MATRILINEAL GENETIC DIVERSITY AND FORENSIC CHARACTERISATION OF HAUSA POPULATION IN NIGERIA

A Thesis Submitted

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN

FORENSIC SCIENCE

By

IBRAHIM EL-LADAN SHEHU
Registration No.-20SBAS3010002

Supervisor

DR. GAURAV KUMAR
Associate Professor
Division of Clinical Research,
Department of Biosciences,
School of Basic and Applied Sciences

GALGOTAS UNIVERSITY

GREATER NOIDA, UTTAR PRADESH

JUNE-2023

Candidate's Declaration

I hereby certify that the work which is being presented in the thesis, entitled **"Matrilineal Genetic Diversity and Forensic Characterisation of Hausa Population in Nigeria"** in fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Forensic Science and submitted in School of Basic and Applied Sciences Galgotias University, Greater Noida is an authentic record of my own work carried out during a period from July, 2020 to May, 2023 under the supervision of Dr. GAURAV KUMAR, Associate Professor, School of Basic & Applied Sciences.

The matter embodied in this thesis has not been submitted by me for the award of any other degree of this or any other University/Institute.

(IBRAHIM EL-LADAN SHEHU)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Dr. GAURAV KUMAR)
Supervisor
Division of Clinical Research
Dept. of Biosciences,
School of Basic and Applied Sciences
Galgotias University, Greater Noida,
UP, INDIA

The Ph.D. Viva-Voice examination of IBRAHIM EL-LADAN SHEHU, Research Scholar has been held on_____

Sign. of Supervisor(s) Sign. of External Examiner

Abstract

Mitochondrial DNA (mtDNA) plays a crucial role in forensic science and population genetics, providing valuable insights into human evolution, ancestry, and disease susceptibility. African genetic diversity holds particular significance in understanding phylogeny, forensic investigations, and precision medicine. This study aims to investigate the control regions (HVR1 and HVR2) of mtDNA among a sample of 100 Hausa individuals from Nigeria.

Buccal swabs were collected, and DNA extraction was performed performed using QiaGen DNA extraction kit according to manufacturer's protocol, followed by quantification and purity assessment using Nanodrop One. PCR amplification was carried out using two primers for HVR1 and HVR2. The forward and reverse primers used for both the HVR1 and HVR2 were F-5'-TTA ACT CCA CCA TTA GCA CC-3' and R-5'-CCT GAA GTA GGA ACC AGA TG-3' and F-5' GGT CTA TCA CCC TAT TAA CCAC3' and R-5' CTG TTA AAA GTG CAT ACC GCCA3' respectively, followed by gel electrophoresis to ensure DNA quality and integrity. BIg dye terminator Sanger Type Sequencing was utilised for sequence analysis. Base calling and multiple alignments (ClustalW) were performed using BioEdit. AMOVA analysis was conducted using Arlequin, population admixture was assessed using STRUCTURE, phylogenetic tree construction utilised MEGA11, pairwise genetic comparisons were performed with SDT, and haplotype networks were created using PopArt and tcs.

Of the total 100 individuals, 93 and 94 successful sequences were generated based on HVR and HVR2 primers respectively. The findings revealed significant within-population genetic variation based on AMOVA analysis in both HVR1 and HVR2, the within population variance was estimated at 99.69% and 96.69% in HVR1 and HVR2 respectively. The population admixture analysis revealed varying degrees of population subgrouping depending on the number of simulation runs. The phylogenetic tree and haplotype network analyses demonstrated complex genetic diversity among the Hausa

population. The results underscore the importance of considering population-specific genetic markers for forensic investigations and population genetics studies in Nigeria.

In addition, we downloaded the mitogenome sequences from the NCBI database of different global populations in order to make comparison with our genetic data. We downloaded from Southern Africa (Angola), Oceania (New Zealand) Tokelau population, South Asia (west Indian Caste), Central Europe (Switzerland), South America (Paraguay) and North America (Canada, Newfoundland) with accession numbers MF3812871-MF3813061, MT9282831-MT9282971, MK0439671-MK0439862, MT0790191-MT0790371, MH9818231-MH9818421 and MF5887941-MF5888111 respectively. We used Arlequin to conduct AMOVA and use the FST values generated to make a pairwise identity matrix between the populations using R statistical tool, SDT was used to make pairwise comparison of individual sequences, and STRUCTURE to make ancestral admixture within and between the population across the continents. We used 15-20 sequences from each of the continental populations. Markov Chain steps were set at 10,000, and a burn-in length of 10,000 was used with K replication values set at 6, and 3 iterations (Evanno et al., 2005). To determine which of the eighteen runs was best suited for inferring ancestral relationships among the study populations, a structure harvester was used (Earl & vonHoldt, 2012)

The AMOVA result based on selected global populations across the continent indicated that the populations can be distinguished from one another with relative certainty based on their genetic sequence. More than 90% of the genetic variation is accounted for among the population with only a less than 10% within population variation. This has indicated that there is genetic variation between populations from Nigeria, Switzerland, New Zealand, West India, Angola, Paraguay and Canada. The overall FST value of above 0.9 indicated strong genetic diversity among the study population, thus inferring genetic variability. In addition, individual FST values that were used to generate a colour-coded matrix indicated genetic distinctiveness of the Nigerian population from the rest of the

world based on the mtDNA HVR2 data. However, it can be observed that statistical significance was only observed between Nigerian and Canadian population.

Sequence Demarcation Tool indicated a pairwise individual genetic distance between the global population with the Nigerian population maintaining a high genetic diversity and distinctiveness from the rest of the population. However, the Angolan population that we assumed to show relationship with the Nigerian population had demonstrated a more genetic relatedness with the rest of the population. This can be attributed to mixed genetic traits between the Angolan and Portuguese populations that are still present in the country even after independence from the Portuguese colonisation. Based on the SDT figure, the South and North American population are furthest from the West African population, it is interesting that the Central European populations are placed closer to the West African populations.

In our effort to determine shared ancestry between the West African and other continental populations, we observed a distinct ancestry of the Nigerian Hausa population from the rest of the world. The targeted Hausa population in our study were selected based on their historical importance in the Hausa kingdom. The targeted Hausa population were from ancient Hausa cities which could be the reason for such genetic uniqueness. The population has further elucidated the much reported high genetic diversity of the African population. This has further indicated the relevance of this population for forensic genetics.

In conclusion, this study provides insights into the genetic diversity and structure of the Hausa population in Nigeria, as inferred from the analysis of mtDNA control regions. The identified patterns contribute to our understanding of the population's evolutionary history and can aid in forensic investigations and precision medicine approaches. Further research and expanded sampling are recommended to enhance the accuracy and robustness of population genetic studies in the region.

Dedication

This PhD work is dedicated to Allah Subhaanahu Wa Ta'aala, who created the universe and creations therein. Whose knowledge is vast that if the oceans in the universe were to be inked to write His words, the ocean water would be exhausted before his words, even if we were to bring more and more oceans.

Acknowledgement

All gratitude and grace go to Allah (SWT) whose infinite wisdom provided me with knowledge, health and strength to carry out this PhD journey. In the pursuit of knowledge and the completion of this thesis, I have been fortunate to receive support and guidance from numerous individuals who have played significant roles in my academic journey. I extend my heartfelt gratitude to each and every one of them, as their contributions have been invaluable.

First and foremost, I would like to express my deepest appreciation to my supervisor, Dr. Gaurav Kumar. Your unwavering commitment, patience, and mentorship have been instrumental in shaping this research. Your expertise and guidance have pushed me to new heights, and I am truly grateful for your continuous support. I am equally indebted to Prof. Arvind Kumar Jain, for your valuable insights and scholarly guidance throughout this endeavor. Your expertise and meticulous attention to detail have enhanced the quality of this work. Your encouragement and motivation have been a constant source of inspiration. To all the faculties in the Division of forensic science, Galgotias university, your influence and support have been invaluable throughout this research journey. I am deeply grateful for your contributions, encouragement, and belief in my abilities.

I would like to extend my sincere gratitude to Umaru Musa Yar'adua University, Katsina, for their sponsorship and financial support, which enabled me to conduct this research. Your belief in my abilities and your commitment to promoting academic excellence are deeply appreciated.

To my friends and colleagues from the workplace that we came to India together for the same academic pursuit, Dr. Usman Lawal Usman and Usman Affan, thank you for your unwavering support, stimulating discussions, and shared camaraderie. Your presence has made this journey both meaningful and enjoyable. I want to send my heartfelt gratitude to Mr Umesh Sharma for helping me in the arrangement of some Research kits.

I would also like to express my heartfelt appreciation to Dr. Kabiru Hamman Joda, Dr. Umar Abdu Sulaiman, and Dr. Neksumi Musa, my friends and neighbours. Your intellectual engagement, encouragement, and valuable insights have enriched my understanding and perspective. It is indeed a memorable journey that I will relive for the rest of my life.

My heartfelt thanks go to Mustapha Adam Habib for providing me with accommodation upon my arrival in India. Your generosity and kindness have made a significant difference in my experience, and I am grateful for your support.

I am deeply grateful to Professor Eugenia Maria D'Amato from the University of Western Cape, South Africa, for her invaluable guidance on literature search and analysis. Your expertise and suggestions have broadened the scope and depth of this research. Special recognition goes to Babalola Favour Olanrewaju, an IT student from the Federal University of Agriculture Abeokuta, Nigeria, who provided invaluable assistance during my lab work at Inqaba Biotec West Africa. Your dedication and expertise have contributed significantly to the success of this research.

To my family, my pillar of strength, I owe a debt of gratitude beyond words. To my mother, Hajia Ummukulthum (Turai) Dauda, and my father, Alhaji Shehu Yunusa, thank you for your unconditional love, unwavering support, and constant encouragement. To my beloved wife, Aisha Salisu Abubakar, and my precious daughters, Ramlat Ibrahim El-ladan and Aisha Ibrahim El-ladan (Afnan), your patience, understanding, and belief in me have been a constant source of motivation and inspiration. I would also like to extend my appreciation to my brothers and other family members, Bashir Shehu El-ladan, Yahya Shehu El-ladan, and Nura Sa’idu Abe, Abdulhamid Balarabe, Abdallah Aminu Galadima, Abdurrahman Umar, Saifullahi Yusha’u Elladan, Muhammad Shehu, Abdullahi Aliyu Maiwada, and my sisters, Hajia Hauwa'u Sa’idu Abe, Saratu Shehu, Safinatu Shehu and Maryam Shehu, for their unwavering support and encouragement. I would like to acknowledge the contributions of my uncles, Barrister Ibrahim Dauda El-ladan, Alhaji

Balarabe Dauda El-ladan, Alhaji Armaya'u Dauda El-ladan, Alhaji Ibrahim Yahya El-ladan, Alhaji Sulaiman Dauda El-ladan and Professor Siraj Abdulkarim may Allah reward all of you and unite us in Jannah (heaven). I owe a debt of gratitude to my Aunties, Hajia Halima Dauda El-ladan, Hajia Mardhiyyah Abbas, Aunty Zainab Sahabi, Hajia Murja Abubakar Maiwada, they have all been a pillar and support in my academic pursuit.

Dr Abubakar Sadeeq Adamu, Muhammaad Mubasshir Muhammad and Mahmood Usman, your support, guide and prayers are appreciated. Other colleagues from work and fellow Ph.D. students, whose names were not mentioned. Your support, encouragement, and intellectual discussions have played a significant role in shaping my ideas and perspectives.

Finally, I express my heartfelt appreciation to all the individuals, mentors, colleagues, and friends who have supported me in ways that cannot be fully captured in words. Your encouragement, constructive criticism, and belief in my capabilities have sustained me during the challenging moments of this research.

I am also grateful to the wider academic community and the anonymous reviewers whose valuable feedback and constructive comments have significantly strengthened the quality and rigor of this thesis. Your expertise and thoughtful insights have undoubtedly shaped the final outcome.

I extend my gratitude to the institutions, organizations, and funding agencies that have provided resources and opportunities for my research. Your support has enabled me to access essential tools, technologies, and research facilities, thereby enhancing the depth and reliability of my findings.

Lastly, I would like to acknowledge the countless individuals whose names may not be mentioned but who have contributed to my personal and academic growth. Whether

through stimulating discussions, intellectual exchanges, or acts of kindness, your impact on my journey has been profound.

This thesis is a culmination of the collective efforts, guidance, and support of all these individuals. I am humbled and honored to have had the privilege of working with such remarkable people. Their influence will forever resonate in my academic and personal endeavors.

Thank you all for being an integral part of this significant milestone in my life.

Date: 22nd May, 2023 **IBRAHIM EL-LADAN SHEHU**

Approval Sheet

This thesis titled entitled **"Matrilineal Genetic Diversity and Forensic Characterisation of Hausa Population in Nigeria"** was written by Mr. Ibrahim El-ladan Shehu for the approval of the award of the degree of Doctor of Philosophy in Forensic Science.

Examiner

……..

……..

……..

Supervisor's Signature

……

……

Date: …….

Place: ……

Table of Contents

Contents

Statement of Thesis Preparation

1. Thesis title: **"Matrilineal Genetic Diversity and Forensic Characterisation of Hausa Population in Nigeria"**
2. Degree for which the thesis is submitted: **Doctor of Philosophy**
3. Thesis Guide was referred to for preparing the thesis.
4. Specifications regarding the thesis format have been closely followed.
5. The contents of the thesis have been organised based on the guidelines.
6. The thesis has been prepared without resorting to plagiarism.
7. All sources used have been duly cited appropriately.
8. The thesis has not been submitted elsewhere for a degree.

Name:

Roll. No.:

List of Publications

1. Matrilineal Genetic Diversity and Forensic Data of Hausa Ethnic Population of Daura Emirate, Nigeria. Journal of Indian Academy of Forensic Medicine: Scopus-indexed…Communicated
2. Population Structure and Stratification among the Hausa Ethnic Group based on HVR1 mtDNA Analysis. Journal of Genetics: Springer; Scopus-indexed...Communicated
3. Mitochondrial DNA Analysis Reveals Complex Genetic Diversity and Population Structure among the Hausa People of Nigeria: Forensic Sciences Research: Taylor and Francis; Scopus-indexed…Communicated
4. Ibrahim El-ladan Shehu and Priyanka Chhabra. African Y-STR haplotyping and Y chromosome profiling: A Review. Indian Journal of Forensic Medicine and Pathology. 2021;14(3 special issue):379-385. Scopus-indexed

International Conferences

5. Ibrahim El-ladan Shehu, Arvind Jain Kumar and Gaurav Kumar. PhD thesis presentation: Matrilineal Genetic Diversity and Forensic Characterisation of Hausa Population in Nigeria. 2nd International Conference on Neo Era of Forensic Science and Law Interface. 22nd – 23rd April, 2023. Forensis Agora 2023. Galgotias University.
6. Ibrahim El-ladan Shehu and Priyanka Chhabra. Review on African Y-STR haplotyping and Y Chromosome Profiling. International e-Conference on Forensic Science and Criminology: Bridging the Gap in Criminal Justice System Conference Series; Forensis Agora. 15th -16th May, 2021. Galgotias University.

List of Figures

List of Tables

List of Terms and Abbreviations

ABBREVIATION	TERM
ATP	Adenosine TriPhosphate
aDNA	ancient DNA
BLAST	Basic Local Alignment Search Tool
CRS	Cambridge Reference sequence
CODIS	Combined DNA Index System
dGTP	Deoxyguanosine triphosphate
D	Genetic Differentiation
DVI	disaster victim identification
DDBJ	DNA DataBank of Japan
POLG	DNA polymerase gamma
EMPOP	EDNAP mtDNA Database
EDNAP	European DNA Profiling Group
ENA	European Nucleotide Archive
ftp	File Transfer Protocol
FINDS	Forensic Information Database Service
HVR1 and 2	hyper-variable regions 1 and 2
HVS-I, HVS-II, and HVS-III	hypervariable sections/segment 1,2 and 3
AI-SNPs	informative single nucleotide polymorphism
ISFG	International Society for Forensic Genetics
Kb	Kilo base
MPS	Massively Parallel Sequencing
mtDNA	Mitochondrial DNA
Mitogenome	mitochondrial Genome
MRCA	Most Recent Common Ancestor
NIH	National Institutes of Health
NJ	neighbour-joining

NUMTs	nuclear-mitochondrial DNA segments
nDNA	Nuclear DNA
PCR	Polymerase Chain Reaction
rCRS	revised Cambridge Reference sequence
rRNA	ribosomal RNA
SWGDAM	Scientific Working Group on DNA Analysis Methods
STS	Sanger-Type Sequencing
SNVs	Single nucleotide variants
SNPs	single nucleotide polymorphisms
SE	Standard Error
Kya	thousand years ago
tRNA	transfer RNA
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WGS	Whole Genome Shotgun

CHAPTER ONE

1.0 INTRODUCTION AND LITERATURE REVIEW

This chapter is segmented into two major components:

I. Introduction: This component provides an overview of the importance of mitochondrial Deoxyribonucleic Acid [DNA (mtDNA)] in studying human evolution, population genetics, and forensic applications. It also provides background information on the Hausa ethnic population, their genetic history, and relationship with other Nigerian and West African Populations. This section also provides the research problems/gaps, aims, objectives, scope, and significance that will be addressed in the study.

II. Literature Review: This section provides review of the existing literature on mtDNA and its application in forensic science, population genetics, and disease. It presents studies that have investigated the genetic diversity, population structure and other forensic data of African and more specifically West African populations. This component also provides an overview of previous mtDNA studies on Hausa genetics and other Nigerian tribes.

1.1 BACKGROUND OF THE STUDY

Mitochondrial DNA (mtDNA) is a unique type of genetic material that is found exclusively within the mitochondria, the organelles responsible for producing energy within eukaryotic cells[1]. Unlike nuclear DNA (nDNA), which is inherited from both parents, mtDNA is only inherited from the mother [2]. The mtDNA molecule is a circular piece of DNA that is significantly smaller than nDNA. It contains genes that encode for proteins, transfer RNA (tRNA), and ribosomal RNA (rRNA) involved in the production of energy within the mitochondria through a process called oxidative phosphorylation [3]. The mtDNA is also involved in other critical functions, such as the regulation of the mitochondrial replication and maintenance of mitochondrial structure [3].

One of the unique features of mtDNA is that it is present in multiple copies within each mitochondrion, and each cell contains many mitochondria. This means that a single cell can contain hundreds or thousands of copies of mtDNA, which makes it easier to detect changes or mutations in the mtDNA sequence [4]. The mitochondrial genome consists of two major segments, the non-coding and coding segments/regions. The latter encodes proteins involved in oxidative phosphorylation, while the non-coding region includes the control segment/region, which constitutes the replication origin and the replication and transcription promoters [5].

Due to its high mutation rate and non-recombination nature, mtDNA has become an essential tool for studying human evolution, forensic science, and disease diagnosis [6], [7]. In human populations, mtDNA analysis is used to track the migration of ancient populations and study the genetic diversity among different groups of people [8]. In forensic science, mtDNA analysis is used in cases where traditional DNA testing is not possible, such as in cases where the DNA sample is degraded, mixed with other DNA, or from a small amount of biological material. The high copy number and stability of mtDNA make it a useful tool for forensic analysis. Furthermore, mtDNA plays a crucial role in the diagnosis of mitochondrial diseases, which are a result of mtDNA mutations or mitochondrial proteins that are encoded by nuclear genes. These diseases can affect different tissues and organs and are presented in varying degrees of symptoms, mild to severe. The mtDNA sequencing analysis is used to identify the mutations responsible for these diseases and can also help in predicting the risk of developing them.

The earliest studies of African mtDNA diversity focused on small, geographically localized populations [9]–[13]. These findings revealed high levels of genetic diversity within and among these populations, consistent with a long history of population differentiation and isolation. Subsequent studies using larger sample sizes and more extensive geographic coverage have confirmed these findings and provided a more detailed picture of African mtDNA diversity [14]–[19].

African mtDNA diversity is characterised by a high frequency of haplogroups L0, L1, L2, and L3. These haplogroups are believed to have originated in Africa and are linked to the modern humans' early migration out of Africa [20]–[23]. Haplogroup L0 is found predominantly in southern Africa[24], [25], while L1 is found primarily in Central and West Africa [21], [26], [27]. Haplogroup L2 is found throughout Africa[28], but it is most prevalent in West and Central Africa. Haplogroup L3 is the most widespread haplogroup in Africa, and it is found in all regions of the continent [20].

The distribution of these haplogroups within and among African populations reflects a complex history of migration and population admixture. For example, the high frequency of haplogroup L3 in eastern and southern Africa suggests that these regions were important centres of early human migration out of Africa. Similarly, the high frequency of haplogroups L1 and L2 in West and Central Africa suggests that these regions were important centres of population expansion and admixture [29].

In addition to the major haplogroups, African mtDNA diversity also includes a wide variety of rare and novel haplotypes [30]. These haplotypes are often restricted to specific regions or populations and provide important clues about the populations' history [26], [31]–[34]. For example, rare haplotypes found in southern African populations suggest a complex history of population migrations and admixture with non-African populations [35].

Nigeria, situated on the Gulf of Guinea in Western Africa, boasts a population that makes up one-sixth of Africa's total populace. As of 2013, the estimated population is over 174 million, as per the National Population Commission's official records [36]. Nigeria is home to more black people than any other country in the world, and it comprises over 250 different ethnic groups [37]. The three largest ethnic groups are the Hausas, Yorubas, and Igbos, respectively [38], [39]. The Fulani-speaking ethnic group, known as Fulanis, have a significant history of intermarriage with the Hausas. They are sometimes collectively referred to as Hausa-Fulani, making up approximately 30% of Nigeria's total population. Hausa is the predominant language of identity in this group [17].

In this study, the target population comprised of Hausa, Fulani and Hausa-Fulani. The Hausa people are a major ethnic group in West Africa, primarily located in Nigeria, Niger, and Ghana [40]. They are one of the largest ethnic groups in Africa, with a population estimated at over 80 million people. There is a growing interest in understanding the genetic diversity of the Hausa population. One area of particular interest has been the analysis of mitochondrial DNA (mtDNA) genetic data, which provides valuable insights into the maternal ancestry of the population.

Several studies have been conducted to investigate the mtDNA genetic diversity of the Hausa population. Several studies found that the Hausa population had high levels of genetic diversity, with a wide range of maternal lineages present. The most common mtDNA haplogroup in the Hausa population was L3, which is also common in other African populations. Other common mtDNA haplogroups in the Hausa population included L0 and L2 [40], [41].

We aim to provide a comprehensive overview of the genetic structure and diversity of this population by studying the non-coding regions of mtDNA. Our findings may have important implications for genetic association studies, forensic investigations, and population health research.

1.1.1 RESEARCH PROBLEMS/GAPS

I. Limited mtDNA data for the Hausa population: Despite the large population of Hausa people in Nigeria, there may be limited existing mtDNA data available for this specific population. This could be due to a lack of previous research, limited sample size, or inadequate sampling methods. Addressing this research gap by generating more mtDNA data for the Hausa population could provide valuable insights into the genetic diversity and ancestry of this group.

II. Limited knowledge of regional genetic variation: While previous studies have found high levels of mtDNA genetic diversity within the Hausa population, there is still limited knowledge about regional genetic variation. A research gap exists in the need for more detailed analyses of mtDNA data at the regional level to

determine if there are differences in maternal ancestry between different regions of the Hausa population.

III. Comparing Hausa mtDNA data to other West African populations: While mtDNA data for the Hausa population may be limited, there may be existing data for other populations in West Africa that could be used for comparison. Comparing mtDNA variation between Hausa individuals and other West African populations could provide insights into the genetic relationships between these groups, as well as shed light on any unique genetic features of the Hausa population.

IV. Investigating the relationship between mtDNA variation and cultural practices in the Hausa population: The rich and diverse cultural heritage of the Hausa population could be reflected in their mtDNA variation. Investigating the relationship between mtDNA haplogroups and cultural practices, such as language, religion, and social organization, could provide insights into the historical and cultural factors that have shaped the genetic diversity of the Hausa population.

V. Examining the forensic utility of mtDNA analysis in the Hausa population: mtDNA analysis is commonly used in forensic investigations to identify individuals and determine their maternal lineage. However, the forensic utility of mtDNA analysis may vary between populations due to differences in genetic diversity and structure. Investigating the forensic utility of mtDNA analysis in the Hausa population could provide insights into the accuracy and reliability of this technique in this specific population.

VI. Lack of forensic mtDNA databases: Despite the potential for mtDNA hyper-variable regions (HVR) 1 and 2 to be used in forensic applications, there is a lack of mtDNA databases specific to the Hausa population. This research gap highlights the need for the development of Hausa-specific mtDNA databases to aid in forensic investigations.

VII. Evaluating the impact of migration and gene flow on Hausa mtDNA variation: The Hausa population has a long history of migration and gene flow with other

populations in West Africa and beyond. Investigating the impact of these historical events on mtDNA variation in the Hausa population could provide insights into the genetic relationships between different populations and provide insight on the factors which conferred the genetic diversity of the Hausa population over time.

1.1.2 AIM OF THE STUDY

The general research aim is to investigate the genetic diversity, structure, and phylogenetic relationships of the Hausa ethnic population from Daura, Nigeria, using mtDNA hypervariable regions 1 and 2, with the goal of providing insights into the evolutionary history, population dynamics, and forensic DNA analysis of this population.

1.1.3 THE OBJECTIVES ARE TO:

I. Assess the genetic structure and diversity of the Hausa ethnic population from Daura, Nigeria using mtDNA hypervariable regions 1 and 2.
II. Determine the frequency and distribution of haplotypes and haplogroups in the Hausa ethnic population.
III. Investigate the phylogenetic relationships among haplotypes and haplogroups in the Hausa ethnic population.
IV. Explore the population stratification of the Hausa ethnic group based on mtDNA hypervariable regions 1 and 2.

