 16264T
42	IND_42	U9a1	73G 200G 263G 315.1C 499A 750G 1438G 1811G 2706G 3290C 3531A 3834A 4769G 5999C 6386T 7028T 8027A 8860G 11467G 11719A 12308G 12372A 14094C 14766T 15077A 15326G 16051G 16193T 16278T 16357C
43	IND_43	M2b1	73G 152C 182T 195C 198T 199C 263G 315.1C 447G 750G 1438G 1453G 1780C 2706G 2831T 3630T 4769G 5420C 5744A 6260A 6647G 7028T 8502G 8701G 8860G 9540C 9899C 10398G 10400T 10873C 11083G 11719A 12705T 13254C 14783C 15043A 15301A 15326G 15670C 16104T 16189C 16223T 16274A 16319A 16320T 16519C
44	IND_44	W3a1	73G 189G 194T 195C 204C 207A 263G 315.1C 709A 750G 1243C 1406C 1438G 2706G 3505G 4370C 4769G 5046A 5460A 7028T 8251A 8860G 8994A 11674T 11719A 11947G 12414C 12705T 13263G 14766T 15326G 15784C 15884C 16129A 16223T 16292T 16519C

45	IND_45	M5a2a2	73G 234G 315.1C 489C 499A 709A 750G 1438G 1888A 2706G 3399G 3921T 4454C 4769G 7028T 8860G 9540C 10398G 10400T 10873C 11719A 12477C 12705T 14323A 14766T 14783C 15043A 15262C 15301A 15326G 16129A 16144A 16223T 16265C 16362C 16519C
46	IND_46	J1d	73G 152C 263G 295T 315.1C 462T 489C 750G 1438G 2706G 3010A 3523G 4216C 4769G 5147A 7028T 7789A 7963G 8860G 10398G 11251G 11719A 12346T 12612G 13708A 14066T 14766T 15326G 15452A 16069T 16114T 16126C 16193T 16519C
47	IND_47	HV+16311	143A 227G 315.1C 593C 750G 1438G 2706G 3540C 4769G 6584T 7028T 8426C 8860G 9592C 15326G 15766G 16311C
48	IND_48	M30b	73G 152C 195A 263G 315.1C 489C 523del 524del 750G 1438G 2706G 4586C 4769G 5147A 7028T 7129G 8650G 8701G 8860G 9428C 9540C 10398G 10400T 10873C 11719A 12007A 12705T 13980A 14766T 14783C 15043A 15301A 15326G 15431A 16192T 16223T 16239T 16278T 16519C
49	IND_49	U2c1	73G 152C 263G 315.1C 750G 1438G 1811G 2706G 4769G 4991A 5790A 7028T 7684C 8023C 8676T 8860G 9767T 11467G 11719A 12308G 12372A 14766T 14935C 15061G 15326G 16051G 16234T 16264T
50	IND_50	M18c	73G 246C 315.1C 489C 750G 1095C 1438G 2706G 3394C 4769G 7028T 7853A 7861C 8108G 8206A 8701G 8860G 9540C 9612A 10373A 10398G 10400T 10873C 11719A 12007A 12405A 12498T 12705T 13135A 14696G 14766T 14783C 15043A 15301A 15326G 16223T 16311C 16318T 16519C

Appendix IX: Whole mtDNA haplogroups for 510 Emirati samples

MPS generated haplogroups in the Emiratis samples.

No.	Haplogroup	Frequency	PM
1	A	0.00784	0.0000615
2	B4h1	0.00196	0.0000038
3	B5b1c	0.00196	0.0000038
4	D4b2b4*	0.00196	0.0000038
5	D4t	0.00980	0.0000961
6	F1a1a	0.00196	0.0000038
7	F3b1	0.00196	0.0000038
8	H1+152	0.00392	0.0000154
9	H1+16355	0.00196	0.0000038
10	H11a	0.00392	0.0000154
11	H13a2c	0.00392	0.0000154
12	H13b1+200	0.00392	0.0000154
13	H14a	0.00392	0.0000154
14	H14b1	0.00980	0.0000961
15	H15a1b	0.00196	0.0000038
16	H1a	0.00392	0.0000154
17	H1ah2	0.00392	0.0000154
18	H1b	0.00392	0.0000154
19	H1e+16129	0.00392	0.0000154
20	H1e1a4	0.00196	0.0000038
21	H1e1a6	0.00196	0.0000038
22	H1e1a8	0.00196	0.0000038
23	H1q3	0.00196	0.0000038
24	H2+152+16311	0.00196	0.0000038
25	H2a*	0.00784	0.0000615
26	H2a1	0.00196	0.0000038
27	H2a2a	0.02549	0.0006498
28	H2a2a1	0.01176	0.0001384
29	H2a2a1c	0.00196	0.0000038
30	H2a2a1d	0.00196	0.0000038
31	H2a2a1f	0.00392	0.0000154
32	H2a2a1g	0.00588	0.0000346
33	H2a2a2	0.00196	0.0000038
34	H32	0.00196	0.0000038
35	H3c2	0.00196	0.0000038
36	H3h7	0.00588	0.0000346
37	H3z	0.00196	0.0000038
38	H5	0.00980	0.0000961
39	H5'36	0.00196	0.0000038