1.1.4 SCOPE OF THE STUDY

The study will be centreed on the Hausa population in Nigeria, specifically analysing the mtDNA variation in HVR1 and HVR2. The study will aim to characterise the mtDNA variation, infer the phylogenetic relationships, assess the association between mtDNA haplogroups and demographic factors, evaluate the forensic utility of mtDNA analysis, and investigate the historical and cultural factors that have shaped the genetic diversity of the Hausa population.

The study involves collecting mtDNA samples from individuals in the historic Daura emirate, predominantly Hausa population, analysing mtDNA sequences using molecular biology techniques, and conducting statistical analyses to infer phylogenetic relationships and assess demographic associations. The forensic utility of mtDNA analysis will be evaluated by comparing the accuracy and reliability of maternal lineage identification using HVR1 and HVR2 data with other forensic markers used in Nigeria. Finally, the study also involves reviewing literature on the historical and cultural factors that have influenced the genetic diversity of the Hausa population and how these factors impact the use of mtDNA in forensic investigations.

1.1.5 SIGNIFICANCE OF THE STUDY

I. Contribution to the understanding of genetic diversity: The study will provide contribution to the understanding of the genetic diversity and ancestry of the Hausa population, revealing insights into their origins and historical relationships with other populations in West Africa and beyond.

II. Forensic significance: The study will evaluate the forensic utility of mtDNA analysis in the Hausa population, providing important information for criminal investigations and identifying missing persons.

III. Health implications: The study may have implications for public health, as certain mtDNA haplogroups have been associated with increased risk of certain diseases.

IV. Implications for population genetics research: The study may have implications for population genetics research, as it will provide new mtDNA data for the Hausa population that can be used for comparative analyses with other populations and for testing hypotheses about the genetic relationships among populations.

In a nutshell, the study has the potential to make significant contributions to the fields of population genetics, forensic science, public health, and cultural anthropology, while also providing important information for the Hausa community in Nigeria.

1.2 LITERATURE REVIEW

1.2.1 THE WHOLE GENOME OF THE MITOCHONDRIA (MITOGENOME)

MtDNA is an important genetic element that is located exterior to the nucleus of a cell in the mitochondria, which are organelles responsible for producing metabolic energy. mtDNA is double-stranded, circular, 16.6 kilobase pairs DNA molecule, it is without histones and is inherited uniparentally, making it useful for evolutionary and population history studies as well as forensic sciences [42]. The nucleotides in mtDNA can be regarded into Heavy (H) and Light (L) strands, the H strand has the higher guanine content which conferred it with greater relative molecular mass in opposition to L strand which is cytosine-rich [43]. The mtDNA replication starts with the H strand in the control region, which does not code for any gene products [34].

In 1840, early scientists described the presence of small, specialized structures within cells that appeared to play a vital role in cellular processes. These structures were referred to as "bioblasts" and were thought to be responsible for important cellular functions. However, in subsequent research, these structures were given the name "mitochondria." [44]. They were later dubbed the "powerhouse of the cell" due to their ability to generate majority of the chemical energy needed to drive the biochemical reactions in the cells. Mitochondria have a variable morphology that can change drastically within the same cell and vary in number and location depending on the cell type [45].

The mitogenome to be first sequenced completely was the human’s in 1981, up to 10 copies of mtDNA can be found in a single mitochondrion [46]. The pioneering work of Anderson et al., (1981) gave rise to the CRS which is termed Cambridge Reference Sequence that the mitochondrial genome is typically compared to when conducting study. The CRS has since been updated and is now known as the revised Cambridge Reference sequence (rCRS). When new mitochondrial DNA sequences are analysed, any differences found in the sequence are assigned as mutations to the rCRS. These differences can be in the form of single nucleotide polymorphisms (SNPs) or

insertions/deletions (INDELS) [47]. Historical numbering conventions were retained to facilitate ease of reference.

Mitochondrial DNA (mtDNA) consists of 37 genes, with 28 and 9 on the H and L strands respectively. Mitochondria play a crucial role in energy production, and 13 of the genes are related to the respiratory chain complexes, which produce enzymes that contribute to oxidative phosphorylation, a mechanism for producing Adenosine TriPhosphate (ATP) [48]. Additionally, two genes encode for rRNA, and 22 encode for tRNA [48].

The mtDNA control region, located between positions 576 and 16024 [49], contains important factors such as transcription and regulation factors and the origin of replication for the H-strand. This region is 1.1 kb long and includes three hyper-variable segments or sections (HVS I,II, and III or HVR1, 2, and 3), which are vital in understanding human genetics. HVR1 and 2 span from positions16,024-16,482 and 21-413 respectively and HVR3 from 438 to 574 [50]. HVR1 and HVR2 regions are more variable compared to HVR3, and this explains why they are the most commonly analysed regions in forensic analysis, but with the advent of complete mtDNA genome sequencing, more precise information has been obtained [50].

1.2.2 MITOCHONDRIAL MUTATION

The mitogenome exhibits greater mutation rate than the nGenome, it is estimated at around 100 to 1000 times higher. However, published rates of mitochondrial DNA mutations vary greatly, likely due to differences in the size of studies and methods used to measure the rate [51]. For example, a study in Iceland reported 0.0043/generation mutation rate, while other publications reported an estimated a rate of 0.030/generation in poly C-tract and 0.0106/generation involving meta-analysis of eleven pedigree-related studies [50]. Although phylogenetic-based mutation rates generally manifest as considerably lower, the calculations derived from ancient DNA studies focusing on "relatively young" human samples indicate a reduction of approximately two-fold in comparison to rates determined through pedigree observations [52], [53]. The range of published phylogenetic rates is believed to be influenced by diverse estimates that take

into account recent human demographic histories, including instances of serial bottlenecks and expansions [19].

While it is often suggested that the lack of mitochondrial DNA repair mechanisms is the reason for the relatively high mutation rate, the issue is more complex and not well understood [54]. Multiple repair pathways have been uncovered through research, with nuclear genes encoding mechanisms such as single-strand break repair and mismatch repair [55]. Additionally, according to a fascinating study, the significant presence of mitochondrial Deoxyguanosine triphosphate (dGTP) in various tissues may potentially impede the proofreading function of the exonuclease domain within the DNA polymerase gamma (POLG) protein. This function is essential for eliminating DNA base-pair mismatches during the replication process [50].

1.2.3 MITOCHONDRIAL HETEROPLASMY

In most cases, all the mitochondrial DNA in a cell has the same sequence, known as homoplasmy. However, mutations in the mitochondrial DNA can cause mitochondrial dysfunction, which can lead to severe inherited diseases [56]. Therefore, maintaining the homoplasmic state is crucial, although the underlying mechanism is not well understood. It is believed that a genetic bottleneck occurs during oogenesis, resulting in a single molecule that is usually free of mutations [57]. However, since mitochondrion can possess multiple mitogenomes, mutations affecting only a portion of the molecules in the cell can occur, which is known as heteroplasmy. Heteroplasmy, the presence of multiple types of mtDNA molecules in an individual's cytoplasm, is an important factor to consider when using mtDNA for forensic purposes [5], [48]. It was previously thought that an individual's mtDNA was homoplasmic, but we now know that heteroplasmy exists in up to 5-10% of the population, albeit at low levels. In essence, heteroplasmy refers to a mix of two or more mtDNA populations within an individual [58].

The level of heteroplasmy differs between various tissues and increases with an individual's age. The highest level of heteroplasmy is typically found in hair [58]. The two types of heteroplasmy that have been identified in regions HVRI and HVRII of

mtDNA are length and point heteroplasmies. Length heteroplasmy, which is characterised by variations in the number of cytosine bases in the homopolymer sequence, is more common in the population than point heteroplasmy. The latter is less frequent in the population than the former [59], [60]. Heteroplasmy can either be inherited or arise from somatic mutations that occur during an individual's lifetime. When heteroplasmy is present in all tissues at a high level, it is likely to have been inherited from the mother. On the other hand, when heteroplasmy occurs only in some tissues, it is the result of somatic mutations [58]. In Sanger-Type Sequencing (STS) involving mtDNA analysis, minor heteroplasmic variants exceeding 20% frequencies are the recommended to be identified [61], [62].Heteroplasmy above background noise can be accurately identified while disregarding heteroplasmy below this threshold. The sensitivity of the mtDNA analysis has improved with adoption of Massively Parallel Sequencing (MPS), allowing heteroplasmy to be identified at levels lower than was previously feasible [63], [64].

1.2.4 MITOCHONDRIAL INHERITANCE

Mitochondrial DNA is inherited maternally. The mechanism behind this maternal inheritance was originally thought to be due to a dilution effect, with the number of mtDNA from the ovum higher than the mtDNA from the sperm by around 100-fold [65]. A study in 2018 reported three families with unusually high heteroplasmy, suggesting the possibility of biparental mtDNA inheritance. This hypothesis challenges the traditional view of maternal inheritance and has received both supporting and critical responses [66]. However, n a recent study encompassing 11,035 trios of mothers, fathers, and children, it was discovered that the observed patterns could be attributed to the existence of nuclear-mitochondrial DNA segments (NUMTs) that imitate heteroplasmy, rather than paternal mtDNA inheritance [65].

In every family included in the study, the fathers displayed at least one unique NUMT that was not present in the mothers. Upon examining a potential contamination incident, researchers detected a megaNUMT in eight members of the family who shared maternal relations. The expected mtDNA haplogroup was observed consistently across the family

tree. Notably, the megaNUMT was not found in hair shaft samples, where nuclear DNA is absent, offering a means to investigate future assertions of paternal mitochondrial inheritance [67].

1.2.5 MtDNA PHYLOGENY

The study of the mitochondrial DNA (mtDNA) phylogeny provides insight into the migration patterns and demographic history of human evolution. Ancestry informative single nucleotide polymorphism also known as AI-SNPs are used in the mtDNA phylogeny to construct a branching structure where each branch represents a specific mutation [67]. These AI-SNPs are responsible for the DNA sequence variation and can undergo a transition or trans-version within different populations. By obtaining AI-SNPs information from mitogenome's coding region, the haplogroups can be assigned based on the samples' genotypes, which helps in inferring the mtDNA phylogenetic tree [67], [68].

It is essential to have an up-to-date mtDNA phylogenetic tree for accurate haplogroup assignment, estimating site-specific mutation rates, and quality control. Phylotree is a tool that constructs a world-wide phylogeny by defining each haplogroup using a set of mutations in each branch based on sequence variation along the entire mitogenome length [34].

In conclusion, the study of the mtDNA phylogeny and the use of tools like Phylotree are crucial for understanding human evolution. These tools help researchers obtain important information regarding haplogroup assignment, site-specific mutation rates, and quality

control.

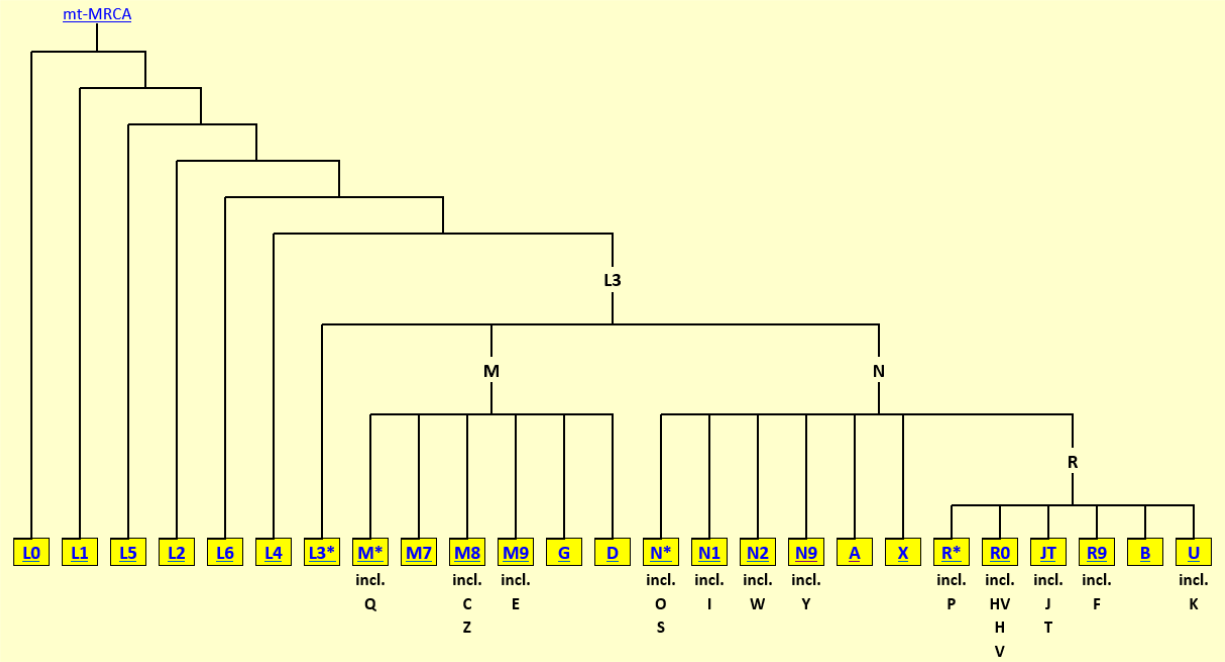


Figure 1.1 mtDNA phylotree build 16 organised into eleven sub-trees generated by comparison with rCRS

The mtDNA phylogenetic tree build 16 is depicted in Figure 1.1, which is organised into 11 subtrees that can be accessed through the website http://www.phylotree.org/tree/main.htm. The tree is rooted using the mitochondrial Most Recent Common Ancestor (mt-MRCA), and it shows how accumulated polymorphisms have led to the formation of different haplogroups among individuals over time. The information in Figure 1.1 was updated on February 18, 2016 [69].

1.2.6 MtDNA DATABASES

Mitochondrial DNA (mtDNA) databases are collections of mtDNA sequences that are used for various purposes, including forensic identification, medical research, and population genetics studies. These databases are designed to store and organize large amounts of genetic information, making it easier to compare and analyse mtDNA sequences from different individuals [70].

Forensic mtDNA databases are one of the most widely used types of mtDNA databases. They are used by forensic scientists to identify individuals or determine the relatedness of individuals in criminal investigations or missing person cases. Forensic mtDNA databases typically contain mtDNA sequences from individuals who have been convicted of a crime, as well as sequences from family members of missing persons or unidentified human remains [71].

One example of a forensic mtDNA database is the United States' Combined DNA Index System (CODIS), which contains DNA profiles from individuals convicted of certain crimes, as well as from evidence collected at crime scenes [72]. Another example is the National DNA Database in the United Kingdom; Forensic Information Database Service(FINDS), which contains DNA profiles from individuals arrested for certain offenses and from crime scenes[73].

Medical mtDNA databases, on the other hand, are used for research purposes in medicine and biology. They contain mtDNA sequences from individuals with specific medical conditions, such as mitochondrial diseases, as well as from healthy individuals. These databases are used to study the genetic basis of these diseases and to identify potential treatments [74].

Population genetics mtDNA databases are used to study the genetic diversity of populations around the world. These databases contain mtDNA sequences from individuals from different populations and regions, allowing researchers to compare the genetic makeup of different groups [75]. Examples of population genetics mtDNA databases include the Human Mitochondrial Genome Database (mtDB) and the GenBank database.

One of the challenges with mtDNA databases is the quality and completeness of the data. The accuracy of mtDNA sequencing can be affected by a number of factors, such as sample quality, sample quantity, and sequencing errors. Therefore, it is important to ensure that the data stored in mtDNA databases are of high quality and completeness [16], [58]. In addition, there are ethical and privacy concerns associated with mtDNA

databases, particularly in the context of forensic databases. Critics argue that the use of forensic databases can lead to racial profiling and discrimination, and that individuals may be unfairly targeted based on their genetic information [76].

In conclusion, mtDNA databases are an important tool for forensic identification, medical research, and population genetics studies. While there are challenges associated with these databases, they provide valuable information that can be used to better understand human genetic diversity and disease. It is important to continue to ensure the quality and completeness of data in these databases, as well as to address the ethical and privacy concerns associated with their use. There are numerous mtDNA databases, some few examples include:

1.2.6.1 EDNAP MtDNA Database (EMPOP)

Regarded as the foremost comprehensive repository of mitochondrial DNA worldwide, EMPOP proves invaluable in forensic investigations. Not only does it exhibit an extensive representation of populations across the globe, but it also stands out for its meticulous focus on data quality. The European DNA Profiling Group (EDNAP) spearheaded this project, which offers a publicly accessible website (www.empop.online). To ensure the accuracy of incorporated data, all contributions undergo rigorous quality assessment filters, including EMPcheck and Quasi-Median Network tools, effectively identifying sequencing errors [77].

Employing the SAM2 string-based search algorithm, EMPOP possesses the capability to convert sequences into alignment-free strings, thereby facilitating accurate definition of haplotypes, regardless of alignment differences. This sophisticated tool not only identifies the haplogroup but also resolves nomenclature ambiguities in the presence of phylogenetically unstable positions, while simultaneously accounting for block INDELS [50], [78].

Quality assessment of mitochondrial DNA databases is critical, as errors ranging from approximately 10% to over 50% have been reported in medical publications [50].

EMPOP aims to address these issues by providing a high-quality database that can be trusted for forensic and research purposes [53].

1.2.6.2 Human Mitochondrial DNA (mtDB; http://www.mtdb.igp.uu.se/)

human mitochondrial DNA (mtDB) is a publicly available database that contains a large number of human mitochondrial DNA (mtDNA) sequences from individuals of different ethnicities and geographic locations around the world [79]. The database is maintained by the Department of Genetics and pathology Uppsala university Sweden section of Medical Sciences and Molecular Anthropology.

The mtDB database includes both complete mtDNA sequences and partial sequences, such as those obtained from the hypervariable regions of the mtDNA. The sequences in the database are obtained from a variety of sources, including research studies, forensic investigations, and medical testing [79].

One of the strengths of mtDB is its focus on haplogroup classification, which provides important information about the evolutionary history and population genetics of individuals. The mtDB database sequences are provided with haplogroup information, giving researchers ability to explore the diversity of mtDNA haplogroups in different populations [79]. In addition to its use in population genetics and evolutionary studies, mtDB is also a valuable resource for forensic investigations. The large number of mtDNA sequences in mtDB provides a useful reference database for forensic investigations, allowing investigators to compare unknown mtDNA sequences to those in the database to determine possible matches [79].

Overall, mtDB is a valuable resource for researchers and forensic investigators interested in the study of human mtDNA variation and haplogroup diversity. Its large collection of mtDNA sequences and haplogroup classifications make it an essential tool for exploring the evolutionary history and population population genetics, as well as for identifying and comparing mtDNA sequences in forensic investigations.

1.2.6.3 Mitomap (https://www.mitomap.org/MITOMAP)

Mitomap is a mitochondrial DNA database that provides comprehensive information on variations and mutations associated with human diseases. As of January 15, 2023, mitomap database have incorporated 2,613 additional full-length (FL) GenBank sequences and 1,043 new control region (CR) sequences into their database. This brings mitomap database current total of FL sequences to 59,389 and CR sequences to 78,884. The total number of Single nucleotide variants (SNVs) in mitomap database is now 19,571. Mitomap database routinely update their GenBank sequences every 4-6 months, while hand curation of variants and references is an ongoing weekly process. They take great care to ensure the accuracy and integrity of their database for the benefit of the users.

Mitomap is a valuable resource for researchers studying mitochondrial diseases, as it allows them to identify potential genetic causes for these diseases. It also provides a tool for the classification of mtDNA haplogroups, which can be useful for population genetics studies. Mitomap's variant analysis and annotation tools enable researchers to identify and interpret genetic variants in mtDNA sequences, which can lead to a better understanding of the molecular mechanisms underlying disease.

1.2.6.4 GenBank

GenBank ® is a comprehensive genetic sequence database managed by the National Institutes of Health (NIH), consisting of annotated DNA sequences that are publicly available. As a member of the International Nucleotide Sequence Database Collaboration, EMPOP actively participates in daily data exchange with the DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA).

GenBank releases are made bimonthly and can be accessed via the File Transfer Protocol (FTP) site. Detailed information on the latest release, including notifications of upcoming changes, can be found in the release notes. Additionally, release notes for previous GenBank versions can be accessed. Each release includes growth statistics for both the conventional GenBank divisions and the Whole Genome Shotgun (WGS) division.

GenBank aims to maintain a reliable and up-to-date database for the benefit of the scientific community.

GenBank's large size and diverse range of data make it a powerful tool for genetic research. Researchers can use GenBank to access and compare genetic sequences from different populations, which can help them understand the genetic basis of disease and human evolution. The database is freely available to the public, making it accessible to researchers and clinicians around the world (https://www.ncbi.nlm.nih.gov/genbank/)

1.2.6.5 HaploGrep (https://haplogrep.i-med.ac.at/)

HaploGrep is a web tool that is used for the classifying and analysing mtDNA haplogroups. It uses the Phylotree database as a reference to classify mtDNA haplogroups based on SNPs (single nucleotide polymorphisms) and other variations in mtDNA sequences. HaploGrep is a powerful tool for researchers studying population genetics, human evolution, and the molecular basis of disease [80], [81].

HaploGrep can be used to classify mtDNA sequences into haplogroups based on their genetic characteristics. It provides an automated classification process, which can save researchers time and resources. The tool also allows researchers to compare and visualize haplogroup data, which can help them identify patterns and relationships among different populations. HaploGrep can be used to analyse large datasets of mtDNA sequences, making it useful for large-scale population genetics studies.

1.2.6.6 Phylotree

Phylotree is a mitochondrial DNA haplogroup classification system that is used by many researchers studying human evolution and population genetics [81]. It provides a detailed phylogenetic tree that classifies mtDNA haplogroups based on SNPs and other genetic markers. Phylotree is regularly updated with new information and is widely used as a reference for mitochondrial DNA haplogroup classification [69].

Phylotree allows researchers to classify mtDNA sequences into specific haplogroups based on their genetic characteristics. It provides a hierarchical structure that can help researchers understand the relationships between different haplogroups and populations.

Phylotree is widely used in studies of human evolution and population genetics, as well as in forensic investigations and medical research [69].

Phylotree's detailed classification system and hierarchical structure make it a valuable tool for researchers studying mitochondrial DNA haplogroups. It provides a standardized and widely accepted classification system, which can help to ensure consistency and comparability across different studies. Phylotree is also regularly updated, which ensures that researchers have access to the most up-to-date information and classification criteria.

1.2.6.7 HmtDB (https://www.hmtdb.uniba.it/)

HmtDB (Human Mitochondrial DNA DataBase) is a comprehensive database of human mtDNA sequences, which is curated and maintained by the Indian Institute of Science. The database contains information on over 25,000 mtDNA sequences from individuals around the world, along with their haplogroup classifications [82].

1.2.6.8 MitoMiner

MitoMiner is a database of mitochondrial protein sequences and functional annotations. The database contains information on over 12,000 mitochondrial proteins from a wide range of species, and can be used for bioinformatics and systems biology research [83].

1.2.6.9 MitImpact

MitImpact is a database of mtDNA variants and their potential impact on protein function. The database contains information on over 2500 mtDNA variants that are predicted to have a functional impact, along with their clinical and phenotypic associations.

1.2.7 FORENSIC APPLICATIONS OF MTDNA

Forensic geneticists first suggested the use of mtDNA analysis in the field in the late 1980s [84]. Since then, the technique has undergone significant evolution. Initially, mtDNA analysis involved the examination of the hypervariable regions. With the passage of time, the research scope has broadened significantly, now encompassing not just the entirety of the control region but also single nucleotide polymorphisms (SNPs) in the

coding region, and most recently, the sequencing of the whole mitochondrial genome [63]. Mitochondrial DNA sequencing is a valuable tool for characterising biological evidence. While mitochondrial DNA (mtDNA) has limited potential for individual identification due to the lack of recombination, it offers significant benefits in cases that require confirmation of maternal lineage or involve limited nuclear DNA, such as the examination of bones, teeth, and hair. This advantage stems from the high copy number of mtDNA. As a result, mtDNA is extensively employed in the analysis of ancient DNA and plays a crucial role in the triage of disaster victim identification (DVI). In forensic mtDNA typing, apparent differences in nomenclature can be misleading and may result in erroneous exclusions. Consequently, the establishment of a well-defined reporting process is of utmost importance. While both point heteroplasmy and length heteroplasmy are relevant to mtDNA typing, reporting the latter is unnecessary in forensic typing since it does not influence the haplogroup assignment. Furthermore, it should be noted that samples displaying variations in point heteroplasmy may not necessarily yield definitive exclusionary evidence. Sequences differing by a single base should be subjected to further evaluation to determine their rate of mutation. Collaboratively, the International Society for Forensic Genetics (ISFG) and the Scientific Working Group on DNA Analysis Methods (SWGDAM) have formulated guidelines that are specifically tailored for forensic mtDNA typing. Adherence to these guidelines promotes uniformity across the forensic community and simplifies the process of reporting and comparing forensic results.

1.2.7.1 SWGDAM Guidance

These guidelines emphasise the importance of controls to avoid contamination, as well as the use of validated thresholds to account for low-level contamination that may be difficult to avoid during mitochondrial analysis. In line with the guidelines, a hybrid approach combining rule-based and phylogenetic methodologies for nomenclature is advocated, with a strong emphasis on leveraging the trustworthy EMPOP database while being cautious of historic data that may have been interpreted differently. The guidelines provide rules for homopolymeric C-stretches, substitution and INDELS, and transition

and transversions in MPS workflows. The reporting guidance provides a framework of three interpretive categories. The first category, labeled as 'exclusion,' is applied when two or more nucleotide differences (excluding length heteroplasmy) are identified. The second category, referred to as 'inconclusive,' is assigned when there is a single nucleotide difference. The third category, 'cannot exclude,' is used to indicate the inability to rule out a specific scenario. Simple counting is employed for frequency reporting, utilising a recognized population database. The SWGDAM is actively planning to revise these guidelines to incorporate reporting strategies for the complete mitochondrial genome.

1.2.7.2 ISFG Guidance

The ISFG provides updated recommendations to supplement the guidance provided in 2000 for forensic mtDNA typing. The recommendations place a strong emphasis on error prevention and the resolution of potential instances of minor contamination, with a focus on independent verification and active engagement in proficiency testing. Additionally, the ISFG recommends reporting in relation to the rCRS reference sequence and conducting typing of the entire mtDNA control region, offering support and tools to assist in making informed decisions regarding nomenclature. Laboratories are encouraged to create their own guidelines for the interpretation and reporting of length and point heteroplasmy, with a choice on whether to report length heteroplasmy. Quality tools are recommended to check the phylogeny for expected results, given the high incidence of mtDNA sequence interpretation errors. In EMPOP, the default practice is to avoid including homopolymeric C-tracts, and they should be disregarded when searching other databases. To mitigate reporting bias, it is advised to conduct alignment-based searches encompassing the entire database. Additionally, the database used to assess the significance of a match should align with the specific circumstances of the case, and different frequency estimates, including the conservative confidence interval situated in the upper 95%, can be considered, taking into account local population variation in mtDNA frequencies. The ISFG plans to review these recommendations to address whole mitochondrial genome reporting.

1.2.8 Degraded and Ancient DNA

The methodologies devised for analysing ancient DNA can be effectively applied to sequence extensively deteriorated substances encountered at crime scenes. To eliminate external contaminants, a meticulous and comprehensive cleaning process of bones is imperative, and the analysis should be performed in laboratories specifically designed for handling DNA at extremely low levels [85]. As the duration progresses, DNA experiences degradation, resulting in the generation of inaccurate sequences. One of the key forms of damage is hydrolytic damage, leading to deamination. Deamination involves the conversion of cytosine to uracil, and during PCR, uracil pairs with adenine, ultimately producing thymine. The majority of sequencing errors arise from miscoding, but the application of N-glycosylase treatment to remove uracils can assist in alleviating this problem [86], [87].

Degraded DNA introduces another complication with the emergence of unexpected long chimeric sequences due to PCR jumping. This phenomenon occurs when an additional adenosine molecule is inserted at the end of a DNA template, causing it to "jump" to another template during polymerization, resulting in the creation of an in vitro recombination product. This issue becomes particularly troublesome if amplification is initiated from a small number of copies, and any unexpectedly successful products should raise doubts or scepticism [88]. The analysis of degraded DNA becomes more intricate due to the contamination of modern DNA and bacterial DNA. In situations where sample quality is compromised, a capture-hybridization technique [89] can be employed to enhance the abundance of mtDNA. Nevertheless, this methodology has the side effect of amplifying NUMTs, which can be mistakenly interpreted as heteroplasmic sequences. Through the elimination of known NUMTs and the application of reliable filters to manage other occurrences, the frequency of sequencing errors can be significantly decreased [67].

1.2.8 INTERESTING HISTORICAL CASES OF MtDNA USAGE

The study of mitochondrial DNA has made noteworthy advancements in our understanding of population migrations and the domestication of animals. Moreover, it has served as a pivotal tool in the examination of mass graves connected to historical conflicts. Moreover, mitochondrial DNA analysis can occasionally provide answers to questions of historic interest, with each case highlighting critical issues that arise during the identification process.

Through analysing mitochondrial DNA, researchers have been able to trace human migration patterns and the origins of various populations. The analysis has also provided insights into the domestication of animals, including horses, cows, and dogs. In forensic investigations, mitochondrial DNA analysis has been used to identify the remains of individuals in mass graves associated with past conflicts, such as World War II. The analysis has helped to provide closure for families and communities, and has aided in holding perpetrators accountable for their crimes.

In some cases, mitochondrial DNA analysis has been utilised to answer questions of historical interest, such as identifying the remains of famous individuals from the past or determining the familial relationships between historical figures as accounted below:

1.2.8.1 Romanov Family Identification

In July 1918, the Koptakyi forest became the final resting place for the Russian Imperial Royal family, which consisted of the Emperor, the Empress, and their five children, following their assassination. The remains were first discovered in 1979 and later in 2007, prompting geneticists from Russia and the United Kingdom to carry out DNA testing [90].

Analysis of mitochondrial DNA (HVRI and HVRII) was instrumental in establishing a genetic connection between the purported remains of the Tsarina and Prince Philip, a distant maternal relative. The results of this analysis exhibited a perfect match. However, when investigating the Tsar's femur, a C/T heteroplasmy at position 16,169 was identified, contrasting with the homoplasmic T observed in living relatives. The presence

of this heteroplasmy was later verified through the exhumation of the body of the Tsar's brother, Georgii. Further confirmation of the identification was obtained by analysing the clothing worn by Nicholas II at the time of his demise [90].

Although the identification process yielded positive results, it encountered challenges along the way. A scientist raised doubts by presenting conflicting findings from a blood-stained handkerchief believed to belong to Nicholas II, and preserved hair samples from Georgii showed no signs of heteroplasmy. However, when the scientist's obtained sequence was published, it became evident that the samples were likely contaminated. Another point of contention regarding the identification arose from the examination of a finger bone purportedly belonging to the Tsarina's sister. According to a scientist, this bone did not align with Prince Philip's haplotype. The authors of the study criticized Gill's utilisation of a nested approach, asserting that it yielded a flawed match. Intriguingly, their own analysis uncovered a mixture, and they reported a non-match outcome based on a consensus sequence [90].

The reports sparked extensive discussions within the scientific community, highlighting the significance of independent analysis, result replication, method validation, genetic consistency, and the availability of verifiable reference materials. These essential aspects were noticeably absent in the work conducted by the critics.

1.2.8.2 RMS Titanic Sinking: 15th April, 1915

The sinking of the Titanic resulted in a survival rate of just over 30% among its passengers. The 'Unknown Child,' a young victim of the tragedy, was initially thought to be Gösta Pålsson based on evidence from the time [91]. However, a project conducted in 2001 employed mtDNA analysis of HV1 to confirm the identity but was unsuccessful in establishing a match with Gösta Pålsson. Interestingly, the analysis did result in a match with the families of two other missing children, Eino Panula and Sidney Goodwin. In 2004, odontological age determination leaned towards Eino Panula, yet doubts lingered due to inconsistencies in shoe sizes. Subsequent analysis, involving the sequencing of an extended control region, brought forth two distinct sequence discrepancies when

compared to the Panula references. These notable variations served as crucial evidence, ultimately leading to the identification of the 'Unknown Child' as Sidney Goodwin. This case underscores the challenges inherent in DVI when materials are limited, degraded, and potentially contaminated, particularly when families seek a prompt resolution while scientists must undertake a lengthy and complex analysis to ensure the most secure identification possible [91].

1.2.8.3 The King of France Louis XVII (1793-1795)

Upon inheriting the throne of France at the tender age of eight, Louis Charles' reign was tragically cut short as he passed away soon after. Following the customary practice, the young boy's heart was laid to rest in the Basilica, a hallowed site reserved for the burial of members belonging to the French Royal Family. Nevertheless, unyielding rumours persisted, hinting at the substitution of his body with that of another. Several individuals came forward, asserting their direct lineage to Louis XVII, including a German clockmaker. In 1998, a comparison of MtDNA between the heart sample and living Habsburg descendants showed no match, leaving uncertainty about the heart's identity. To address this question, researchers analysed the heart's HV1 and HVII MtDNA and obtained consensus sequence that exhibited a strong correspondence with living descendants of the Habsburg lineage, establishing a clear maternal relationship. This sequence was not observed in 1700 Europeans, providing support for the identification of the heart as that of Louis Charles. This study highlights the importance of adhering to strict guidelines when analysing ancient DNA, as the analysis of degraded and potentially contaminated material can be challenging [92].

Adhering to the prescribed guidelines, the researchers thoroughly examined the heart samples using two different institutions. As anticipated, most of the sequenced fragments were of limited length, yet they consistently yielded reproducible outcomes. However, when attempting to analyse longer sequences, distinct results emerged, likely attributable to minor contamination associated with PCR artefacts, specifically PCR jumping. This phenomenon can result in the preferential amplification of erroneous longer sequences, contributing to the discrepancies observed [92].

1.2.9 POPULATION STUDIES

Population studies refer to the scientific study of populations, including their characteristics, behaviour, and patterns of distribution [93]. This interdisciplinary field of study encompasses a range of disciplines such as demography, sociology, epidemiology, economics, geography, and anthropology, among others. Population studies focus on various aspects of population dynamics, including population growth and decline, migration, fertility rates, mortality rates, and aging patterns [94]. This information is critical in understanding the social, economic, and health implications of changes in population demographics. Population studies are essential for policymakers, researchers, and planners in making informed decisions and policies.

1.2.9.1 Archeogenetics

Archaeogenetics is an emerging field that aims to understand the history of human populations by combining molecular genetics, archaeology, and anthropology. It involves studying the genetic material of ancient individuals to reconstruct their past and understand their evolutionary relationships with modern-day populations [95]. The emergence of archaeogenetics dates back to the 1960s, marking its nascent phase and the commencement of investigations in this field when the geneticist Cavalli-Sforza conducted research on human blood groups and lactase persistence to study ethnic and linguistic groupings [96]. However, it was not until archaeologist Colin Renfrew coined the term that this field of study was officially recognized [96].

The landscape of archaeogenetics underwent a paradigm shift in the 1980s as new technologies, including Polymerase Chain Reaction (PCR) and Sanger DNA sequencing, reshaped the field. The advancements continued in the 2000s with the introduction of next-generation sequencing techniques, paving the way for more comprehensive and efficient analyses in archaeogenetics. These technologies have enabled researchers to analyse DNA from small and degraded samples found at archaeological sites, also known as ancient DNA (aDNA) [33]. This has greatly enhanced our understanding of ancient civilisations and human migration patterns [97].

In conclusion, archaeogenetics is a rapidly evolving field that has made significant contributions to our understanding of the history of human populations. The use of modern genetic techniques in combination with archaeological and anthropological evidence has allowed researchers to reconstruct the past with unprecedented accuracy and detail.

1.2.9.2 Phylogeography

Phylogeography is an interdisciplinary field that combines phylogenetics and geography to study the historical relationships among populations. It aims to establish the evolutionary history and distribution of lineages over time and space [98]. Key instruments in the study of phylogeography include the construction of phylogenetic trees, analysis of the geographic distribution of established lineages, and the application of molecular clocks to estimate divergence times [99]. Other fields, such as archaeology, paleoanthropology, paleoclimatology, and ethnology, are also integrated to provide a more comprehensive understanding of population history [100].

By combining these different fields, researchers can gain insights into the modern distribution patterns of species, including their origins and dispersal patterns. Phylogeography provides a powerful framework for investigating the genetic and ecological factors that shape the diversity and distribution of species, making it a valuable tool for conservation and management efforts.

Overall, phylogeography is a critical field for understanding the evolutionary history and distribution of populations, and its insights have implications for a broad range of areas, including biodiversity conservation, biogeography, and evolutionary biology.

Phylogenetic Tree

By presenting a graphical depiction, a phylogenetic tree showcases the intricate evolutionary relationships shared among individuals or organisms [101]. IA phylogenetic tree consists of interconnected branches and nodes, where branches represent distinct lineages and nodes signify the ancestral points where lineages split. The presence of a root, if applicable, establishes the orientation of the tree and represents the shared

ancestor from which all individuals or organisms within the tree originated. Each branch terminates in a terminal node or leaf, which represents modern sampled sequences. The relationships depicted in a phylogenetic tree are based on genetic mutations, or the lack thereof, and provide insights into the evolutionary history of the studied organisms or individuals [102], [103].

Phylogenetic trees can be either rooted or unrooted, depending on whether a common ancestor is specified or not. In a rooted tree, an outgroup is used to identify the most ancestral node. On the other hand, an unrooted tree only displays the relationship between branches without indicating ancestry [104].

The construction of phylogenetic trees involves the utilisation of two primary methods: distance-based and character-based. In the distance-based method, the emphasis is placed on assessing the distance between sequences, enabling the determination of genetic dissimilarity or similarity and forms the tree using methods such as UPGMA and Neighbour-Joining (NJ) clustering. UPGMA is a straightforward clustering method, known as Unweighted Pair Group Method with Arithmetic Mean. It operates under the assumption of a constant rate of evolution, known as the molecular clock hypothesis. To perform the clustering, a distance matrix of the taxa being analysed is required, which can be derived from a multiple alignment. NJ clustering is frequently employed because it yields the smallest total branch length among the available options. However, character-based methods are preferred for phylogeography analysis as they take into account the occurrence of mutations and alterations in repetitive motifs. These methods, such as Maximum-Parsimony, Maximum-Likelihood (ML), and Bayesian methods, organize aligned sequences by similarity of characters and better present evolutionary processes [23], [104]–[107].

Maximum-Parsimony calculates the length of each branch based on the number of polymorphisms along it, minimizing the total number of character-state changes [106], [108]. However, this method may lead to statistical inconsistencies called "long branched attraction". Maximum-Likelihood on the other hand generates all possible combinations

of unknown parameters and estimates which tree is more likely to represent the data, but it is computationally demanding. The Bayesian Inference method employs Markov Chain Monte Carlo simulations to generate a series of trees, allowing for the determination of probabilities assigned to each tree [108], [109]. This method is also computationally demanding and is increasingly used in phylogenetics, including in phylogeography. Understanding these different methods is crucial for constructing accurate phylogenetic trees and understanding evolutionary relationships between organisms. [104], [110], [111].

Creating the most accurate phylogenetic tree is a challenging task, and there is no single method that can provide a perfect result. To increase the accuracy of the trees, researchers often use an integration of various methods. By combining different methodologies, the strengths of each approach can be utilised to overcome their limitations [112]. For instance, distance-based methods may be employed to obtain a rough estimate of the evolutionary distance between sequences, while character-based methods can be used to refine the tree by accounting for specific changes in the sequences. In addition, Bayesian Inference can provide a probabilistic framework to evaluate the confidence of the obtained trees. This integration of methods can lead to more robust and reliable results, as it reduces the biases and errors that might be introduced by relying on a single method [113]. Therefore, using an integration of various methodologies is a common and effective approach to construct more accurate phylogenetic trees.

Median Networks

In instances where homoplasmy is observed, indicating recurring mutations in the same mtDNA position, a single phylogenetic tree may not accurately capture the complexity of the data. This is because it can result in loops or reticulations, where multiple equally viable evolutionary reconstructions exist. To enhance the visualization of the various potential evolutionary routes, networks are employed instead of trees. Median networks, which transform variant sites into binary characters and link samples based on their relative distances, are commonly employed in the analysis of mtDNA data [105]. However, when a large amount of data is involved, reticulations can result in large hyper-

dimensional cubes, so rules have been developed to eliminate some of the more likely occurrences [114]. In order to mitigate such complications, specific guidelines have been implemented, including a weighting scheme that assigns higher probabilities of recurrence to fast-evolving sites. This approach aims to minimize the potential challenges associated with the formation of loops or reticulations in phylogenetic analyses [115]. Integrating networks with trees and other methodologies can also provide a more reliable representation of the data [116].

Founder Analysis

The application of founder analysis offers a powerful quantitative approach to unravelling the phylogeographic dynamics of populations. Its central objective revolves around identifying, quantifying, and dating the migrations of a population from a source region to a distinct territory, by examining the lineage diversity derived from the "source" population and its subsequent establishment in a specific "sink" location [117].

First introduced by Richards and co-authors, this methodology has seen extensive application in the investigation of the ancestral lineages that serve as the roots of global populations [118]. However, its complex mathematical nature has limited its widespread use. To address this issue, a user-friendly program was recently developed based on the principles of the original methodology [99].

In founder analysis, the primary step involves the recognition of shared sequences or clades between the source and sink populations. After identification, the subclades within the sink population are isolated, and by evaluating the diversity accumulated in the newly colonised regions, the migration time is estimated using a molecular clock. By quantifying the diversity of lineages that migrated from the source to the sink, it is possible to infer the scale and timing of these migration events [119]. This method is particularly useful in understanding the historical patterns of human migration and colonisation, and has provided valuable insights into the origins and dispersal of populations around the world.

Founder analysis can be challenged by common mtDNA mutations and reverse gene flow. To overcome this issue, multiple criteria have been established to ascertain the characteristics that define a founder group, distinguishing what falls within its scope and what does not [95], [99], [118], [120].

Molecular Clock

Understanding the timing and frequency of mutations is critical in phylogeography, and a reliable molecular clock is necessary for dating the divergences between branches and subclades. The evolution of molecular sequences is generally assumed to occur at a steady rate, allowing accumulated diversity to be used to infer temporal events [121], [122]. Founder analysis can be used in conjunction with a molecular clock to facilitate the assessment of the earliest feasible time-frame for the arrival of each founder cluster in the sink population [118].

Most molecular dating models assume a known and constant mutation rate, often using either the coding or non-coding regions of mtDNA as a reference. For example, according to the research conducted by Forster et al. (1996), the substitution rate in the non-coding HVS-I section of the mtDNA control region was estimated to be 1.80×10-7 substitutions per nucleotide per year [123]. Additionally, Mishmar et al. (2003) reported a lower substitution rate of 1.26×10-8 substitutions per nucleotide per year in the coding region of mtDNA [124], [125]. However, using only a single region can introduce issues, such as different mutation rates in coding and non-coding regions (with the non-coding region having a higher rate) and the effect of natural selection on mutation rates over time. Due to the influence of purifying selection on slightly harmful mutations, the oldest branches in a phylogenetic tree may exhibit a higher rate of synonymous mutations, resulting in a non-linear relationship between mutation accumulation and time [126].

Soares et al. introduced a recalibrated molecular clock as a solution to these challenges, which encompasses the entire mitochondrial genome, including coding and non-coding segments. This refined approach also integrates the influence of purifying selection. They put forth a mutation rate of 1.665×10-8 substitutions/nucleotide/year, which corresponds

to an estimated occurrence of one mutation every 3,624 years. Furthermore, this rate has been precisely adjusted through mathematical techniques to account for the effects of natural selection [98], [125].

The recalibrated molecular clock proposed by Soares et al. (2009) has been used in various studies to estimate the divergence times of different human populations and to trace their migration routes. It has also been applied to founder analysis, which aims to identify the source population and estimate the time of migration to a new location. The use of the entire mitochondrial genome in this approach ensures a more accurate estimation of the mutation rate and accounts for the effect of natural selection.

In summary, the understanding of mutations and their rate is essential for molecular dating and phylogeographic analysis. While the mutation rate is theoretically constant, different regions of the mitochondrial genome may have different rates, and the effect of natural selection must be taken into account. The recalibrated molecular clock proposed by Soares et al. (2009) using the entire mitochondrial genome and adjusted for purifying selection provides a more accurate estimation of the mutation rate and is widely used in phylogeographic studies.

1.2.10 AFRICAN ARCHEOGENETICS AND PHYLOGEOGRAPHY

Africa is widely recognized as the birthplace of modern humans, and it played a significant role in the global expansion of our species (Batini & Jobling, 2011; Fregel et al., 2019; Teresa Rito et al., 2013). The ancient remains of Homo sapiens, discovered in Ethiopia, are considered the earliest evidence of modern humans and are estimated to be around 190,000 years old [129]. Contemporary fossil discoveries in Morocco, dated to 315,000 years ago, suggest that Ethiopia played a significant role in the evolution of modern humans, contributing to their development alongside East Africa [129]. These discoveries suggest that modern humans had a longer history in Africa than previously anticipated, underscoring an extended period of their existence on the continent prior to spreading across the globe [127], [130]. Despite the ongoing discussions surrounding the emergence of modern humans, the prevailing agreement is that Africa serves as the

established cradle of humanity, with humans having an extensive history on the continent that surpasses any other location [131]. However, the exact location of the origin within Africa is still debated, with some researchers suggesting South Africa [10], [132], , in recent times, Namibia, Botswana and Zimbabwe [133] and East Africa [11], [134].

Understanding the origin and evolution of modern humans is a complex and challenging task. One of the main theories is the Out-of-Africa theory model, which suggests that the cradle of modern humans is believed to be Africa and subsequently migrated and substituted other hominin populations in other regions of the world. Another theory is the Multi-regional evolution theory model, that proposes evolution of modern humans from multiple regions simultaneously through a coalescence of existing archaic populations. Both theories have their strengths and weaknesses and are likely to have occurred in some combination.

To determine the origin of modern humans, scientists have used a variety of methods, including genetic, paleoanthropological, and archaeological evidence. For example, fossil findings from Omo I and Herto support the idea that modern humans evolved in eastern Africa, while the Jebel Irhoud remains in Morocco provide evidence of early modern humans outside of Eastern Africa.

Genetic evidence, particularly DNA analysis, has played a significant role in understanding modern human evolution, especially in South African populations [28], [135]–[137]. According to the multi-regional hypothesis, the development of modern humans took place through the evolution of diverse populations in different parts of the world, stemming from a common ancestor [128], [138], [139]. Based on compelling fossil and chronological evidence, the prevailing view is that modern humans emerged in Africa according to the Out of Africa theory, and subsequently populated different areas of the world through migration [28], [127], [133], [140]. Although the precise mechanism of evolution is yet to be elucidated, it is apparent that modern humans emerged over time, displaying a continuous morphological spectrum from archaic H. erectus to modern H. sapiens [127], [141].

During a period approximately 150-200 thousand years ago, when the global population of anatomically modern humans hovered around 10,000 individuals, a figure named Mitochondrial Eve became the focal point of the female lineage. As the most recent common ancestor, she stands as the solitary woman from the past whose direct female descendants persist in the present, representing a convergence point within the maternal tree of life [142]. Nevertheless, the precise whereabouts of Mitochondrial Eve remain elusive, primarily due to the complexities of phylogeography and the absence of deeper lineages to follow. Nonetheless, the analysis of nucleotide diversity within the HVS-I region of mitochondrial DNA points towards a likely location in Central Africa [98].

The concept of Mitochondrial Eve was first introduced by evolutionary geneticist Allan Wilson and his colleagues in 1987, based on the study of mtDNA variation in contemporary human populations [143]. The hypothesis put forth states that a single female progenitor, located in Africa, is responsible for transmitting her mitochondrial DNA to all individuals in the human population and was part of a larger population of anatomically modern humans [144].

The discovery of Mitochondrial Eve sparked a great deal of excitement and interest in the scientific community, as it offered a new and powerful tool for understanding the origins of the human species. Through the study of mtDNA, researchers could trace the maternal lineages of human populations back to a common ancestor and gain insights into the migration patterns and evolutionary history of our species.

Despite its significance, the concept of Mitochondrial Eve has also been subject to some controversy and misunderstanding. For example, some people have interpreted the term to suggest that Mitochondrial Eve was the only woman living at the time, or that she was the first anatomically modern human. However, these interpretations are not accurate, as Mitochondrial Eve was just one member of a larger population of early humans who lived in Africa during that time.

Nonetheless, Mitochondrial Eve remains a critical concept in the study of human evolution and genetics. The evolutionary journey of our species has been unveiled by

researchers through the examination of mitochondrial DNA within contemporary human populations and gain insights into the migration patterns and population dynamics that have shaped our genetic diversity.

Understanding Mitochondrial Eve is crucial in our understanding of human evolution and the origin of modern humans. Despite the challenges, researchers continue to use genetic data to gain a better understanding of our past and the origin of humanity. By analysing the genetic information and developing new techniques, researchers may be able to uncover new insights into the origin of humanity and the lineage of our ancestors.

The study of human lineage and the ancestral history of our species has been illuminated through the exploration of the Y-chromosome, which tracks the male lineage. The findings point towards a probable central or western origin of the male lineage [145], [146] dating back approximately 350,000 years to the point of coalescence [100]. In contrast, the information conveyed by mtDNA, which pertains to maternal lineages, presents mixed signals. The lineages of mtDNA diverge into two separate branches: L1’6 and L0 [28], [134], [147]. L1'6 is indicative of the Y-chromosome's origin or its association with the northern territories of the continent [148]. However, due to depopulation resulting from climate changes around 75-65kya, the actual historical records of the northern regions of Africa are absent. Interestingly, this time period coincides with the emergence of L3 and the subsequent migration out of Africa [98]. Around 40-20ka, the region saw a resurgence in population as individuals from Eurasia migrated into the area [149], [150]. On the other hand, The origin of L0 can be traced back to the southern part of the continent, where it is predominantly found among Khoisan populations and Bantu speakers [151], [152]. These discrepancies between the male and female lineages make the origin of AMH still a mystery. A holistic comprehension of the origins and progression of modern humans necessitates the integration of multiple approaches and continue to study and analyse new evidence as it becomes available. By doing so, we can gain a better understanding of our own evolutionary history and the factors that have shaped our species over time.

In conclusion, understanding the origin of AMH is critical to our understanding of human evolution. Although genome-wide studies suggest a southern African origin, gender-specific markers reveal contrasting results, making it challenging to determine a clear picture of human evolution. It is essential to continue investigating and analysing the genetic data to shed light on our evolutionary past.

1.2.11 HISTORY OF HUMAN MIGRATION IN AND OUT OF AFRICA

1.2.11.1 Human Migration in Africa

Africa has a complex history of migration, with different time points and geographic areas involved. According to Campbell et al., (2017), one such migration occurred over 20,000 years ago, when the predecessors of existing African hunter-gatherer populations and Khoisan speakers dispersed across central, eastern, and southern Africa in their migratory movements. However, Chan et al., (2019) suggested that the distribution of Khoisan and Khoisan ancestors diverged much earlier, around 69,000 years ago, in a region southwest of the Zambezi River. This indicates that the migration and diversification of populations in Africa may have occurred much earlier than previously thought. The relevance of this finding lies in its potential to deepen our understanding of the evolutionary history and genetic diversity of African populations, which has important implications for health and disease research, as well as social and cultural studies.