40	H57	0.01176	0.0001384
41	H5a1+152	0.00196	0.0000038
42	H5a5	0.00196	0.0000038
43	H5r	0.00196	0.0000038
44	H6	0.01373	0.0001884
45	H6b2	0.00784	0.0000615
46	H8+(114)	0.00392	0.0000154
47	HV1	0.00196	0.0000038
48	HV14	0.00392	0.0000154
49	HV15	0.00392	0.0000154
50	HV15*1	0.00196	0.0000038
51	HV21	0.00392	0.0000154
52	HV2a	0.00196	0.0000038
53	HV6	0.00784	0.0000615
54	I*	0.00392	0.0000154
55	I1	0.00196	0.0000038
56	I1c1	0.00196	0.0000038
57	I2'3	0.00196	0.0000038
58	I5a	0.00196	0.0000038
59	I6b	0.00196	0.0000038
60	J1	0.00196	0.0000038
61	J1b	0.00392	0.0000154
62	J1b1a1	0.00588	0.0000346
63	J1b1a1a	0.00196	0.0000038
64	J1b1b	0.01176	0.0001384
65	J1b3a	0.00196	0.0000038
66	J1b5a	0.00588	0.0000346
67	J1b8	0.00980	0.0000961
68	J1c	0.01373	0.0001884
69	J1c17a	0.00196	0.0000038
70	J1c2	0.00392	0.0000154
71	J1d	0.00588	0.0000346
72	J1d1	0.00196	0.0000038
73	J1d1a	0.00588	0.0000346
74	J2a	0.00196	0.0000038
75	J2a2a	0.00196	0.0000038
76	J2a2b	0.01765	0.0003114
77	J2b	0.00392	0.0000154
78	K1a	0.00392	0.0000154
79	K1a*1	0.00392	0.0000154
80	K1a1	0.00392	0.0000154
81	K1a2a1	0.00392	0.0000154
82	K1a4a*2	0.00196	0.0000038
83	K1a4a1h*	0.00392	0.0000154

84	K1a4c1*1	0.00392	0.0000154
85	K1a4c1*1a	0.00392	0.0000154
86	K1b1c	0.00392	0.0000154
87	K1c	0.00196	0.0000038
88	K2	0.00392	0.0000154
89	K2a	0.00784	0.0000615
90	K2a5	0.00196	0.0000038
91	K2b1b	0.00392	0.0000154
92	L0a1a2	0.00196	0.0000038
93	L0a2a2	0.01176	0.0001384
94	L0d1'2	0.00196	0.0000038
95	L1b	0.00196	0.0000038
96	L1b1a+189	0.00196	0.0000038
97	L1c2a1a	0.00196	0.0000038
98	L1c2b1b*1	0.00196	0.0000038
99	L1c2b2	0.00196	0.0000038
100	L1c3a	0.00392	0.0000154
101	L1c3b1a	0.00196	0.0000038
102	L2a1+143+16189+(16192)+@16309	0.00196	0.0000038
103	L2a1a2	0.00196	0.0000038
104	L2a1a3	0.00196	0.0000038
105	L2a1b+143	0.00196	0.0000038
106	L2a1b1a	0.00196	0.0000038
107	L2a1h*	0.00784	0.0000615
108	L2b1a	0.00588	0.0000346
109	L3'4'6	0.00196	0.0000038
110	L3b	0.00392	0.0000154
111	L3b1a+152*2	0.00196	0.0000038
112	L3d1a1a	0.01373	0.0001884
113	L3e'i'k'x	0.00196	0.0000038
114	L3e1	0.00196	0.0000038
115	L3e1b2	0.00196	0.0000038
116	L3e2b	0.00980	0.0000961
117	L3f1b4*	0.00392	0.0000154
118	L3h2	0.00196	0.0000038
119	L4	0.00196	0.0000038
120	M	0.00588	0.0000346
121	M1	0.00196	0.0000038
122	M18'38	0.00196	0.0000038
123	M1a1	0.00196	0.0000038
124	M1a5	0.00196	0.0000038
125	M23	0.00196	0.0000038
126	M2b	0.00196	0.0000038
127	M2b1a	0.00196	0.0000038