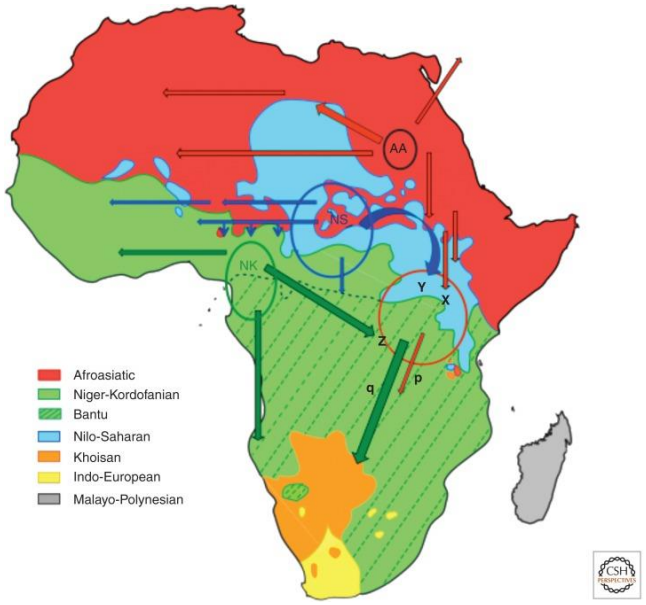


Figure 1.2: Geographic distribution of main African linguistic groups [153]

A visual representation in the form of an image showcases the spatial distribution of the major linguistic groups across Africa provided by Gomez et al., (2014). The figure shows three major groups: Afro-Asiatic (AA), Nilo-Saharan (NS), and Niger-Kordofanian (NK). The image provides a helpful overview of the linguistic diversity across the continent. Understanding linguistic diversity is crucial for research into African cultures, history, and society, as well as for communication and translation purposes.

Archaeological evidence suggests that migrations played a significant role in shaping the genetic landscape of Africa. One contributing factor was the expansion of people engaged in agriculture from Central-West Africa, along with migrating pastoralist communities from North-Eastern Africa, contributed to the population dynamics and movement across regions [153]. This is supported by a range of archaeological findings that provide insight into the movement of people and cultures across the continent. According to Gomez et al. (2014), several populations migrated in Africa, which further underscores the complexity and diversity of the continent's history. Understanding these migrations and their genetic impact is crucial for the development of accurate and

comprehensive models of African history and the study of the evolution of human populations.

Over thousands of years, several ethnic groups migrated into Ethiopia, forming a network of genetic diversity in the African human genome. The first group to migrate were the Afro-Asiatic-Agro pastoralists from the Nile Valley between 8000-5000 years ago [154]. Originating from Chad/Sudan, the Nilo-Saharan pastoralists embarked on a westward migration across the Sahel roughly 7000 years ago. Subsequently, they extended their movement eastwards into Kenya and Tanzania around 3000 years ago [155].

The Bantu speakers, who belong to the Niger-Kordofanian group, migrated across sub-Saharan Africa from Cameroon/Nigeria about 5000 years ago, as evidenced by linguistic and archaeological evidence [136]. Lastly, Southwestern Asians migrated into Africa north- and eastwards about 3000 years ago, leaving traces of their DNA in the African genome [30], [119].

The arrival of these different ethnic groups into Ethiopia resulted in the formation of a diverse genetic network within the African human genome. This diversity has contributed to the unique cultural and genetic heritage of Ethiopia, making it an important centre of human diversity and a place of interest for genetic research [156].

1.2.11.2 Human migration out of Africa

Around 40,000 years ago, it is believed that Anatomically Modern Humans (AMH) migrated out of Africa, traversing various regions including India, China, Central Asia, Indochina, Sunda, Sahul, and eventually reaching America [157]. Researchers propose a single southern dispersal out of Africa by African ancestry for AMH, moving from the Red Sea towards the Indian Ocean coastline to Bali, and then southwest to Melanesia and Australia [95].

However, morphogenetic models suggest primarily two major pathways used for migration. The initial route followed by early humans involved traveling along the southern path, starting from the Red Sea approximately 50,000 to 45,000 years ago. They moved along the South Asian coastline towards Australia, where populations like the

Andaman islanders, Southeast Asian Negritos, Melanesians, and Australians were later discovered [158]. Genetic traces of Japan and North America were also present. This wave of migration is known as the Australian wave, as it culminated in the settlement of Australia.

The unique genetic signatures of various populations in the regions mentioned above are evidence of human migration and colonisation patterns. Understanding these patterns is crucial for understanding human history, population genetics, and the evolution of the human species.

Another significant migration route out of Africa was the Asian wave, which occurred after the Australian wave. The migration took place along a North-Eastern trajectory, as populations from Eurasia moved from Africa to the Levant approximately 45,000 years ago. Around 45,000 years ago, there was a migration following a North-Eastward path, as populations from Eurasia moved from Africa to the Levant. From there, they dispersed westward towards Europe and eastward towards Eurasia around 40-20 kya [135]. This migration led to the development of unique genetic signatures in the populations of Europe and Asia, which can be traced back to their African ancestry.

Understanding the migratory patterns of humans out of Africa is vital in understanding the evolutionary history of the human species. The distinct genetic markers of different populations can help us to trace their migration patterns and understand how they have adapted to different environments over time. Such knowledge can provide insights into the history of human populations and their interactions, as well as inform modern medical research and genetic studies.

According to Oppenheimer, (2012), all non-African groups today are descendants of the out-of-Africa migration, apart from a small minority comprising 7% of these populations resulting from genetic mixing with prehistoric non-African groups. Both European and Asian populations might have been established by a common African exit group. Around 60-65,000 years ago, the M and N lineages, derived from L3, diverged and led to the emergence of all non-African populations [123], [124], [150], [160].

This finding suggests that all non-African populations share a common ancestor, and that the genetic diversity observed among these populations is a result of subsequent migrations, genetic drift, and natural selection. The knowledge of the origins of non-African populations is important in understanding the genetic makeup of humans and how they have migrated and adapted to different environments over time. This information can be applied to various fields, including genetics, anthropology, and medicine.

The Atlantic slave trade of the 16th century played a significant role in shaping the genetic makeup of Brazilians. Long-distance gene flow resulted in Brazilians harbouring African maternal lineages that hold particular significance beyond Africa [20]. Nearly half of these African lineages were identified as L1c and L3e lineages. Studies have indicated that the genetic makeup of Afro-Americans in South America reveals a predominant Central African origin, accounting for about 65% [101].

The genetic diversity of populations and its connection to the historical transatlantic slave trade is a topic of great significance in genetics and anthropology. It highlights how historical events have shaped the genetic makeup of populations, and the need to understand the origins and movement of populations [35]. This knowledge can help identify genetic risk factors for diseases and develop more effective treatments that are tailored to specific populations. Additionally, it can also aid in the identification of populations that are at risk of genetic diseases and help develop targeted public health interventions.

1.2.12 AFRICAN GENETIC DIVERSITY

African populations exhibit remarkable levels of genetic and phenotypic diversity compared to other human populations. Such diversity provides valuable information for understanding the demographic history of populations in Africa, including periods of growth, contraction, migration, and admixture [10]. In addition to this, genetic variation also allows for the identification of genetic adaptations in response to diverse environmental conditions [131]. The characterisation of human genetic variation and

phenotypic diversity in contemporary African populations is therefore critical in identifying the genetic basis of functional traits, adaptation, and susceptibility to complex diseases.

Africa is known for its linguistic diversity, with over 2000 indigenous languages spoken on the continent. The linguistic landscape of Africa is characterised by four major language families: Niger-Kordofanian, Nilo-Saharan, Afroasiatic, and Khoisan. The Niger-Kordofanian family, prevalent among agriculturalist populations, is widely spoken across sub-Saharan Africa, spanning from West Africa to eastern and southern Africa. The Nilo-Saharan family, on the other hand, is spoken primarily by pastoralist populations in central and eastern Africa [148]. In northern and eastern Africa, the Afroasiatic family of languages is mainly spoken by agro-pastoralist and pastoralist populations. On the other hand, the Khoisan family, which features click consonants, is prevalent among hunter-gatherer San communities in southern Africa and among the Hadza and Sandawe hunter-gatherers in Tanzania [161].

1.2.12.1 African mtDNA Diversity and Haplogrouping

Clusters of variants with a shared ancestry, known as haplogroups, are organized in a hierarchical manner into clades based on common mutations. Haplotypes, however, pertain to a set of alleles present at a linked locus and can be found on one of the two homologous chromosomes. The mutations along the mtDNA molecule define the haplotypes. Assigning a haplogroup can provide insights into the maternal lineage's likely geographic origin. Furthermore, it can be utilised for quality assurance purposes [77]. As the phylogenetic patterns and motifs for haplogroups are well-established, the provided information can be utilised for retrospective analysis. If unexpected variants are detected in a haplotype that has been assigned to a haplogroup, it may indicate errors in sequencing or the interpretation of the data.

The phylogenetic tree in Figure 1.3 is divided into four major branches, referred to as Macrohaplogroups (L, L3, M and N), which are further divided into sub-branches known as haplogroups. In total, there are eight major haplogroups identified in the tree. The

position of an individual's ancestry on the mtDNA tree can be determined using a diagnostic tool designed based on multiplexes of SNPs [162]. This tool helps to assign haplogroups and to screen for specific mutations associated with certain diseases. These haplogroups can provide valuable information about human migration patterns and population history.

Torroni et al., (1993) pioneered identification of mtDNA haplogroups in Native Americans by utilising the letters A, B, C, and D to designate the observed haplogroups. However, as research progressed, the haplogroup nomenclature changed over time. Currently, the African mtDNA haplogroups, excluding L3 (M and N), are referred to as African Hg L (xM, N). The evolving sides of the root (L0 and L1′6) are referred to as branches. Within the mitochondrial DNA phylogenetic tree, the L0 branch represents haplogroup L0 exclusively, while the L1′6 branch encompasses haplogroups L1-L6 and includes the ancestral L3-sub-branch, which played a key role in the out-of-Africa migration. The tree illustrates two main branches: L0 (found in Southern Africa) and L1′6 (found in Central and Eastern Africa). Rito et al., (2019) conducted an mtDNA analysis specifically examining the L1′6 and L2′3′4′5′6 sub-branches. Macrohaplogroups L, M, and N are classified based on their distinct geographic origin. The highest percentage of diversity within the human gene pool is found in Africa, primarily comprising L haplogroups. Understanding the diversity and distribution of mtDNA haplogroups can provide valuable insights into human migration patterns and population history.

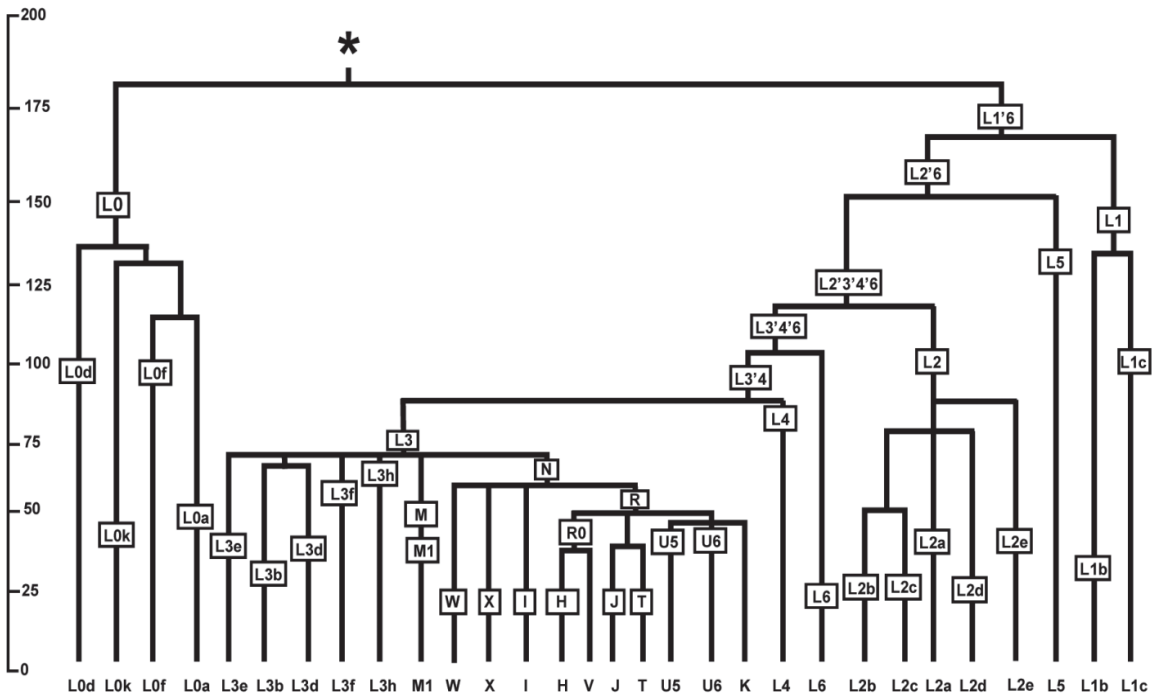


Figure 1.3: Main haplogroups and migration routes over time based on phylotree build 16

African haplogroups have emerged due to various evolutionary forces such as climate changes, migrations, and other factors. These haplogroups are classified based on their estimated time of emergence and are arranged in the following order: L0, L1, L5, L2, L6, L4, L3. These haplogroups can be grouped into two clusters, L0 and L1'6. Several researchers, including (Behar et al., (2008); Rito et al., (2019; 2013); Rosa & Brehm, (2011); Salas et al., (2002); and Soares et al., (2009, 2012, 2016) have contributed to the understanding of the genetic ancestry and evolutionary tree of human mtDNA, which is depicted in Figure 1.3.

The superhaplogroup L, which is an important part of African genetic diversity, began to differentiate around 180-190 thousand years ago. The first branch to emerge in sub-Saharan Africa around 150 thousand years ago was L0, which has distinct subhaplogroups known as L0a'b'd'f'g'k. Among them, L0d appeared approximately 100,000 years ago, and when combined with L0k, they represent a significant portion of

the maternal genetic diversity. Notably, these subhaplogroups are predominantly found among the Khoisan population in South Africa, as well as the click-speaking populations in Tanzania and Angola (Tishkoff et al., 2007).

The understanding of the emergence and distribution of L0 subhaplogroups has provided insights into the migration and genetic history of early human populations in Africa. Furthermore, the study of these subhaplogroups has helped in the identification and diagnosis of certain genetic disorders prevalent in specific African populations.

In summary, L0 is a significant branch of the superhaplogroup L that emerged around 150 thousand years ago in sub-Saharan Africa. The subhaplogroups of L0 provide valuable information regarding the genetic diversity and evolution of early human populations in Africa.

The classification and understanding of African haplogroups have helped in the study of human genetic diversity and evolution. These haplogroups have provided insights into the migration patterns of humans across the African continent and beyond. They have also been useful in the study of various diseases and have been linked to certain genetic disorders.

Based on genetic evidence, we can discern the proposed origins, coalescence ages, and relevant demographic and migratory events associated with African mtDNA haplogroups and their subclusters. It is important to note that coalescence ages in this text are derived from various sources, including HVS-I, coding region, and full mtDNA information, and as such may not be coincident.

It is important to acknowledge that when analysing the mitochondrial DNA (mtDNA) haplogroups in Africa, it is crucial to consider both ethnolinguistic connexions and geography. Research conducted by several researchers indicate that there is significant variability in the distribution and sub-structuring of mtDNA haplogroups across the continent [9], [27], [33], [60], [140], [164]. These findings emphasise the need for careful consideration of these factors when conducting genetic research in African populations.

Haplogroup L0

Studies have identified the macrohaplogroup L, predominantly found in sub-Saharan Africa, which further branches into haplogroups L0-L6. Extensive research has focused on this subdivision, particularly highlighting haplogroup L0 as one of the earliest mitochondrial DNA haplogroups. It is noteworthy that haplogroup L0 serves as the sister clade to all other extant haplogroups in anatomically modern humans (AMH) [27], [124], [156], [165]. The substructure of haplogroup L0 includes several sub-haplogroups, including L0a, L0d, L0f, and L0k. Of these, L0d emerges as the earliest sub-clade branching from the L0 node, estimated to have a chronological age of around 100,000 years. Its distribution appears to be limited to specific populations, notably the Khoisan people in South Africa, along with populations in Tanzania and Angola, as reported by various studies, including[12], [102], [161], [166]–[169]. The most recently accepted tree topology allows for the identification of L0k as a sister clade to L0abf, and this sub-haplogroup is predominantly observed among the Khoisan populations in South Africa. These genetic lineages are frequently found in this region, whereas their prevalence among click-speaking Tanzanian groups is relatively low, as reported by [24], [27], [102], [161]. These findings highlight the need for further research into the distribution and diversity of mtDNA haplogroups in Africa.

The shared L0d and L0k maternal lineages in the population suggest an ancient link predating the existence of click-speaking communities. This connection may be a remnant of an East African proto-Khoisan population, as proposed in earlier research. According to Behar et al. (2008) L0k lineages have been dated to approximately 40,000 years ago. Meanwhile, the rare L0f lineages are found mainly in East Africa, with the highest incidence in Tanzanians, and are thought to have originated around 85-90,000 years ago [24], [27], [105], [139], [167].

The lineage of L0a is believed to have originated in eastern Africa during Paleolithic times, approximately 40-55 thousand years ago. In modern-day Africa, the L0a lineages are extensively dispersed across eastern, central, and southern regions of the continent.

These lineages make up a significant portion, surpassing 25%, of the maternal lineages in certain geographical areas [170]–[172]. The L0a1 sub-clade is predominantly found in eastern and southeastern Africa, including Nubia, Sudan, and Ethiopia, and is estimated to have originated about 30,000 years ago [125]. While L0a1 is also present in West Africa, it is found at relatively low frequencies [20], [26], [173]. L0a phylogenetic tree suggests recent population growth as evidenced by several short branches. The specific L0a2 lineages are believed to track the Bantu-speaking people's migration to South Africa about 3,000 years ago [174].

L1 Haplogroup

The L1 lineage of mtDNA is particularly interesting as it coalesces to a common ancestor around 140-150 kya, according to studies by [45], [139]. Several offshoots originated from this particular lineage, with one notable branch being haplogroup L1b. This haplogroup demonstrates a pronounced concentration in the coastal areas of western-central Africa [157].

Multiple studies have documented the distribution of haplogroup L1b in various populaces (Johnson et al., 2015; Olivares et al., 2021; Osman et al., 2021; Rosa & Brehm, 2011) all reported the prevalence of L1b in different populations in western-central Africa. The highest frequency of haplogroup L1b is found in the Wolof and Mandenka of Senegal, according to [141], [178]. Similarly, Černý et al., (2006) found significant frequencies of L1b in Fulani populations in Chad, South Cameroon and Burkina-Faso [179]. The evolutionary trajectory of haplogroup L1b has been shaped by a significant bottleneck event, resulting in the current variation and genetic characteristics of this lineage. As a result, only one clade expanded after this event, which occurred around 30,000 years ago [33]. This is a clear example of how a bottleneck can shape the genetic variation of a population over time. By studying such events, we can better understand the mechanisms that drive evolutionary processes and the ways in which they impact the diversity of life on our planet [179].

L1c, which is the sister clade of L1b, is found frequently in Central and West Africans. The genetic heritage of many Pygmy populations is overwhelmingly influenced by haplogroup L1b, representing more than 70% of their maternal lineage (Montano et al., 2011). Interestingly, in more recent studies, Angola's Bantu ethnic groups has been documented with frequencies ranging from 18% to 25%. [157], [181]. These findings highlight the complex distribution and diversity of mitochondrial DNA haplogroups across different populations, and underscore the importance of studying genetic variation to gain insights into human evolution and migration patterns.

In their notable study, Quintana-Murci et al., (2008) put forward a significant reevaluation of the L1c phylogeny, providing further insights into the cultural shift from hunting-gathering to agriculture. This research served to support previous findings that indicated a connection between Central African Bantu-speaking farmers and their Neighbouring Pygmy hunter-gatherers, as previously suggested by Batini & Jobling, (2011) and Antonio Salas et al., (2002). According to the proposition put forth by the authors, it is highly probable that both groups shared a common ancestral population from Central Africa that exhibited a significant presence of L1c mtDNAs. This ancestral population began to diverge and evolve in isolation no later than 70,000 years ago. Over time, the diverse forms of L1c observed among modern agricultural populations, including L1c1a, L1c1b, L1c1c, L1c2-6, and more, emerged from this population. Noteworthy is the fact that the only surviving clade among the western Pygmies is L1c1a, as elucidated in the study conducted by Quintana-Murci et al., (2008). It is interesting to note that L1c lineages have also played a role in tracing gene flow between the ancestors of both farming Bantu and hunter-gatherer Pygmy groups, occurring around 40,000 years ago. Both L1b and L1c have been proposed as autochthonous lineages of Central Africa, with L1c estimated to have originated between 85-100 thousand years ago [125], [139], [167]. The existence of these lineages along the West Atlantic coast suggests a westward expansion, and it is plausible that their introduction to Northwest Africa occurred comparatively later, either during the Neolithic epoch or during periods linked to the

slave trade [33], [164], [182]. These findings provide insight into the complex migration patterns and evolutionary history of human populations in Africa.

L5 Haplogroup

Haplogroup L5, which was formerly known as L1e, is believed to have emerged around 120-140 thousand years ago and assumes an intermediate stance between L1 and L2'3'4'6 (Gonder et al., 2006; M. Silva et al., 2021; N. M. Silva et al., 2012). In numerous eastern African nations, including Egypt, Sudan, Ethiopia, Kenya, Rwanda, and Tanzania, the presence of this haplogroup has been documented, albeit in small frequencies. There have been sporadic instances of gene flow introducing these lineages into the Mbuti Pygmies and North Cameroon Fali populations (Davidovic et al., 2020; Kloss-Brandstätter et al., 2021; Lorente-Galdos et al., 2019; S. A. Tishkoff et al., 2007). L1c and L5 lineages are notable components of the genetic repertoire found among Central African Pygmies, underscoring their potential "relict" status akin to the Khoisan groups [139], [141], [167], [178], [186]. The presence of haplogroup L5 in eastern Africa and its association with Pygmy populations highlights the complexity of human genetic diversity and the need for further studies to fully comprehend it.

L2 Haplogroup

Recent studies have suggested that the expansion of the Bantu-speaking populations during the sub-Saharan agricultural spread, and later on, may be associated with the star-like demographic bursts observed in L2a1a and L2a2, and their subsequent expansion into southeast populations [131], [141], [187]. This highlights the importance of considering demographic events and cultural practices in addition to geographical factors when investigating the distribution and evolution of haplogroups and their sub-clades in human populations.

Haplogroup L2, along with L3, constitutes approximately 70% of the sub-Saharan maternal variation. Ground-breaking research conducted by Chen et al., (1995) made significant contributions to the understanding of haplogroup L2 by delineating its sub-clades, including L2a, L2b, L2c, L2d, and L2e [178]. Of these sub-clades, L2a has

emerged as the most predominant and widely distributed mtDNA cluster within Africa. It has been documented to account for over 40% of the genetic makeup in various populations, such as the Tuareg populations in Niger/Nigeria and Mali, Fali populations in North Cameroon, Western Pygmies in Gabon, and Mozambique Bantu populations [9], [12], [22], [27], [170], [188]. The elusive geographic origin of L2a poses a significant challenge, as researchers continue to grapple with uncovering its exact location, approximately 45,000 to 55,000 years ago, exhibits parallel patterns in both East and West Africa (Behar et al., 2008; Antonio Salas et al., 2002; Soares et al., 2009). One possible explanation for this is that haplogroup L2a originated in central Africa and subsequently dispersed westward and eastward along the Sahel corridor after the LGAM, the coalescence of shared founder types took place at an estimated age of 14,000 years [33], [189].

Haplogroups L2b-L2d are prominent among the sub-Saharan maternal variation, and their distribution is mainly limited to West and West-Central Africa. Studies suggest that these haplogroups most likely, it is from this region that these populations have their roots [26], [119], [151], [166]. The estimated coalescence time for L2c and L2b is around 30-25*kya* (Behar et al., 2008; Antonio Salas et al., 2002). Interestingly, specific lineages of L2b and L2c in West Africa are estimated to have expanded around 18 kya, similar to L1b in the same region [20], [175]. On the other hand, L2d is a less frequent clade in West and Central Africa, which around 100-120 thousand years ago, a divergence from the L2 root occurred [20], [119], [139]. However, the extant variation of L2d does not exceed 25-30 kya [139], which suggests that it most likely originated in Central Africa rather than West Africa.

L6 Haplogroup

The maternal variation of haplogroup L6, as proposed by Kivisild et al., (2004), is mainly found in Yemeni individuals, with only a few samples in Ethiopian Amhara and Gurages. L6 has a relatively recent coalescence time of about 22 kya [139], likely due to past variation being wiped out or never expanded. The existence of L6 among Ethiopians,

along with the abundant presence of its related clades, strongly suggests that its origin is rooted in East Africa, where it is likely to have undergone prolonged isolation. Nonetheless, the precise ancestral homeland of L6 remains elusive [182].

L4 Haplogroup

Haplogroup L4 is mainly found in East and Northeast Africa, although at low frequencies. It is a sister clade of haplogroup L3 and has been studied in various populations such as those in Sudan, Tanzania, Ethiopia, and other regions (Kelly et al., 2022; Osman et al., 2021; Sirugo et al., 2019). L4 nomenclature has been updated, with L7 now referred to as L4a and L3g/L4g as L4b2. The occurrence of L4a in Sudan and Ethiopia, originally mistaken for L3e4, has been documented. L4b2 is frequent in Tanzania, Amhara, and Gurages from Ethiopia. The coalescence estimates of L4b2 and L4a are around 90 and 55 kya respectively, whereas L4 clade is projected to have originated around 95 kya, possibly in East Africa. These findings suggest that haplogroup L4 has an East African origin and for several millennia, it has been safeguarded and kept separate from other populations [139], [156].

L3 Haplogroup

The origin of superhaplogroup L3 is generally believed to be in East Africa, around 60-75 kya, based on various studies (Fähnrich et al., 2023; M. B. Richards et al., 2016; Antonio Salas et al., 2002; Soares et al., 2012) . L3 is present throughout Africa and its diversity and the frequencies observed suggest that the sub-clades of this lineage spread from sub-Saharan Africa to West Africa [139], [164], [192]. The superhaplogroup exhibits a hierarchical structure comprising various clades, with the M and L superhaplogroups being the primary ones identified in populations outside of Africa.

The distribution of L3b and L3d is predominantly concentrated in the western region of sub-Saharan Africa, with a mean frequency of 10% [20], [22], [164], [177]. Studies conducted by [144] and [102] demonstrate the presence of L3b in the Hutu population of Rwanda and the South African Kung, respectively, with substantial frequencies. L3d is

abundant in the maternal gene pool of South Africa and exhibits significant prominence in Angola and Tanzania, as evidenced by studies [12], [161], [193].

Approximately 70,000 years ago, L3b and L3d diverged from a common ancestral node, followed by L3b coalescing around 20,000 to 30,000 years ago and L3d coalescing around 30,000 to 40,000 years ago (Behar et al., 2008; A Salas et al., 2007; Antonio Salas et al., 2004; Antonio Torroni et al., 2006). As observed with other haplogroups, a particular lineage within L3b is prevalent among Bantu speakers in south-western Africa, implying its potential connection to the Bantu expansion and migration [135], [192]. The L3e cluster has been fragmented into several subclades, including L3e1, L3e2, L3e3, and L3e4, as researchers began analysing the specific details of the HVS-I [121]. The ancestral branches of L3e are estimated to have originated around 40-50 thousand years ago in what is presently Central Africa, specifically in the area that encompasses Sudan [20], [27], [45], [95], [139].

L3e1 probably originated about 16,000 years ago in central Africa and is now prevalent in Mozambique, the L3e cluster is prominent among southeast Bantu speakers, indicating its widespread presence in this region [56], [189], indicating that Bantu must have migrated through the eastern route. The L3e2b lineage, a subset of the L3e2 haplogroups, demonstrates broad geographical distribution and high frequency, with its highest prevalence observed in West and Central Africa [33], [151], [164], [175].

The L3e2b and L3e2a lineages are widespread and the early Holocene witnessed significant population movements in the Sahara, and these haplogroups seized the opportunity to hitchhike and propagate, contributing to their success in terms of geographical expansion, with an estimated range expansion at about 9 kya. Despite their MRCA arising around 25,000-35,000 ya, their expansion highlights their ability to adapt and thrive in new environments. L3e3 is primarily found in West African people, with an estimated coalescence age of 14 kya, while L3e4, with an age of nearly 24 kya, is essentially restricted to Atlantic West Africa[49]. The existence of L3e lineages among

Southeast Africans suggests a potential link to the eastern branch of Bantu people, signifying their involvement in population movements and cultural exchanges.

Earlier studies detected similar sequences in Tunisian Berbers, but these were not named and were suggested to have originated from North Africa. Conversely, Černý et al., (2006) established a clear genetic ancestry for L3e5 subgroups, primarily observed in the western regions of Central Africa. Although there has been some migration into North Africa, the original variant remains more widespread among populations in the Chad Basin and is projected to have expanded approximately 12,000 years ago. The distribution of haplogroup L3f spans across a range of regions, encompassing Ethiopia in the east, Angola and Mozambique in the south, the Chad Basin in Central Africa, Guinea-Bissau in the west, and Tunisia in the north. Multiple studies have shed light on this widespread distribution, offering robust evidence for the presence of L3f in these diverse locations (Cerny et al., 2016; Johnson et al., 2015; Lambert & Tishkoff, 2009; Quintana-murci et al., 1999; A Salas et al., 2007; Trovoada et al., 2004; Watson et al., 1996). By analysing coalescence estimates and identifying a limited number of matches to L3f1 founder lineages in Central and West Africa, it becomes apparent that these lineages experienced a local dispersal event. This dispersal is believed to have occurred approximately 30,000 years ago according to HVS-I estimates or around 50,000 years ago based on full mtDNA data [139]. Contrastingly, the expansion of L3f1 only began in East Africa approximately 10 kya [100]. Interestingly, L3f2 is a rare genetic subgroup that is predominantly found within Chadic-speaking populations residing in the Chad Basin, while it is scarcely present among Niger-Congo and Nilo-Saharan groups [22], [49].

The MRCA of haplogroup L3f2 is projected to be around 60,000 years old, based on full mtDNA data or 29ky old if we focused on HVR1 alone [22], [49], [139]. This indicates that L3f2 and its sister clade L3f1 likely originated in the Chad Basin region. However, L3f2 is also found in northern Cushitic groups from Somalia and Ethiopia, the findings strongly point to a migration event around 8,000 years ago, in which proto-Chadic

pastoralists moved from East/Northeast Africa to the Chad Basin, as suggested by the data [49], [100], [192].

Rosa & Brehm, (2011) first identified the specific set of variations that define L3h lineages, particularly L3h1b, during a survey of Guinean populations using HVS-I-coding region RFLPs. Similar HVS-I variants have been found at low frequencies in Cape Verdeans and Ethiopian Amharans, but the highest known frequency is found in populations such as the Zriba in Tunisia, Ejamat in Guinea-Bissau, and the Datoga in Tanzania [119], [131]. However, it's essential to exercise caution when assigning samples to L3h exclusively using the HVR-I motif, without considering coding region diagnostic polymorphisms in the analysis [139].

Rosa and Brehm, (2011) estimated coalescence time for the lineages within L3h is around 70 kya, as reported in studies by [139] and [125]. However, there are still several L3* lineages that have yet to be properly classified, as they remain within the L3* paragroup.

NON-L HAPLOGROUP

M1 haplogroup

The M1 lineage is primarily distributed in Northeastern and Eastern Africa, with greater diversity observed in these regions [11], [181], [182]. Occasional occurrences of M1 are also observed in Northwest and West Africa [98], [151], [164], [195]. However, the distribution of M1 is not limited to Africa, as it is relatively elevated among populations in the Mediterranean region, its highest levels in the Iberian Peninsula, with its presence extending to the Basque country [195]. Additionally, the M1 lineage has a well-documented presence of this haplogroup, encompassing territories from the Arabian Peninsula to Anatolia and from the Levant to Iran [134]. Findings from Central Asian studies have indicated that this haplogroup has been observed in regions as remote as Tibet [196]. The origin of M1 lineage has been a subject of intense debate for many years, as it is the only M representative found in Africa. The origins of the M1 lineage have sparked a contentious discussion, presenting two divergent theories. One hypothesis proposes that M1 lineages originated within East Africa and subsequently dispersed

across the continent, while the other suggests an origin outside of Africa, with a subsequent migration back into the continent referred to as "back-to-Africa [123], [156], [194].

Recent investigations have provided corroborating evidence indicating that the origins of M1's molecular ancestors can be linked to West Eurasia or the Near East, indicating that M1 lineages trace back 35-40 kya [98], [131], [149], and The coalescence age of M1 lineages is comparatively younger than that of Asian-exclusive M lineages, pointing to a more recent origin. Moreover, the research conducted by Fregel et al., (2019) reveals a significantly younger coalescence age for the entire M1 clade, estimated to be around 20,000 to 30,000 years ago (kya). The presence of the basal M1c lineage in Jordanians and its occurrence in regions as distant as Tibet supports the theory of a Near Eastern origin, with subsequent dispersals into Central Asia and westward migrations into Africa, potentially facilitated through the Sinai Peninsula.

U6 Haplogroup

Northwest Africans, particularly Algerian Berbers, Moroccans, and Mauritanians, exhibit a significant prevalence of haplogroup U6. This genetic lineage is highly frequent in the population of Northwest Africa but is also present in Eastern Africans [100], [182], [192]. Many researchers hold the belief that this haplogroup represents the first migration of ancient Caucasoid lineages back to Africa, occurring 40-50 kya [149]. A plausible scenario for the origin of haplogroups M1 and U6 is a joint diffusion from a West Eurasian/Near Eastern source, and The Early Upper Paleolithic era is thought to have marked the time when this haplogroup migrated and settled in North Africa around 30,000ya, possibly as a result of a harsh glacial period [33].

The U6 haplogroup's U6a sub-clade is particularly widespread in Northwest Africa, and its diversity and frequency increase towards this region, indicating a prehistoric autochthonous lineage around 38 kya [197]. The distribution pattern of U6b and U6c sub-clades mirrors that of M1b and M1c1, indicating a similar geographic spread, indicating a more recent local dispersal [33]. One of the most prevalent U6a lineages started

expanding around 11,000ya and partly dispersed encompassing the Sahel region [49], [119], [192].

U5 Haplogroup

The temporal and spatial overlap observed for the emergence of haplogroups M1 and U6 is also evident for U5. This suggests the possibility that the molecular ancestors of these haplogroups originated from the same geographic area in Southwest Asia, potentially in separate regional enclaves. U5 underwent its main radiation in Europe after arriving During the early Upper Paleolithic period, approximately 40,000 to 50,000 years ago, there were migratory movements of populations believed to have originated from the Middle East/Caucasus region, based on findings presented in studies by [147], [149]. A fascinating discovery has been made regarding the U5b1b branch, revealing a connection between the Saami people of Scandinavia, the Berber population of North Africa, and the Fulani ethnic group in sub-Saharan Africa. However, this association did not emerge until approximately 9,000 years ago [94].

The Mediterranean coastal regions of Africa exhibit a high prevalence of haplogroups H1, H3, V, and U5 and are believed to be post-LGM signatures [195]. The Franco-Cantabrian refuge area is considered the ancestral homeland of these lineages. This region served as a sanctuary for surviving hunter-gatherer lineages during the late-glacial period. Subsequently, these lineages dispersed into Europe and eventually contributed to the mtDNA pool of North Africans after crossing the Strait of Gibraltar [195]. While Sub-Saharan West African populations exhibit limited occurrences of specific sub-clades within the U5b haplogroup, they provide support for the hypothesis that North Africans crossed the Sahara and established commercial networks [49], [95], [155].

Other N- and R-derived Haplogroup

Haplogroup X is an ancient lineage present in West Eurasia, with origins dating back to pre-Holocene times around 30,000 years ago, soon after its divergence from superhaplogroup N, along with its sister clades W, N1, and I [20]. In West Eurasian, North African, and Near Eastern populations, haplogroup X is observed at relatively low

frequencies, but Native American populations harbor distinct sub-clades [123]. This suggests a possible diffusion of haplogroup X1 along the Mediterranean and Red Seas, particularly among Afro-Asiatic populations in North Africa, such as Moroccans and Ethiopians [100]. The coalescence age of this lineage in North Africa is comparable to that of the M1 and U6 mtDNA lineages, indicating a similar time frame for their origin in the region, which are also ancient Paleolithic lineages in North Africa. The maternal lineage of populations inhabiting the Mediterranean coast of Africa underwent several migratory events that led to an enrichment of Eurasian lineages. Research suggests that H lineages, except for H1 and H3, closely matches that of Jordanian H mtDNAs. This suggests the occurrence of parallel events in North Africa and the Middle East, followed by southward migrations resulting from the cold climate during the Last Glacial Maximum, which took place from 20 to 14 kya [20]. The X2 lineage likely emerged and spread after the glacial period, not earlier than 25 kya. The geographic scope of haplogroup X extends across Europe, the Near East (where it demonstrates greater variation), and North Africa, encompassing a broader territory than its X1 sister clade [149].

The origins of haplogroups J, T1, and R0 are firmly rooted in the Middle/Near East, as evidenced by their estimated ages and decreasing frequencies spans the southern Caucasus, Near East, Europe, and North and East Africa [20], [119], [120]. These maternal lineages spread to Europe and North Africa during the Neolithic period, with the migration of societies that adopt a mixed economy of farming and herding around 8 to 10 kya. The question of how much the Neolithic culture from the Near East impacted the native Capsian Neolithic culture in Northwest Africa is still being debated [198]. Additionally, it is possible that these lineages were introduced to North Africa through more recent events, such as the migration of Phoenician traders [182]. The N1 mitochondrial DNA haplogroup has been sporadically observed in different populations, albeit at low levels across Europe, the Near East, India, and East Africa, particularly in Semitic speakers [100]. While the coalescence of this haplogroup is considerably ancient in the Near East and Southwest Asia, there is speculation that the arrival of N1, as well as

U (excluding U5 or U6) and W lineages, occurred relatively recently with the spread of Semitic languages. This may be particularly true for the Ethiopian population [156].

1.2.13 NIGERIA: GEOGRAPHY, DEMOGRAPHY AND ETHNIC POPULATIONS

Resting in West Africa with its coastline adjoining the Gulf of Guinea, Nigeria stretches across a substantial land area of 923,768 square kilometers. Accommodating over 200 million inhabitants, it claims the title of Africa's most populous country and stands as the seventh most populous nation on the planet. Nigeria is known for its diverse geography, demography, and major ethnic populations[164].

Geography: Nigeria has a diverse geography, the southern part is a blessed with tropical rainforests and the northern part savannah grasslands. The country is segmented into 36 states and one Federal Capital Territory, which is home to the capital city of Abuja [17]. The country's coastline stretches for over 850 kilometers and is home to many sandy beaches and lagoons. The Niger River, which is the third-longest river in Africa, flows through Nigeria and empties into the Gulf of Guinea. The country also has many natural resources, including oil, gas, coal, and minerals such as tin, iron ore, and gold[199].

Demography: Nigeria has a large and growing population, with an estimated population of over 200 million people. The demographic makeup of Nigeria encompasses a rich tapestry of various ethnic groups, each with its own language, customs, and heritage. Among these groups, the Hausa-Fulani, Yoruba, and Igbo emerge as the three predominant ethnicities in the country [200]. Other notable ethnic groups include the Kanuri, Tiv, and Ijaw. Christianity and Islam are the two dominant religions in Nigeria, with approximately equal numbers of adherents [199].

Major Ethnic Populations: The Hausa-Fulani is the largest ethnic group in Nigeria, accounting for about 29% of the population. They are predominantly Muslims and are concentrated in the northern part of the country. The Hausa-Fulani are known for their distinctive dress, music, and food. The Yoruba is the second-largest ethnic group in Nigeria, accounting for about 21% of the population. The majority of individuals

belonging to these ethnic groups in Nigeria adhere to the Christian faith, with a particular concentration of Christians in the southwestern region of the country. They are known for their rich cultural heritage, including music, art, and festivals such as the Osun Osogbo festival. The Igbo is the third-largest ethnic group in Nigeria, accounting for about 18% of the population. They are predominantly Christians and are concentrated in the southeastern part of the country. The Igbo are known for their entrepreneurial spirit and are involved in various economic activities, including trade and manufacturing[199], [201], [202].

Other notable ethnic groups in Nigeria include the Kanuri, who are predominantly Muslims and are concentrated in the northeastern part of the country; the Tiv, who are predominantly Christians and are concentrated in the central part of the country; and the Ijaw, who are predominantly Christians and are concentrated in the southern part of the country [199].

In conclusion, Nigeria is a country with a diverse geography, demography, and major ethnic populations. It has a large and growing population, with over 200 million people. The country is known for its many ethnic groups, each with its own language, culture, and traditions, and its diverse geography, which includes tropical rainforests, savannah grasslands, and a long coastline. Nigeria is a country with over 250 ethnic groups, but three major ethnic groups dominate the population each having their distinct cultural, linguistic, and genetic characteristics.

Mitochondrial DNA is an important genetic marker for studying human population genetics and evolutionary history. Studies have shown that there is high genetic diversity among African populations, including those in Nigeria. In terms of mtDNA genetic variation, studies have been carried out on the three major ethnic groups in Nigeria. The results of these studies have shown that there are significant differences in mtDNA genetic variation among these ethnic groups [17].

The Hausa-Fulani ethnic group is predominantly found in Northern Nigeria, and they are the largest ethnic group in the country. There are reported literatures on the Hausa-Fulani

having greater number of haplogroups L1c, L2a, and L3b[40], [41]. These haplogroups are common among African populations and are associated with the Bantu expansion. The Hausa-Fulani also have a low frequency of haplogroups M and N, which are common among East and North African populations[200].

Yoruba: The Yoruba ethnic group is predominantly found in Southwest Nigeria, and they are the second-largest ethnic group in the country. There are reported literatures on the Yoruba having greater number of haplogroups L1b, L2b, and L3e [203]. These haplogroups are common among West and Central African populations. The Yoruba also have a moderate frequency of haplogroups H and U, which are common among Eurasian populations [151].

Igbo: The Igbo ethnic group is predominantly found in Southeast Nigeria, and they are the third-largest ethnic group in the country. There are reported literatures on the Igbo having greater number of haplogroups L0a, L1c, and L3d [203]. These haplogroups are common among West and Central African populations. The Igbo also have a low frequency of haplogroups M and N, which are common among East and North African populations[164].

The mtDNA genetic variation of the three major ethnic groups in Nigeria reflects their distinct genetic histories and geographic locations. The Hausa-Fulani have a higher frequency of haplogroups associated with the Bantu expansion, while the Yoruba have a higher frequency of haplogroups associated with West and Central Africa. The Igbo, on the other hand, have a higher frequency of haplogroups associated with West and Central Africa, with some unique haplogroups.

CHAPTER TWO

2.0 MATERIALS AND METHODS

This chapter will provide comprehensive **reproducible** details of the materials used and the methodologies employed. This chapter will display some of the equipment and consumables used. It will entail the sample population, sample size and geography of the study population. It will also present the methodologies employed during the buccal swab collection, storage and DNA extraction technique employed. It will provide details of the spectrophotometric technique used to check the integrity of the extracted DNA, it will then further dissect the process of mtDNA amplification via Polymerase Chain Reaction (PCR) and the primers used in the process. Furtherance to that, the process employed in integrity and size estimation of the extracted mtDNA through Gel Electrophoresis. The technique through which the estimated mtDNA is sequenced via Big Dye Terminator will also be provided in detail for reproducibility.

The chapter will also provide the details of the data analysis tools used and the technical details of each analysis. This section will provide the details of the software used to conduct base calling, file conversion from fasta format to other formats such as Arlequin, structure, phylip, and nexus files. It will then provide the details of software used in generating Neighbour-Joining Tree, Sequence Demarcation Tool, Haplotype Network and population admixture among other techniques. The haplotype and haplogroup assigning tools will also be provided in the chapter.

2.1 METHODOLOGY

The research was carried out in Daura Local Government in Nigeria, the city holds historical importance to the Hausa population globally. The research was carried out by collecting buccal swab from consented individuals. DNA was extracted, the quality and integrity were checked prior to amplification. Other downstream molecular genetic techniques were employed as detailed in the preceding sections. Bioinformatics tools were used to study and analyse the mtDNA dataset generated.

2.1.1 SAMPLING

The total 100 samples were collected from informed and consented Hausa ethnic populations living in four towns (n=25 from each town) from three Local Government Areas (LGAs) of the ancient and historic Daura emirate, Nigeria. The sample collection locations are shown with blue geotag icon in figure 2.1. The towns selected represent ancient towns of the emirate whom according to the locals have been in existence for centuries. Samples from Daura town of Daura LGA are termed as Group 1, the town is located on 13^0 00'27"N:$8^0$20'25"E. Group 2 is represented by individuals from Koza (Tsohon Birni; meaning ancient city) of Mai'adua LGA located on $13^0$06'04"N:$8^0$17'46"E, while groups 3 and 4 were assigned to individuals from Sandamu ($12^0$57'31"N:$8^0$22'32"E) and Rijiyar Tsamiya ($12^0$55'02"N:$8^0$17'44"E) towns of Sandamu LGA respectively.

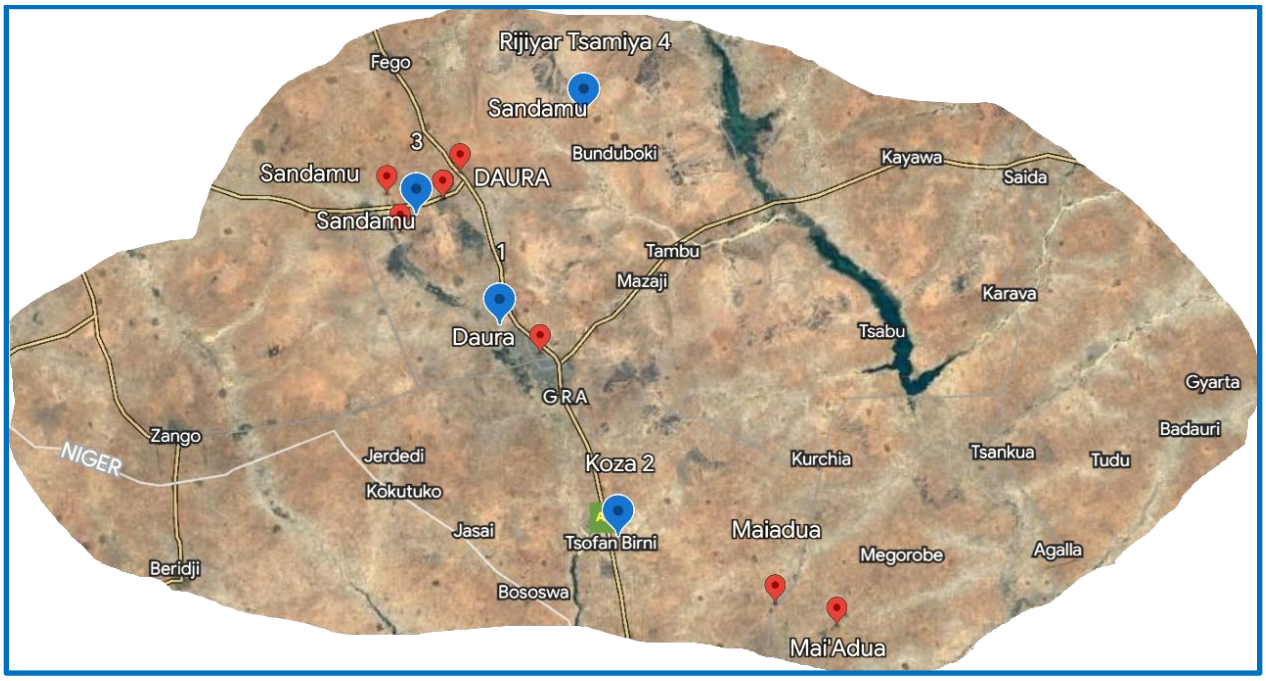


Figure 2.1: Google earth map of the sample collection villages indicated with blue geotag

The figure above shows the map of the areas where the samples were collected.

The maximum number fixed for each cluster was 25 in concordance with the guidelines set by SWGDAM, the guidelines state that the maximum number for a reliable data for comparative study and the participation and successful profiling of samples are contingent upon the willingness of individuals to take part.

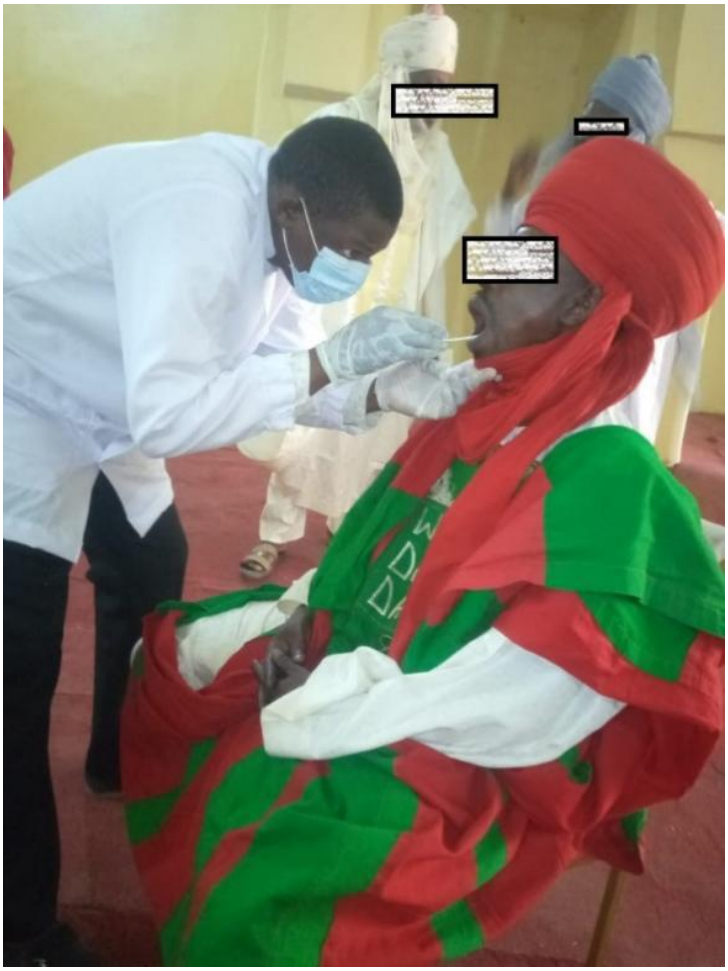


Figure 2.2: buccal swab collection

2.1.2 DNA SAMPLE COLLECTION

Buccal swab samples were collected using a sterile swab stick. The study participants that didn't consume food for at least an hour prior to the collection were used in the study. The swab was collected by gently scratching both the left and the right cheeks with separate sterile swab stick, this was done for about 20 seconds and allowed to air-dry

before inserting in the swab collecting tube. The collected swabs were immediately stored in cold storage system.