128	M3	0.00196	0.0000038
129	M30	0.00980	0.0000961
130	M30*1	0.00392	0.0000154
131	M30+16234	0.00980	0.0000961
132	M30b	0.00196	0.0000038
133	M30c1	0.00196	0.0000038
134	M30d	0.00588	0.0000346
135	M30g	0.00196	0.0000038
136	M33a1b	0.00196	0.0000038
137	M33a2a	0.00196	0.0000038
138	M36a	0.00196	0.0000038
139	M38+195	0.00196	0.0000038
140	M39b1*1	0.00196	0.0000038
141	M3a1+204	0.00196	0.0000038
142	M3c2	0.00196	0.0000038
143	M3d	0.00196	0.0000038
144	M40a1a	0.00196	0.0000038
145	M45a	0.00196	0.0000038
146	M49	0.00196	0.0000038
147	M4a	0.00392	0.0000154
148	M57+152	0.00196	0.0000038
149	M5a1b*	0.00196	0.0000038
150	M5a2a1a	0.00392	0.0000154
151	M5a2a2	0.00196	0.0000038
152	M5a2a2*	0.00196	0.0000038
153	M5a2a4	0.00196	0.0000038
154	M5b2	0.00196	0.0000038
155	M65b	0.00196	0.0000038
156	M6a1a	0.00196	0.0000038
157	N	0.00980	0.0000961
158	N1a1a	0.00588	0.0000346
159	N1a1b1	0.00196	0.0000038
160	N1a3a	0.00980	0.0000961
161	N1b1	0.03922	0.0015379
162	N1b1a+16129	0.01569	0.0002461
163	N1b1a4	0.00196	0.0000038
164	N2a	0.00196	0.0000038
165	P2	0.00392	0.0000154
166	R0a	0.00196	0.0000038
167	R0a1+152	0.00784	0.0000615
168	R0a1a	0.01176	0.0001384
169	R0a1a1a	0.00196	0.0000038
170	R0a1a3*	0.00588	0.0000346
171	R0a2h	0.00588	0.0000346

172	R1	0.00196	0.0000038
173	R12'21	0.00196	0.0000038
174	R2	0.00392	0.0000154
175	R2+13500+195	0.00392	0.0000154
176	R2c	0.00196	0.0000038
177	R30a1b	0.00588	0.0000346
178	R30a1c	0.00196	0.0000038
179	R30b2	0.00196	0.0000038
180	R5a1	0.00196	0.0000038
181	R5a2	0.00392	0.0000154
182	R6a2*	0.00196	0.0000038
183	R8	0.00196	0.0000038
184	R8a1a1a2	0.00196	0.0000038
185	R8a1a1d	0.00196	0.0000038
186	R8b2	0.00588	0.0000346
187	R9	0.00196	0.0000038
188	T	0.00588	0.0000346
189	T1	0.00392	0.0000154
190	T1a	0.00588	0.0000346
191	T1a1'3	0.00784	0.0000615
192	T1a1m	0.00196	0.0000038
193	T1a2a	0.00392	0.0000154
194	T2a1a*	0.01176	0.0001384
195	T2b	0.00196	0.0000038
196	T2b7a2	0.00196	0.0000038
197	T2c1a	0.00196	0.0000038
198	T3	0.00392	0.0000154
199	U1a	0.00196	0.0000038
200	U1a3	0.00196	0.0000038
201	U2a	0.00784	0.0000615
202	U2b	0.00196	0.0000038
203	U2b2	0.00588	0.0000346
204	U2c'd	0.00392	0.0000154
205	U2d	0.00196	0.0000038
206	U2e1	0.00588	0.0000346
207	U2e1'2'3	0.01765	0.0003114
208	U2e1f	0.00196	0.0000038
209	U2e3	0.00196	0.0000038
210	U3	0.00196	0.0000038
211	U3a	0.00196	0.0000038
212	U3a2a1	0.00196	0.0000038
213	U3b3	0.00392	0.0000154
214	U4b1+146+152	0.00392	0.0000154
215	U5a1+@16192	0.00196	0.0000038

216	U5a2a	0.00196	0.0000038
217	U5b	0.00392	0.0000154
218	U5b1d1a	0.00588	0.0000346
219	U5b2a1a+16311	0.00392	0.0000154
220	U5b2b4a	0.00392	0.0000154
221	U6	0.00196	0.0000038
222	U6a+16189+(103)	0.00196	0.0000038
223	U7	0.00196	0.0000038
224	U7a	0.01176	0.0001384
225	U7a3b*	0.01373	0.0001884
226	U7a4	0.00196	0.0000038
227	U9a	0.00392	0.0000154
228	U9a1	0.00196	0.0000038
229	U9b1	0.00196	0.0000038
230	V1a1b*	0.00392	0.0000154
231	W	0.00196	0.0000038
232	W+194	0.00392	0.0000154
233	W1+119	0.00392	0.0000154
234	W6	0.00392	0.0000154
235	X	0.00196	0.0000038
236	X2+225	0.00784	0.0000615
237	X2+225+@16223	0.00196	0.0000038
238	X2b+226	0.00196	0.0000038
239	X2c1b	0.00196	0.0000038
240	X2d1	0.00196	0.0000038
241	X2j	0.00196	0.0000038

Appendix X: Whole mitochondrial DNA sequences using MPS for 510 UAE samples with the haplotypes nomenclatures in a supplementary Excel file provided separately.