2.1.3 DNA EXTRACTION

The whole genomic DNA (gDNA) including the mtDNA was extracted using QIAamp DNA mini-kit (www.qiagen.com/HB0329; QIAamp cat. no. 51304, QIAGEN Heidelberg Germany) and Zymo Research gDNA MiniPrep Kit Cat. no. D4068 (Zymo Research, USA). The kits employ spin column extraction process; the process starts with lysis, then binding to the column, then washing and finally elusion.

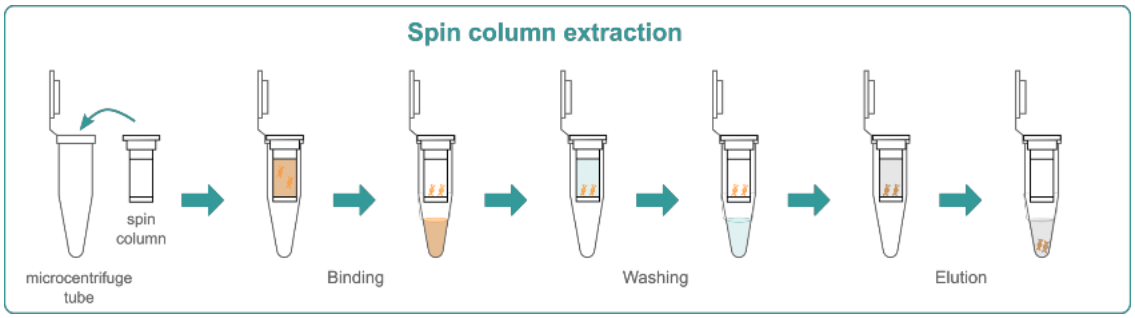


Figure 2.3: DNA Extraction Flow Chart

The manufacturer's recommendations were followed when using the QIAamp DNA mini-kit which requires specific protocol for buccal swab as follows:

NOTE: A 56^0C heating block was set in preparation for incubation stage. The whole process was observed at room temperature (15-25^0C). Additional AL buffer was required to extract swab.

Figure2.4: Vortexing Machine

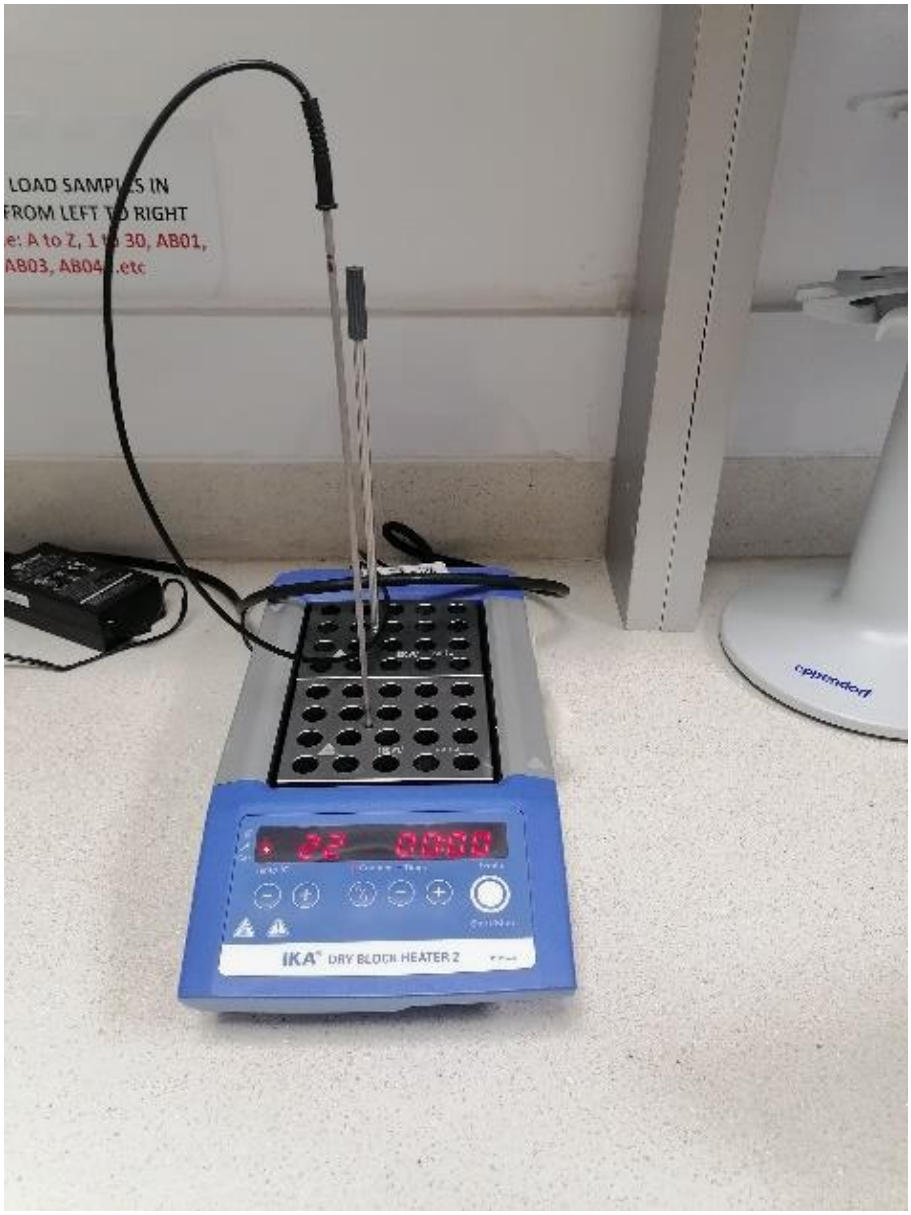


Figure 2.5: The Heating Block

Lysis stage: The cotton swab was gently scratched against the wall of the swab container containing phosphate-buffered Saline (PBS). 400µl of the swab sample was placed in 2ml centrifuge tube. 20ul of proteinase K was then added, 400µl of AL buffer was also added, the stock was then vortexed immediately for 15secs. Vortexing was done with the aid of IKA Vortex 2 S0000 (fig. 2.4) model. Precaution was taken not to add proteinase K directly into buffer AL as recommended by the manufacturer. The 2ml centrifuge tube

containing the stock was then incubated at 56^0 for 10mins using IKA dry block heater 2 (fig. 2.5). The sample was then briefly centrifuged for 15secs at 8,000rpm which aided the removal of drops from inside the lid caused by the heating process.

Binding: A total of 400ul of pure ethanol was introduced and vortexed for 15secs. The concoction was then briefly centrifuged to aid removal of drops from the lid. The centrifuge was achieved using Eppendorf Centrifuge Machine 5420 model (fig. 2.6).

Figure 2.6: Eppendorf Centrifuge Machine 5420

700µl of the mixture above was then pipetted using Eppendorf pipette, the content was then emptied into QIAamp minispin Column without wetting the rim. (provided by the manufacturer). The cap was then closed and centrifuged for 1 minute at 8,000rpm. The

flow through was discarded, while the spin column was placed in a clean sterile 2ml collection tube. This step was repeated by pipetting the remaining 700µl and centrifuged for 1min at 8,000rpm.

Washing: the QIAamp minispin column was carefully opened and 500µl of AW1 was added without wetting the rim. After closing the cap, the sample was subjected to centrifugation at 8,000rpm for 1 minute, and the resulting liquid was discarded. In a clean 2ml collection tube, we introduced 500µl of AW2 into the spin column, which was then subjected to centrifugation at 14,000rpm for 3 minutes. Afterward, the filtrate was discarded, we improvised by empty spinning the spin column in empty collection tube at 14,000rpm for 2mins.

Elution: the spin column was placed in 1.5 micro-centrifuge tube, 100µl of The AE buffer was included, and a 1-minute incubation at room temperature followed. it was then centrifuged at 8,000rpm for 1 min. we also improvised by adding the filtrate back into the column and repeating the elution process above in order to elute more DNA. The centrifuge tube was closed and the eluate (extracted DNA) was then taken for spectrophotometry using Nanodrop (fig. 2.7).

Figure 2.7: Nanodrop One v3.7

2.1.4 DETERMINATION OF NUCLEIC ACID CONCENTRATION AND PURITY USING NANODROP

Spectrophotometry was conducted using thermo scientific Nanodrop One v3.7 (Thermofischer Scientific). Before measuring the concentrations, blanking operations were performed using wash buffer or distilled water. 1µl of wash buffer was loaded on the Nanodrop sensor, the arm was lowered and blank option was pressed on the display to measure the blanking solution as if it were a sample. After the blanking, the sampling arm was opened and the sensor was wiped, the sample was then pipetted upon the lower pedestal meant for measurement (sensor). Following the lowering of the lower sampling

arm the measurement via the spectrum was initiated. The measurement read was then stored and the sampling arm was raised, the lower pedestal wiped with lint-free laboratory wipe, and another sample loaded in the same manner as described above without repeating the blanking. Through a straightforward wiping action, the risk of sample carryover in subsequent measurements is mitigated, especially when dealing with samples that differ significantly in concentration (more than 1000-fold).

2.1.5 AMPLIFICATION OF THE TARGETED MTDNA USING PCR.

Kary Mullis pioneered the PCR technique in 1984, marking a significant breakthrough in molecular biology, it involves the use of enzymes to target and multiply copies of nucleic acid with aid of other reagents such as primers, magnesium chloride ions, and deoxynucloetide triphosphates (dNTPs) as substrates. Within the scope of this research, the study focused on analysing specific regions of the mitogenome, namely HVS-I and HVS-II, which encompassed positions 16,024 to 16,482 and 21 to 413 [49] respectively were the target DNA molecule for the amplification. The final reaction volume of 25µl was used in the following composition (table2.1).

Table2.1: Summary of the PCR Reaction Volumes and Concentrations

Component	**25µl Reaction**	**Final Concentration**
Forward primer(10Um)	0.5µl	0.2 µM
Reverse primer (10Um)	0.5µl	0.2µM
PCR Master mix	12.5µl	1x
Template DNA	2µl	
Nuclease-free water	9.5µl	
Final volume	25µl	

The PCR master mix was ordered from New England Biolabs with model number M0486S. The HVS-I region was amplified using specific primers, with the forward primer sequence as F-5'-TTA ACT CCA CCA TTA GCA CC-3' and the reverse primer sequence as R-5'-CCT GAA GTA GGA ACC AGA TG-3'. For the HVS-II region, the forward primer sequence was F-5' GGT CTA TCA CCC TAT TAA CCAC3' and the reverse primer sequence was R-5' CTG TTA AAA GTG CAT ACC GCCA3' [26]. They were designed by Inqaba Biotec West Africa Limited. The human reference genome was compared to the query using the Basic Local Alignment Search Tool (BLAST) provided by the National Centre for Biotechnology Information (NCBI) and the primer sequences were assembled to rule out the possibility of annealing with segments of nuclear genome. For PCR amplification, the following conditions were used: 35 cycles of initial denaturation at 94°C for 5 minutes, followed by denaturation at 94°C for 30 seconds, annealing at 53°C for 30 seconds, extension at 72°C for 1 minute, and a final extension at 72°C for 7 minutes and held at 4°C. The reaction volume was 25µl, comprising 12.5 µl of One Taq Quick-Load 2X Master mix with standard buffer (obtained from New England Biolabs Inc., www.neb.com/M0486), 0.5 µl each of the forward and reverse primers, 2µl of diluted DNA template, and 9.5µl of nuclease-free water. Quality control was ensured by incorporating negative control in each PCR run. To ensure clean and accurate sequencing results, Exonuclease I and Shrimp Alkaline Phosphatase were used to remove any unused primers and dNTPs from the PCR products before initiating the sequencing process (ExoSAP, Biochemical, USA).

Figure 2.8: PCR Setup Displaying the Samples Loaded Before PCR Run and During Running PCR

2.1.6 SIZE ESTIMATION AND INTEGRITY CHECK VIA GEL ELECTROPHORESIS

The size and integrity of the extracted mtDNA was checked via agarose gel electrophoresis. This technique is useful in separating the extracted molecules based on size and net charge. One percent agarose gel formation involved dissolving 1g of agarose in 2ml of Tris Acetate EDTA (TAE) buffer, and 98ml nuclease-free water. The solution was allowed to dissolve by incubating in microwave oven at medium temperature for 5mins. The solution was removed and 5ul safeView Fire Red with cat. no. G926 (abmGood.com, Canada) was added to enable UV capture by the gel documentation system.

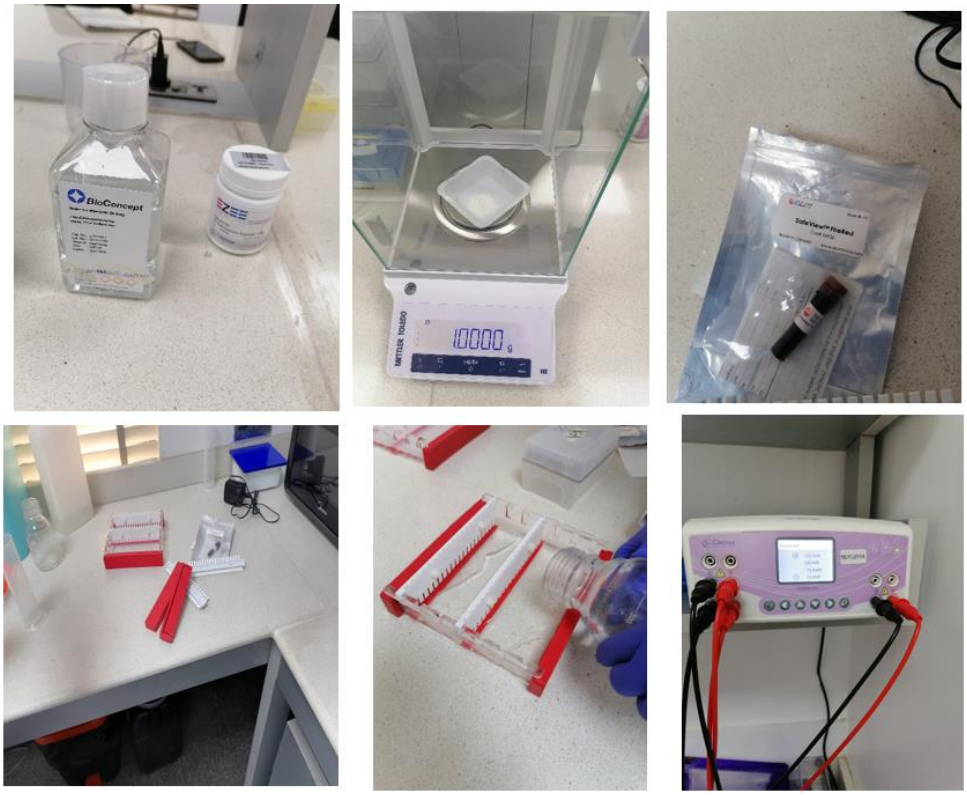


Figure 2.9: Gel Electrophoresis Setup and Workflow

It was mixed properly and allowed to cool (about 60^{o}C), the solution was then poured into the tray with casting dams fit on both ends of the tray and combs in position. This was done carefully to avoid formation air bubbles, it then allowed to solidify. The combs and the casting dams were gently removed after the gel solidified. The gel tray was then placed in gel tank filled with TAE buffer. 1kb DNA ladder with cat. no. N0552S (New England Biolabs) and the samples were loaded into the wells. The tank was covered with a lid, and the electrodes were connected to the power supply. The gel was run using Cleaver Scientific powerPro. The machine was set at 100volt, 138mA, 14watt for 60mins.

Figure 2.10: Gel Documentation System Loading and Image Capture

Sanger Type Sequencing (STS)/ Big Dye Terminator

Big dye terminator Sanger Type Sequencing was conducted to determine the sequence of the HVR I and II of the mtDNA. Adherent to the principle of Sanger Sequencing, capillary electrophoresis was done for separation and detection of fragments using ABI prism 3500xl genetic analyser (Applied Biosystems, USA). Of the 100 collected swabs, only 93 and 94 successful sequences were achieved for primers 1 and 2 respectively.

Figure 2.11: Genetic Analyser 3500xl

2.2 DATA ANALYSIS

After successful completion of the sequencing, the data generated by the genetic analyser was saved in fasta file format. The file needed to be converted to other formats for different types of analysis. Nexus file format was created using Mesquit v 3.7.0, while arlequin, structure, and phylip files were converted from fasta haploid data format, for the analysis, the R program's ape package (accessible at http://cran.r-project.org/package=ape) was employed [204]. However, prior to the conversion, Bioedit v1.3.7 was used to do the base calling which involved the removal of ambiguous bases such as K, R, W by assigning the appropriate bases (ACGT) at the respective positions of the ambiguous bases, base “R” can be seen at position 66 in figure 2.12.

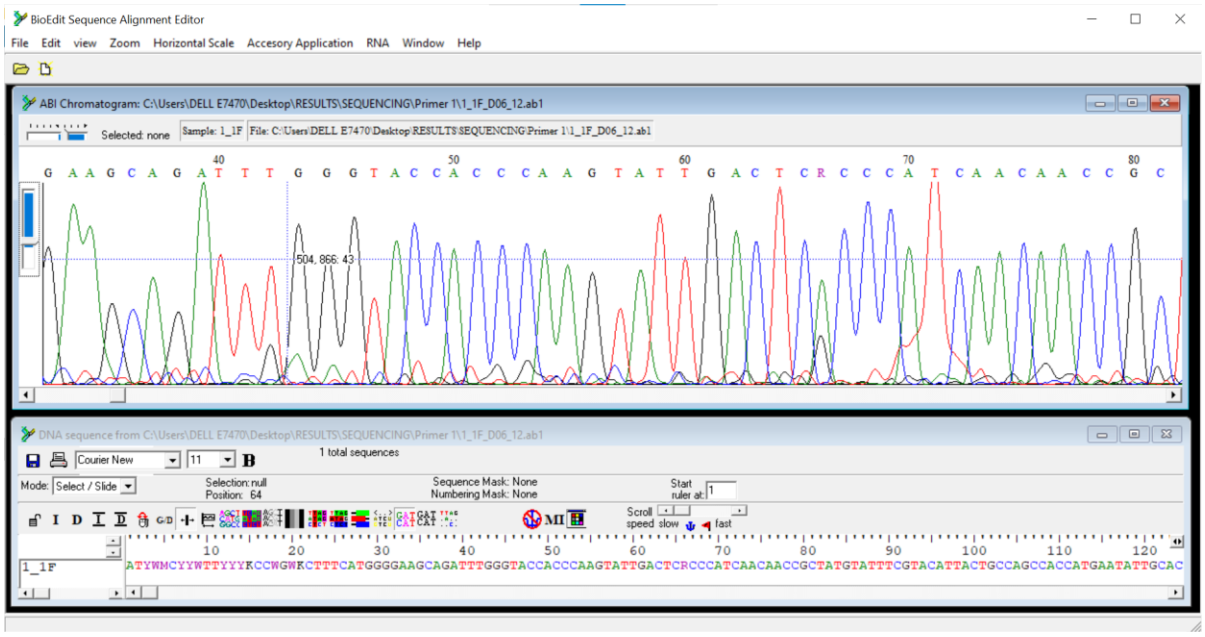


Figure 2.12: The BioEdit Interface Showing the Ambiguous and Rightly Captured Bases

The study's genetic data was submitted to GenBank with accession numbers OQ388355-OQ388447 for HVS-I. Haplogroups were determined using online tools such as mitotool and haplotracker. These tools compare the test samples to the revised Cambridge Reference Sequence (rCRS) and PhyloTree build 17.0 to assign haplogroups. Consensus haplogroups were taken when both tools indicated the same haplogroup. Manual confirmation using Bioedit v.7.2.5 and comparison with rCRS was used to resolve ambiguous haplogroups.

Microsoft Excel 2010 was used to manually count the shared haplotype, Most Recent Common Ancestor (MRCA), and macrohaplogroups. Multiple alignments were conducted using Bioedit v.7.2.5, and ClustalW was used in the multiple alignments with a bootstrap value of 1000 [205]. The resultant files were saved in fasta format before conversion to other file formats for downstream analysis.

Using Molecular Evolutionary Genetics Analysis v. 11, a Neighbour-joining tree was built (MEGA 11; Tamura et al., 2021) with the Jukes-Cantor model and 1000 bootstrap replications. Locus-by-locus Analysis of Molecular Variance (AMOVA), pairwise

genetic distance, and haplotype diversity were calculated using Arlequin v.3.5.2.2 [207]. To conduct additional analysis on the haplotype network, we employed Population Analysis with Reticulate Trees (PopART) software, which can be found at http://popart.otago.ac.nz. A minimum spanning network was constructed to further elucidate the ancestral relationship among the population data set [115], [208]. To compute individual-level pairwise identity scores, our analysis involved the utilisation of the Sequence Demarcation Tool (SDT) v1.2 [209]. To estimate genetic genealogies from DNA sequences and measure statistical parsimony, we employed TCS software release 1.21 [210] for the construction of the haplotype network and subsequent visualisation with the aid of tcs beautifier (tcsBU) [211]. The ancestral relationships of a population were analysed through the study of population admixture using Structure v.2.3.4. The population was divided into four groups, as previously noted. To perform the analysis, we ran 18 sets of iterations, each with a Markov Chain length of 10,000 and a burn-in period of 10,000. Six different K replication values were used with three iterations. To determine the optimal run for inferring ancestral relationships among the study populations, a structure harvester was employed [107].

CHAPTER THREE

3.1 INTRODUCTION

Forensic human identification relies on establishing an individual's unique genetic profile. This profile, also known as a genetic fingerprint, is a phenotypic description of genomic loci specific to an individual. The only genetic fingerprint that is admissible in forensic cases, as recommended by the European DNA Profiling Group (EDNAP), is the one obtained from autosomal Short Tandem Repeats (STRs) [212]. In cases where the nuclear DNA is highly degraded or unavailable, By studying mitochondrial DNA (mtDNA), researchers can establish the genetic profile of the individual(s) under investigation [213]. Mitochondrial sequence analysis is typically performed on samples that are unlikely to yield successful results for nuclear DNA, such as teeth, bones, hair shafts, and nuclear DNA that is severely degraded or significantly environmentally exposed [58]. This is possible because of the high copy number of mtDNA per cell, which provides abundant amplifiable mtDNA in samples with trace amounts of amplifiable nuclear DNA [29].

There is a growing interest in using mtDNA diversity for phylogenetic analysis as a means of unravelling the history of modern humans. This is due to the unparalleled genetic diversity resolution provided by mtDNA, which is aided by the high degree of conservation of the mtDNA non-coding hypervariable region and the resultant evolutionary rate of the mitogenome [42]. While there is a relatively well-known mtDNA diversity of the African population, some regions and ethnic groups within Africa have not been sufficiently sampled. The L-type haplogroup is the phylogenetically classified sub-Saharan African haplogroup [214].

Africa is known to have the highest level of genetic diversity among populations and is believed to be the ancestral home of modern humans. When compared with non-African populations, Among all populations, African populations demonstrate the highest frequency of mitochondrial, X-chromosomal, and autosomal haplotypes that are specific to their population [215]. The pairwise genetic distance estimation of mtDNA (measured

by FST), which is a classic index of population diversity, shows that African populations have significantly higher values than other global populations [164]. Nigeria is a West African country situated in the Gulf of Guinea. It is the most populous country on the African continent and boasts over 500 distinct ethnic groups, which has led to high levels of language and cultural diversity [203]. The ethnic composition of Nigeria is characterised by the dominance of three key groups: the Hausa, Igbo, and Yoruba, with the Hausa being the most populous. Hausa is a member of the Chadic branch of the Afro-Asiatic family and is spoken by over 50 million people as their first language. The language has spread to almost all major cities in Central, North, West, and Northeast Africa [216].

The origin of the Hausa language is largely unknown, but there is a legendary view that Bayajidda was the origin of the Hausa kingdom [201]. According to legend, Bayajidda was a prince from Baghdad who emigrated from the Middle East to Nigeria and was later betrothed to a queen named Queen Daurama. Her maternal lineage ruled the Daura Kingdom for centuries, and their union led to the birth of 7 legitimate (with the queen) and 7 illegitimate (with concubines) sons who are believed to be the rulers of present-day Hausa Kingdoms in Northwestern Nigeria [217].

Despite the rich genetic diversity of African populations, very little African population genetic data is publicly available. Furthermore, the undocumented history or rather myth of Bayajidda has been the driving force in studying the matrilineal genetic diversity among the indigenous Daura emirate population. This study will contribute to our knowledge of population structure and demographic history by analysing mtDNA data, and it will provide a valuable genetic database for forensic and human identification purposes, as well as population genetic reference. This research aims to conduct a comprehensive analysis of the HVS-I within the D loop control region of mtDNA sequences from the Hausa ethnic population in the Daura emirate of Nigeria, and draw conclusions about the population's phylogenetic and forensic traits.

3.2 METHODOLOGY

3.2.1 ETHICS AND CONSENT STATEMENT

Before collecting samples, individuals provided written and informed consent. Additionally, the study obtained ethical clearance from the Research and Ethics Committee of the Katsina State Ministry of Health in Nigeria reviewed and approved this study, referencing number MOH/ADM/SUB/1152/1/558. Furthermore, the consent of traditional rulers in all areas of study was also obtained.

3.2.2 POPULATION AND SAMPLES

Consenting individuals from four different locations in Daura emirate, Nigeria, provided biological samples (buccal swab). Samples were collected from three Local Government Areas (LGAs) that are believed to have been ancient cities of the emirate, that existed for centuries. A total of 100 individuals were sampled, out of which 93 samples were successfully sequenced. The sample size for each location was as follows: Daura (n=22) from Daura LGA, Koza (n=24) from Mai'adua LGA, Sandamu (n=24) and Rijiyar Tsamiya (n=23) from Sandamu LGA. Individuals with known maternal ancestry were excluded from the study.

3.2.3 LABORATORY METHODS

To collect the buccal swab, the swab stick was gently rubbed on both cheeks for 20 seconds and then stored in a collection tube with 1ml of phosphate buffered saline (PBS) on ice. The manufacturer's recommended protocol for the QIAamp DNA mini kit (www.qiagen.com/HB0329; QIAamp cat. no. 51304, QIAGEN Heidelberg Germany) was used to extract genomic DNA, including mtDNA. The Nanodrop was utilised to examine the concentration and purity of the DNA and validate its quality. After verifying the integrity of the DNA, the forward primer (5'-TTA ACT CCA CCA TTA GCA CC-3') and reverse primer (5'-CCT GAA GTA GGA ACC AGA TG-3') were utilised in the Polymerase Chain Reaction (PCR) to amplify the HVS-I region [175]. To ensure the primers did not anneal with segments of nuclear genome, the primer sequences were

checked against the human reference genome using the Basic Local Alignment Search Tool (BLAST) provided by the National Centre for Biotechnology Information (NCBI).

The PCR amplification procedure began by denaturing the sample at 94℃ for 5 minutes. Following this, a total of 35 cycles were performed, including denaturation at 94°C for 30 seconds, annealing at 53°C for 30 seconds, and extension at 72°C for 1 minute. A final extension step was conducted at 72°C for 7 minutes, and the reaction was then held at 40°C. The reaction mixture, with a volume of 25µl, consisted of 12.5 µl of One Taq Quick-Load 2X Master Mix containing standard buffer from New England Biolabs Inc., 0.5 µl each of the forward and reverse primers, 2µl of diluted DNA template, and 9.5µl of nuclease-free water. Negative controls were included in each PCR run to ensure quality control.

Prior to sequencing, the PCR products were carefully prepared by removing unbound primers and dNTPs through treatment with Exonuclease I and Shrimp Alkaline Phosphatase. This step ensured the purity and integrity of the DNA samples for accurate sequencing analysis (ExoSAP, Biochemical, United States). The expected amplicon sizes were confirmed by running agarose gel electrophoresis on all the amplicons in a 1% agarose gel. Once the amplicon size was confirmed, the forward strand was sequenced using big dye terminator, and capillary electrophoresis was performed for fragment separation and detection on an ABI prism 3500xl genetic analyser (Applied Biosystems, United States).

3.2.4 DATA ANALYSIS

The genetic data obtained from the study were deposited in GenBank with accession numbers OQ388355-OQ388447. Haplogroups were assigned using online tools, namely mitotool (www.mitotool.org/cgi/dlooprCRS.pl) and haplotracker (www.haplotracker.cau.ac.kr), which compare the test samples with the revised Cambridge Reference Sequence (rCRS) using PhyloTree build 17.0 [69]. The consensus haplogroup was determined when both tools showed the same haplogroup. In cases

where ambiguous haplogroups were presented, haplotracker and manual confirmation through comparison with rCRS using Bioedit v.7.2.5 were used as standards [218].

Shared haplotype, Most Recent Common Ancestor (MRCA), and macrohaplogroups were manually counted with the aid of Microsoft Excel 2010. Base calling and multiple alignment were performed using Bioedit v.7.2.5 [205], with ClustalW used for multiple alignment, and the bootstrap value set at 1000 [219]. The files were saved in fasta format. A nexus file was created from the fasta format using Mesquit v 3.7.0, while arlequin, structure, and phylip files were converted from fasta haploid data format using the Analysis of Phylogenetic and Evolution (ape) package of the R program (http://cran.r-project.org/package=ape) [204]. Using the MEGA v.11 software, we employed the Neighbour-joining method to construct a phylogenetic tree. The Jukes-Cantor model was used, and 1000 bootstrap replications were performed for reliability [220]. We also performed a Locus-by-locus Analysis of Molecular Variance (AMOVA), pairwise genetic distance and haplotype diversity calculations using Arlequin v.3.5.2.2. Additionally, The haplotype network was visualized using the Population Analysis with Reticulate Trees (PopART) software, accessible at http://popart.otago.ac.nz, and a minimum spanning network was constructed to provide further insight into the ancestral relationships among the population dataset [105].

3.3 RESULTS

3.3.1 HAPLOGROUPING

The haplogroup of each sample was determined using online tools: mitotool and haplotracker. Ambiguous results were resolved by manual inspection and comparison with the rCRS. The majority of the study participants exhibited the African L macrohaplogroup, with the L3 subclade being the most prevalent. Some samples also showed subclades other than L, such as G, M, and U macrohaplogroups. Out of the 93 individuals tested, a total of 67 haplotypes were identified, with 52 being unique and the remaining 15 being shared haplotypes. The most commonly shared haplotypes were

L3b1a1a, which was shared by five individuals, followed by L3d1a, L3e2b1a, and L4b2b, which were each shared by four individuals (Table 3.1).

Table3.1: Distribution, Frequency and Percentages of Haplotype among the Study Population

SN	Haplotype	Sample ID	Frequency	Percentage
1	G2a1c1	Koza19	1	1.1%
2	G2a1g	Rijtsam17	1	1.1%
3	G3a2a	Koza20	1	1.1%
4	L0a1a+200	Daura19	1	1.1%
5	L0a1a1	Sandamu5,12	2	2.2%
6	L0a1a2	Daura21	1	1.1%
7	L0a1b2	Koza4	1	1.1%
8	L0a'g	Rijtsam7	1	1.1%
9	L1b1a4a	Daura1, Koza11	2	2.2%
10	L1b1a9	Koza16,22	2	2.2%
11	L1b1a15	Rijtsam11	1	1.1%
12	L1c2b1b	Koza3	1	1.1%
13	L1c2b1c	Koza14	1	1.1%
14	L1c3a1a	Rijtsam1	1	1.1%
15	L1c3b1b	Daura18	1	1.1%
16	L1c5	Koza15	1	1.1%
17	L2a1a2a1a	Sandamu8	1	1.1%
18	L2a1c3b2	Daura10, Sandamu10,20	3	3.2%
19	L2a1c4a1	Daura25	1	1.1%
20	L2a1c5	Daura11	1	1.1%
21	L2a1d1	Daura3, Rijtsam21	2	2.2%
22	L2a1d2	Daura6	1	1.1%
23	L2a1i1	Sandamu18	1	1.1%

SN	Haplotype	Sample ID	Frequency	Percentages
24	L2a1j	Koza1,12	2	2.2%
25	L2b1a2	Koza2	1	1.1%
26	L2b1a4	Koza5	1	1.1%
27	L2b2	Daura14	1	1.1%
28	L2c1a	Daura9, Daura13	2	2.2%
29	L2e1a	Sandamu1	1	1.1%
30	L3b1a+@16124	Rijtsam8	1	1.1%
31	L3b1a1a	Koza13,Sandamu4,11,24,RijTsam19	5	5.4%
32	L3b1a3	Sandamu15,16	2	2.2%
33	L3b1a7a	Sandamu9	1	1.1%
34	L3b1b	Rijtsam13,15	2	2.2%
35	L3d1a	Daura16, Rijtsam9,12,23	4	4.3%
36	L3d1a1b	Sandamu23	1	1.1%
37	L3d1b1a	Daura22	1	1.1%
38	L3d1b2	Rijtsam10	1	1.1%
39	L3d1b3	Koza18	1	1.1%
40	L3d1d	Koza6,7,Sandamu6	3	3.2%
41	L3d3a1a	Koza23	1	1.1%
42	L3e1a3a	Koza9	1	1.1%
43	L3e1a3b	Sandamu2	1	1.1%
44	L3e1b1	Daura23	1	1.1%
45	L3e1b2	Koza17	1	1.1%
46	L3e1f	Daura15	1	1.1%
47	L3e1f2	Koza8	1	1.1%
48	L3e1g	Sandamu14	1	1.1%
49	L3e2a1	Koza24	1	1.1%
50	L3e2a1a	Koza21	1	1.1%
51	L3e2a1b2	Rijtsam14	1	1.1%

SN	Haplotype	Sample ID	Frequency	Percentages
52	L3e2b1a	Daura24,Sandamu3,17,21,RijTsam2	5	5.4%
53	L3e2b1a1	Rijtsam2	1	1.1%
54	L3e3b1	Rijtsam16	1	1.1%
55	L3f1b+16292	Rijtsam6	1	1.1%
56	L3f1b1a1	Daura12	1	1.1%
57	L3f1b4a1	Rijtsam18	1	1.1%
58	L3f2b	Daura5	1	1.1%
59	L4b2b	Rijtsam4,5,20,22	4	4.3%
60	L4b2b1	Koza10,Sandamu19	2	2.2%
61	M1a1a1	Rijtsam3	1	1.1%
62	M30+16234	Sandamu13	1	1.1%
63	M39b	Daura17	1	1.1%
64	U1a1d	Daura20	1	1.1%
65	U5b1b1b	Sandamu7	1	1.1%
66	U6a1a1	Daura4	1	1.1%
			93	

In this study, a total of 13 macrohaplogroups were identified, with the highest observed macrohaplogroups being L3, L2, and L1, comprising 43.9%, 19.7%, and 12.1% of the total population, respectively, accounting for 75.7% of the total population. The haplotracker analysis revealed 41 Most Recent Common Ancestor (MRCA), of which 21 were unique. The most shared MRCA was L3, which was observed 15 times, accounting for 16.1% of the total MRCA in the study population. Other L3 subclades, including L3e2, L3e2b, L3f, L3e1a, and L3d, accounted for 30.5% of the MRCA in the population studied (Table 3.2).

Table 3.2: Distribution, Frequency and Percentages of Observed MRCAs

SN	sample ID	MRCA	Frequency	Percentage
1	Koza19	G2a1c1	1	1.1%
2	Koza20	G3a2a	1	1.1%
3	Rijtsam7	L0	1	1.1%
4	Daura19,21,Sandamu5,12	L0a1'4	4	4.3%
5	Koza4	L0a1b	1	1.1%
6	Koza14	L1'2'3'4'5'6	1	1.1%
7	Daura1,koza11,16,22,RijTsam11	L1b	5	5.4%
8	Koza3	L1c2b	1	1.1%
9	Rijtsam1	L1c3a	1	1.1%
10	Daura18	L1c3b1	1	1.1%
11	Koza15	L1c5	1	1.1%
12	Koza21,Rijtsam17	L2'3'4'6	2	2.2%
13	Daura6,Sandamu20	L2a1'2'3'4	2	2.2%
14	Daura11,25,Koza1,12	L2a1	4	4.3%
15	Sandamu8	L2a1a2	1	1.1%
16	Daura10,Sandamu10	L2a1c3b2	2	2.2%
17	Daura3,RijTsam21	L2a1d1	2	2.2%
18	Sandamu18	L2a1i	1	1.1%
19	Daura14	L2b	1	1.1%
20	Koza2,5	L2b1a	2	2.2%
21	Daura9,13	L2c1	2	2.2%
22	Sandamu1	L2e	1	1.1%
23	Daura4,15,23,24,Koza17,Sandamu9,17 ,21,22,RijTsam3,8,10,13,15,16	L3	15	16.1%
24	Koza8	L3'4	1	1.1%

SN	Sample ID	MRCA	Frequency	Percentage
25	Koza13,Sandamu4,11,24,RijTsam19	L3b	5	5.4%
26	Sandamu15,16	L3b1a	2	2.2%
27	Daura16,22,Koza18,RijTsam9,12,23	L3d	6	6.5%
28	Koza6,7,Sandamu6,22	L3d1	4	4.3%
29	Koza23	L3d3a	1	1.1%
30	Koza9	L3e1a	1	1.1%
31	Sandamu2	L3e1a3b	1	1.1%
32	Sandamu14	L3e1g	1	1.1%
33	Koza24	L3e2	1	1.1%
34	Rijtsam14	L3e2a1b2	1	1.1%
35	Sandamu3,RijTsam2	L3e2b	2	2.2%
36	Daura5,RijTsam18	L3f	2	2.2%
37	Daura12,RijTsam6	L3f1b	2	2.2%
38	Koza10,Sandamu19,RijTsam,4,5,20,22	L4b2	6	6.5%
39	Daura17,Sandamu13	M	2	2.2%
40	Daura20	U1a1d	1	1.1%
41	Sandamu7	U5b1b1b	1	1.1%
			93	100.00%

3.3.2 PHYLOGENETIC ANALYSIS AND POPULATION COMPARISON

Figure 3.1 illustrates a phylogenetic tree generated using MEGA11, with eight distinct clusters identified. The beginning of the tree appears at the top left of the figure3.1, with Sandamu17 to Koza20 forming cluster one and having the closest relationship to the root ancestry. Clusters two, three, four, five, six, and seven are represented by specific sets of samples, while clusters eight includes RijTsam16 to Koza17 and has the highest number of sub-branches, indicating a wider spread of the population from the root of ancestry. Notably, shared haplotypes from table 3.1 are mostly clustered together, often belonging

to the same or neighbouring branches, indicating a shared genetic trait between individuals.

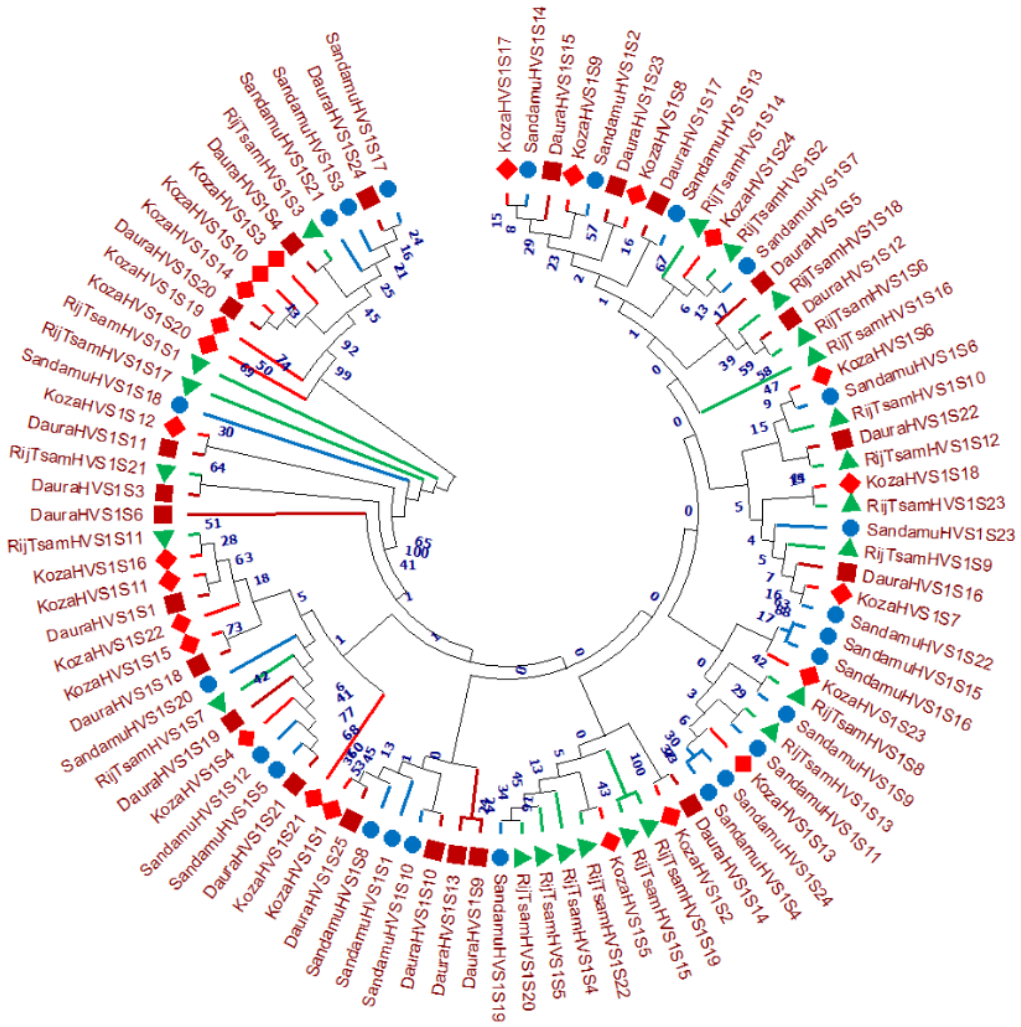


Figure 3.1: Phylogenetic Tree/ NJ Tree generated using Bayesian Clustering Analysis to show Ancestral and Sequence Relatedness Between the Study Population

The AMOVA analysis revealed that the majority of the variation is found within the study population group rather than between the groups. The population was divided into

four groups, namely Daura (1), Koza (2), Sandamu (3), and Rijiyar Tsamiya (4). The results indicated that 99.69% of the variance was within the study population while only 0.31% was found among the groups (see Table 3.3).

Table 3.3: Population Comparison Using AMOVA to Determine Genetic Variation within and between the Study Population

Source of Variation	Sum of squares	Variance components	Percentage variation	FST	p value
Among Populations	65.59	0.079	0.307	0.00307	0.03
Within Populations	2030.94	25.48	99.69		
Total	2096.52	25.56			

A genetic variance measure based on Wright's F-statistics (FST) was conducted using Arlequin. The results showed that there were negative FST values and p-values greater than 0.05, indicating no significant differences between populations (Table 3.4). The relationship between populations was further visualized using the R package "pkg" (http://cran.r-project.org/package/pkg) in Figure 3.2, where an "x" represents no significant differences between populations.

Table3.4: matrix of F wright's statistics Test and their respective p values

p value / **F_{ST}**	**1**	**2**	**3**	**4**
1	0	0.61261	0.71171	0.45045
2	-0.01424	0	0.36036	0.21622
3	-0.01888	-0.00541	0	0.69369
4	-0.00473	0.00985	-0.01499	0

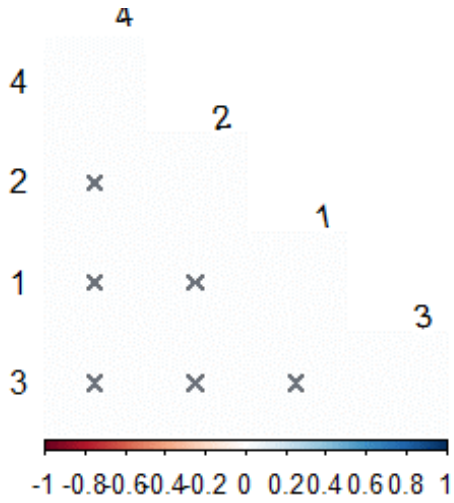


Figure 3.2: Distance matrix computation for population comparison

Figure 3.3 displays a haplotype network created using PopART, a tool developed by scientists at Otago University, New Zealand. The network was constructed using a neighbour-joining tree to show ancestral relationships among the study population. The network reveals that although the HVS1 region of mtDNA is conserved, mutations have created some degree of genetic variability between the study samples. Each small bar on the line connecting ancestral nodes represents a mutation, while larger nodes indicate shared ancestry between the samples. In the depicted picture, only three nodes are enlarged, and the remaining nodes are small or moderate, indicating little or no common ancestry.

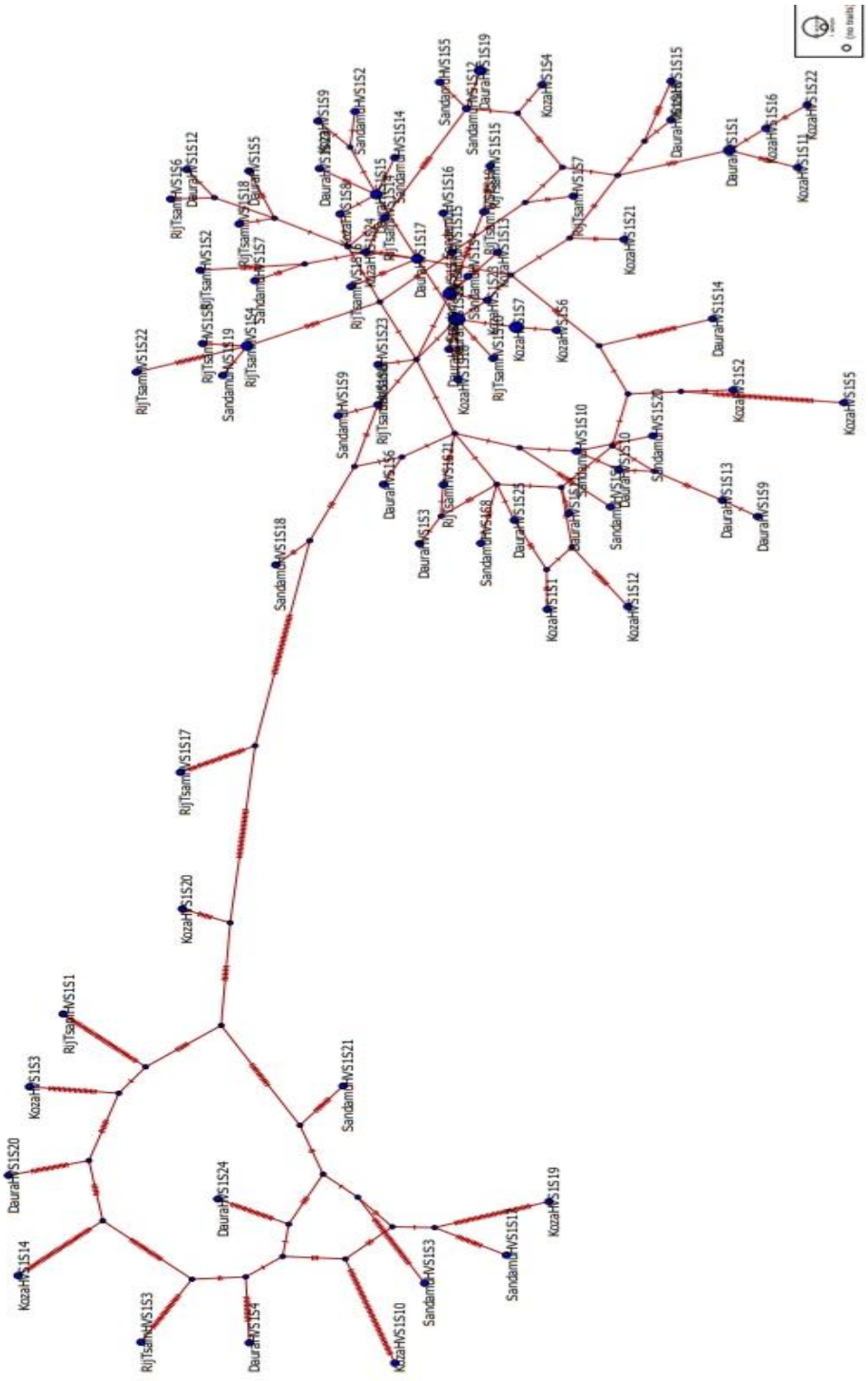


Figure 3.3: Haplotype Network using Maximum Likelihood to show haplotype and sequence relationship among the study population

3.4 DISCUSSION

The study's results are in line with those of other researchers who have found that the L Haplogroup is the most prevalent haplogroup in Africa, believed to have emerged in Africa approximately 150,000 years ago (150kya). It is not surprising, however, to find this haplogroup in other regions of the world, such as Central and West Asia and Europe [158]. This study categorized the population into four groups based on their geographical location: Daura, Koza, RijTsam, and Sandamu, representing population groups from the ancient towns of the kingdom that are believed by locals to be the roots of their ancestry for hundreds of years. Haplogroup L, including L0, L1, L2, L3, and L4 clades, was observed in the study. L3, one of the most recent clades believed to have originated around 50kya, was the most commonly observed macrohaplogroup, accounting for over 43% of the study population. According to the MRCA data, L3 and its subclades such as L3e2, L3e2b, L3f, L3e1a, L3d, and others make up over 46% of the population, making it the most frequent subclade in the study. Interestingly, this subclade is distributed evenly among all the study population groups, regardless of the area or town where the samples were collected. The study also showed a high level of genetic diversity among the population, with unique haplotypes accounting for 77.27% of the population. The second most observed haplogroup is L2 subclade, which accounts for about 19.8% of the population. These findings suggest that the most frequent haplogroups in the Hausa ethnic population in Daura Emirate, Nigeria are L3 and L2. Similarly, a study on African American population showed that L2a (19.8%) and L1b (10.2%) were the most commonly observed haplogroups, which are also commonly found in West Africa [175].

Similar to previous studies, the current research found L3 to be the most frequent subclade in the study population[26], [200], accounting for more than 46%. Additionally, L2 subclade-associated haplotypes were the second most observed, making up approximately 19.8% of the population. This suggests that L3 and L2 haplogroups are the most common among the Hausa ethnic population in Daura Emirate, Nigeria.

Other studies have also reported the presence of L1b, L2b, L2c, L2d, L3b, and L3d in the Fulani nomads, who are close Neighbours to the study population and often intermarry

with the Hausa. Veeramah et al., (2010) observed L3e as the most frequent subclade of L3 macrohaplogroup in the Cross River, Nigeria population, which is in line with the current study's findings.

Similarly, Martínez et al., (2019) conducted a study on the Yoruba population in Nigeria and reported L0 to L4 macrohaplogroups, with L3e being the most frequent subclade. This suggests that L3e subclade dominates the Nigerian population across the three major ethnic groups in the country.

The study identified several haplogroups that are not specific to African populations, including G2a1c1, G2a1g, G3a2a, M1a1a1, M30+16234, M39b, U1a1d, U5b1b1b, and U6a1a1, each observed only once in the study. Notably, all these non-L haplogroups are believed to have originated from L3 out of Africa haplogroups, which gave rise to anatomically modern humans (AMHs). TMRCA of the haplotypes was projected to have arisen between 20-25kya, except for G2 and G3, which were estimated to have originated about 10kya. These haplogroups are thought to have a common origin in the Near East, where G2 and G3 are primarily found in Europe and the Middle East [139], [184]. Samples Koza19 and 20 and RijTsam17 were found to have G2a1c1, G3a2a and G2a1g haplotypes, respectively. This suggests the possibility of an ancestral gene flow between the study population and people of Middle East, possibly through the spread of Islam. The M haplogroups are primarily found in South, Southeast and Central Asia, but M1 has a supra-equatorial African presence, particularly in Eastern and Northeastern Africa, with occasional occurrences in Northwest and West Africa, as observed in this study. The U1 and U5 haplogroups are primarily found in Europe, with a population percentage of 5% and 15-20%, respectively. In contrast, U6 is primarily found in North Africa, the Caucasus, and the Iberian Peninsula. The presence of non-L haplogroups indicates that there was gene flow between the study population and corresponding regions in ancient times, possibly during the colonial era, Islamic spread in Africa, or trade routes. Johnson et al., (2015) also reported the presence of M and U macrohaplogroups in the African American population.

The findings from locus-by-locus AMOVA revealed that the genetic variation was present within the population group and not between them. The genetic variance within the study population was observed at 99.7%, indicating that there was no significant genetic variation that could differentiate between individuals based on their location in Daura, Sandamu, Koza or Rijiyar Tsamiya. Additionally, the FST values suggested that the populations from these locations were not genetically distinct enough to be differentiated at a population level. As a result, the individuals were highly polymorphic, making them ideal for use in forensic reference and characterisation.

To further explore the genetic variation, a haplotype network was generated using PopART software. The network was created to compare the genetic sequences of the Hausa ethnic population from Daura, Koza, Sandamu, and Rijiyar Tsamiya, with the aim of inferring their relationships based on mtDNA sequence similarities and differences. The haplotype relationship between the study population was visually represented by the network. The network displayed two distinct star-like patterns on either side, connected by a long line. These circles and star patterns are connected by lines that signify evolutionary relationships among the haplotypes. Based on this, it can be inferred that the study population comprises two major groups. The length and number of steps on the lines are proportional to the evolutionary distance between the haplotypes, with shorter or fewer lines indicating a closer relationship. The bars on the lines represent mutations, with a higher number of bars indicating a greater number of mutations. In this study, RijTsam17(G2a1g) was observed in the middle of the bridge, while Sandamu18(L2a1i1) was the closest to the right group. As a result, it can be inferred that the right network branch exhibits a higher degree of genetic similarity to the L haplogroup.

The haplotype closest to the left major provenance is Koza20(G3a2a). Furthermore, the network suggests that this branch has a close evolutionary relationship with non-L haplogroups, as indicated by the presence of Koza19(G2a1c1), Daura4&20 (U6a1a1 & U1a1d), and RijTsam3(M1a1a1). The network comprises circles (nodes) that represent different haplotypes, with larger nodes indicating shared haplotypes and ancestry. It can

be observed that there are fewer large nodes, which suggests less shared haplotypes and matrilineal ancestry, thus supporting African matrilineal genetic diversity.

CONCLUSION AND RECOMMENDATION

This study presents mitochondrial DNA (mtDNA) genetic data based on HVS-I on the Hausa ethnic population from the historic Daura emirate in Nigeria. The study has demonstrated high genetic diversity among the African population, with the most frequent haplogroup being L-Haplogroup. Other non-L haplogroups such as G, M, and U macrohaplogroups were observed in a few individuals. The genetic variance was found to be mainly within the population, supporting the suitability of the genetic data for forensic and population genetics reference purposes. The nodes and branches of the haplotype network have shown fewer shared haplotypes and ancestry among the study population, indicating African matrilineal genetic diversity.

Given the importance of genetic data in forensic and population genetics, this study emphasises the need for more comprehensive genetic studies on the African population. Furthermore, the findings of this study can be used as a reference for future studies on the African population, including those on evolutionary history and population genetics. Finally, future research should aim to explore other genetic markers, such as Y-chromosome and autosomal DNA, to provide a more comprehensive understanding of the African population genetics.

CHAPTER FOUR

4.1 INTRODUCTION

The mitochondria contain their own genome, which bears a strong resemblance to that of bacteria. It has a double-stranded, circular configuration that is covalently closed and lacks histones. The absence of introns in genes within the mtDNA genome is a distinguishing attribute of prokaryotes, which lends support to the notion of bacterial ancestry (Ludes & Keyser, 2015). Mitochondrial DNA accounts for roughly 0.3% of an individual's genomic DNA, and each mitochondrion typically contains 2-5 complete copies of the genome. The year 1981 marked a significant milestone in the field of genetics when Anderson et al. successfully sequenced the complete mitochondrial genome. This seminal work produced the Cambridge Reference Sequence, an essential benchmark for subsequent research in mitochondrial DNA analysis [46], it was subsequently verified and amended for accuracy (Chinnery & Lightowlers, 1999). The original numbering conventions were retained for historical continuity. Variations from the rCRS with GenBank reference number NC 012920 are employed to describe sequences for ease of reference. The mitochondrial genome is comprised of H and L strands, and the density of each strand is dictated by the distribution of nucleotides in the corresponding region (Court, 2021).

The H-strand of mtDNA consists primarily of purine bases, while the L-strand has a higher proportion of pyrimidine bases. The entire nucleotide sequence of mtDNA, known as the mitogenome, has been identified and spans 16,569 bp in length (Sanger et al., 1977). The mtDNA contains conservative coding segments that house 37 genes, as well as a variable non-coding control region (Sanger et al., 1977).

Within the non-coding region, which includes the D-loop and a transcription promoter region called the "control region," there are 1121 base pairs. It should be emphasised that the D-loop consists of three distinct DNA strands, with one strand complementing another strand, keeping it separate and generating the displacement (D) loop (Kanti et al.,

2021). The rCRS sequence assigns a specific number to each nucleotide in the whole control region, which spans positions 16,024 to 576, for convenience (Court, 2021).

The mtDNA numbering system originates from an MboI digestion point within the control region and then proceeds around the molecule for a total length of 16,569 base pairs, which are relatively invariant (Court, 2021). Within the control region, three areas have the most sequence variability, and these are known as hypervariable regions (HVR). HVR1 (16,024-16,365), HVR2 (73-340), and HVR3 (438-574) are these HVR regions. HVR1 and HVR2 are roughly 300-400 bp in length and display more variation than HVR3. This is why they are widely used in forensic analyses. People who do not have the same maternal lineage have significant polymorphism in both HVR1 and HVR2 (Irena, 2020).

The population of Nigeria is composed of three primary ethnic groups: the Hausa, Yoruba, and Igbo. The Hausa tribe is mostly found in the northern region, while the Yoruba and Igbo people reside in the western and eastern regions of the country, accordingly (Sabiu et al., 2018). The Hausa ethnic group is the most populous in West Africa, with around 30 percent of the population residing in the northwest region of Nigeria, known as "Hausaland" (Liman, 2019). Hausa people are widely dispersed across Nigeria's northern region and remain one of Africa's most significant indigenous ethnic groups, with their language, Hausa, spoken by over fifty million people and serving as a lingua franca in West Africa (Sabiu et al., 2018). The research aims to examine the population structure and stratification of the Hausa ethnic group by examining their HVR2 mtDNA. The primary objective is to advance our understanding of the genetic diversity, haplotype diversity, and level of population substructure within the community, which we accomplish through HVR2 mtDNA analysis.

4.2 METHODOLOGY

4.2.1 CONSENT AND ETHICAL CLEARANCE

The individuals who participated in the study gave their consent and were fully informed about the research. The study's ethical aspects were carefully reviewed and authorized by

the Research and Ethics Committee of the Ministry of Health in Katsina State, Nigeria (MOH/ADM/SUB/1152/1/558), ensuring compliance with ethical standards. Local traditional leaders also gave their consent.

4.2.2 SAMPLE POPULATION

Following obtaining their consent, individuals provided buccal swabs for the study. The research team collected samples from four different areas in three Local Government Areas (LGAs) that encompass the ancient cities of the Daura emirate, namely Daura, Koza, Sandamu, and Rijiyar Tsamiya. The study recruited 100 individuals, and out of those, 94 individuals underwent successful sequencing. The population samples were categorized into four groups, which were Groups 1, 2, 3, and 4 that corresponded to Daura (n=24), Koza (n=23), Sandamu (n=24), and Rijiyar Tsamiya (n=23), respectively. The study excluded individuals who shared a known maternal ancestor during the selection process.

4.2.3 LABORATORY METHODS

The participants in the study had their left and right cheeks gently scratched for 10 seconds with two separate swab sticks, and the swabs were then placed in a container and stored on ice after being dipped in buffer. To obtain the complete genome, including mtDNA, the QIAamp DNA mini kit (Cat. No. 51304, QIAGEN Heidelberg, Germany) was employed, adhering to the manufacturer's protocol. The concentration and purity of the extracted DNA were measured using Nanodrop spectrophotometry. To prevent annealing with segments of the nuclear genome, the primer sequences were checked against the human reference genome using the BLASTn suite of the NCBI. The study targeted and amplified mtDNA HVR 2 using forward and reverse oligo sequence primers Forward-5' GGT CTA TCA CCC TAT TAA CCAC3' and Reverse-5' CTG TTA AAA GTG CAT ACC GCCA3' (Černý et al., 2006).

To amplify the target, Polymerase Chain Reaction (PCR) was employed with a total of 35 cycles. The amplification process initiated with an initial denaturation step at 94^{o}C for 5 minutes, followed by denaturation at 94^{o}C for 30 seconds, annealing at 53^{o}C for 30

seconds, extension at 72°C for 1 minute, and final extension at 72°C for 7 minutes. Subsequently, the reaction mixture was maintained at 4°C. The final reaction volume of 25µl was prepared, comprising of 12.5µl of one Taq Quick-load 2x master mix with standard buffer obtained from New England Biolabs Inc. (website: www.neb.com/M0486), 0.5µl for each of the forward and reverse primers, 2µl of the diluted DNA template, and 9.5µl of nuclease-free water. Quality control was maintained in each run with the inclusion of a negative control. Exonuclease I and shrimp Alkaline Phosphatase (ExoSAP) treatment were used to eliminate any unincorporated dNTPs and primers left in the PCR reaction before sequencing. To estimate the amplicon size on a 1% agarose gel, agarose gel electrophoresis was used. After amplicon size estimation, the forward strand was sequenced using a big dye terminator, and fragments detection and estimation were done on the ABI prism 3500xl genetic analyser (Applied Biosystems, United States).

4.2.4 DATA ANALYSIS

The process of haplogroup assignment was conducted using two online tools, mitotool and haplotracker, which were then compared to PhyloTree build 17.0. The consensus haplogroup of the test samples was assigned by comparison with the rCRS, with both tools used to resolve any ambiguous haplogroups. Microsoft Excel 2010 was used to manually count shared haplotypes, while Bioedit v.7.2.5 was used for base calling and multiple alignment, with ClustalW used to align the files in fasta format. The Analysis of Phylogenetic and Evolution (ape) package of the R program was used to generate Arlequin, Structure, and Phylip files from the fasta haploid data format. Arlequin v.3.5.2.2 was used to perform locus-by-locus calculations of molecular variance (AMOVA), while the Sequence Demarcation Tool (SDT) v1.2 was utilised to determine pairwise identity scores at the individual level. To examine ancestral relationships and population admixture, Structure v.2.3.4 was used, with the study population divided into four groups, and Markov Chain steps and burn-in lengths set at 10,000. The best run for inferring ancestral relationships among the study populations was determined using structure harvester.

In addition, we downloaded the mitogenome sequences from the NCBI database of different global populations in order to make comparison with our genetic data. We downloaded from Southern Africa (Angola), Oceania (New Zealand) Tokelau population, South Asia (west Indian Caste), Central Europe (Switzerland), South America (Paraguay) and North America (Canada, Newfoundland) with accession numbers MF3812871-MF3813061, MT9282831-MT9282971, MK0439671-MK0439862, MT0790191-MT0790371, MH9818231-MH9818421 and MF5887941-MF5888111 respectively. We used Arlequin to conduct AMOVA and use the FST values generated to make a pairwise identity matrix between the populations using R statistical tool, SDT was used to make pairwise comparison of individual sequences, and STRUCTURE to make ancestral admixture within and between the population across the continents. We used 15-20 sequences from each of the continental populations. Markov Chain steps were set at 10,000, and a burn-in length of 10,000 was used with K replication values set at 6, and 3 iterations (Evanno et al., 2005). To determine which of the eighteen runs was best suited for inferring ancestral relationships among the study populations, a structure harvester was used (Earl & vonHoldt, 2012).

4.3 RESULT

4.3.1 HAPLOGROUP

The haplotype data of mtDNA HVR2 from the Nigerian Hausa population is presented in Table 4.1. Out of the identified 62 haplotypes, 44 were found to be unique, while 18 were common among individuals. The table also includes the frequency and percentage of each haplotype. The unique haplotypes make up the majority, representing 71% of the total haplotypes. The L haplogroup is the most prevalent, accounting for 87% of the study sample, with subclades L2 (25%) and L3 (46%) being the most common. Other haplogroups observed include G, M, and U, which together constitute 13% of the haplotypes. The most frequently shared haplotypes were L3a1a and U5a2b3a1, each shared by 5 individuals, while L3b1b and L3c were shared four times each. Haplogroups

L0, L1, and L4 were rare, present in only a few individuals, and accounted for 11% of the total haplotypes.

Table4.1: Haplotype Distribution, Frequency and Percentage among the Study Population

SN	sample ID	Haplotype	frequency	Percentage
1	Koza19	G2a4	1	1.1
2	Daura20	L0a1a+200	1	1.1
3	Daura7, Sandamu5	L0a1a2	2	2.1
4	Sandamu12	L0a1a3	1	1.1
5	Daura18	L0a1d	1	1.1
6	Koza11	L1b1a4	1	1.1
7	Daura1	L1b1a4a	1	1.1
8	RijTsam11	L1b1a15	1	1.1
9	Daura17	L1c3b1b	1	1.1
10	Daura16	L2a1	1	1.1
11	RijTsam16,17	L2a1a	2	2.1
12	Sandamu23	L2a1a1	1	1.1
13	Sandamu8	L2a1c1a1	1	1.1
14	Daura11,Koza1,Sandamu7	L2a1c2	3	3.2
15	Daura19	L2a1c3	1	1.1
16	Daura10,24,Sandamu20	L2a1c3b2	3	3.2
17	Daura3	**L2a1c4a1**	1	1.1
18	RijTsam21,Koza10	L2a1d1	2	2.1
19	Daura6	L2a1d2	1	1.1
20	Koza21	L2a1g	1	1.1
21	Sandamu18	L2a1i1	1	1.1
22	Koza12,Sandamu6	L2a1k	2	2.1
23	Sandamu21	L2a2b1a	1	1.1
24	Koza2	L2b1a2	1	1.1
25	Koza5	L2b1a4	1	1.1
26	Daura9,13	L2c1a	2	2.1

SN	sample ID	Haplotype	frequency	Percentage
27	Koza22,Sandamu9,RijTsam8,12,15	L3a1a	5	5.3
28	Sandamu4,11,22	L3b1a+152	3	3.2
29	RijTsam19	L3b1a1a	1	1.1
30	Sandamu15,16	**L3b1a3**	2	2.1
31	Daura2,RijTsam3,13,18	L3b1b	4	4.3
32	Sandamu13,RijTsam1,4,10	L3c	4	4.3
33	RijTsam9	L3d1a	1	1.1
34	Sandamu24	L3d1b3	1	1.1
35	Koza16,RijTsam23	L3d1b3a	2	2.1
36	RijTsam2	L3d1c1	1	1.1
37	Koza6,7	L3d4a	2	2.1
38	Sandamu14	L3e1	1	1.1
39	Sandamu2	L3e1a3	1	1.1
40	Koza9	L3e1a3a	1	1.1
41	Koza17	L3e1b2	1	1.1
42	Koza8	L3e1e	1	1.1
43	Daura14	L3e1f	1	1.1
44	Koza20	L3e2a	1	1.1
45	Daura8	L3e2a1a	1	1.1
46	RijTsam14	L3e2a1b2	1	1.1
47	Koza23,Sandamu3	L3e2a2	2	2.1
48	Daura23,Sandamu17	L3e2b1a	2	2.1
49	Sandamu1,RijTsam20,22	L3e2b6	3	3.2
50	RijTsam6	L3f1b	1	1.1
51	Daura12	L3f1b1a1	1	1.1
52	RijTsam7	L3f2a1a	1	1.1

SN	sample ID	Haplotype	frequency	Percentage
53	Daura5	L3f2b	1	1.1
54	RijTsam5	L4b2b	1	1.1
55	Sandamu19	L4b2b1	1	1.1
56	Koza4	M13a2	1	1.1
57	Koza18	M39	1	1.1
58	Daura15	M45a	1	1.1
59	Sandamu10	U4a1b2	1	1.1
60	Daura21,22,Koza3,14,15	U5a2b3a1	5	5.3
61	Daura4	U6a1a1	1	1.1
62	Koza13	U7a5	1	1.1
			94	100

4.3.2 POPULATION STRUCTURE AND STRATIFICATION

In addition to haplogrouping, we employed AMOVA to categorize the population based on their genetic sequence variation, both within and between the population groups. The genetic variation among the study populations is presented in Table 2, which indicates that the majority of the genetic variation (97%) was within the population, with only 3% attributed to differences between the study population groups. The FST and p values obtained from the analysis indicate a high level of confidence in our results.

Table 4.2: Population Comparison Using AMOVA to Determine Genetic Variation within and between the Study Population

Source of Variation	Sum of squares	Variance components	Percentage variation	FST	p value
Among Populations	82.08	0.65287	3.30576	0.03306	0.01
Within Populations	1371.869	19.09646	96.69424		
Total	1453.95	19.74932			

The AMOVA in table 4.3 has shown that the genetic variation is conferred between the population groups. The higher FST value of 0.9 indicate high genetic diversity with a p-value that spells confidence in the analysis.

Table 4.3: AMOVA making population comparison to determine how further or close the populations are related

Source of Variation	Sum of squares	Variance components	Percentage variation	FST	p value
Among Populations	3346.430	29.77095	90.32312	0.90323	0.0001
Within Populations	392.600	3.18955	9.67688		
Total	3739.030	32.96050			

Further to that figure 4.1 generated based on the individual population group FST values has shown genetic distinctiveness between the population groups, however statistical significance was markedly shown between population group 7 (North America) and group 1 (West Africa).

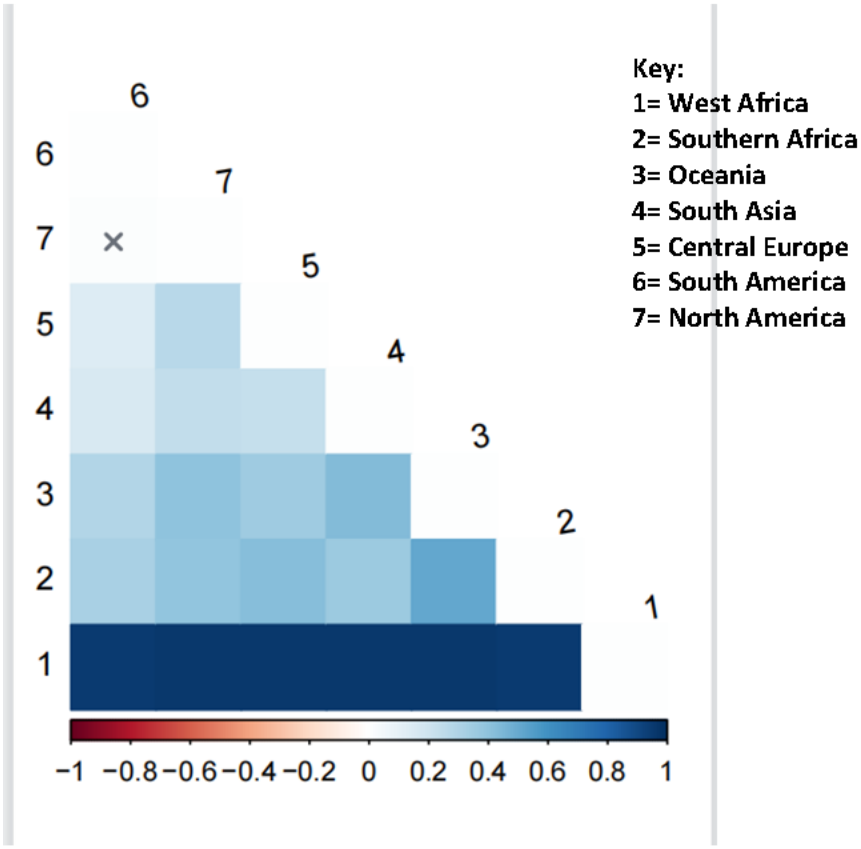


Figure 4.1: Matrix of FST values showing the genetic relatedness between the populations across continents

To structure the population based on pairwise identity comparison between individuals, we employed SDT, as depicted in Figure 4.2. The darker colours in the figure indicate higher percentage identity, implying lower genetic uniqueness. It is noteworthy that a considerable number of individuals have lower identity scores when compared to others within the study population.

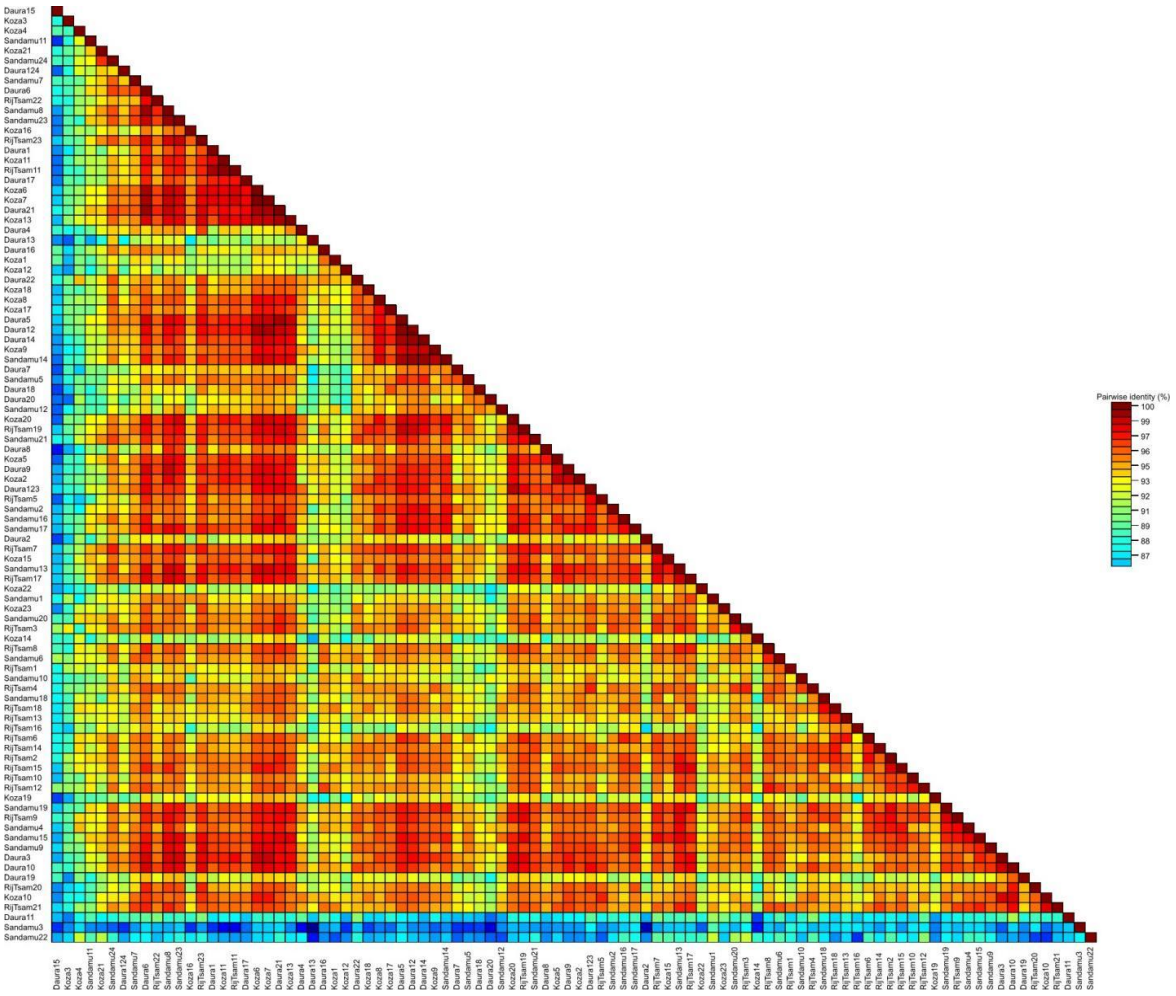


Figure 4.2: The SDT interface is presented, featuring a pairwise identity matrix that is colour-coded to represent the mtDNA HVR2 of the Nigerian Hausa population. Each cell in the matrix is coloured and indicates the percentage identity score between two sequences. The sequences are positioned horizontally and vertically at the bottom of the matrix. A coloured key is also included to illustrate the relationship between the pairwise identity scores and the colours presented in the matrix.

The distribution of percentage pairwise identities is depicted in Figure 4.3. On the graph, the horizontal axis denotes the percentages, while the vertical axis shows the frequency distribution corresponding to these percentages. The distribution displays peaks and troughs, indicating desirable and undesirable thresholds, respectively. These thresholds may be provisionally utilised to establish cut-offs for operational taxonomic unit demarcation that are relatively free of inconsistencies.

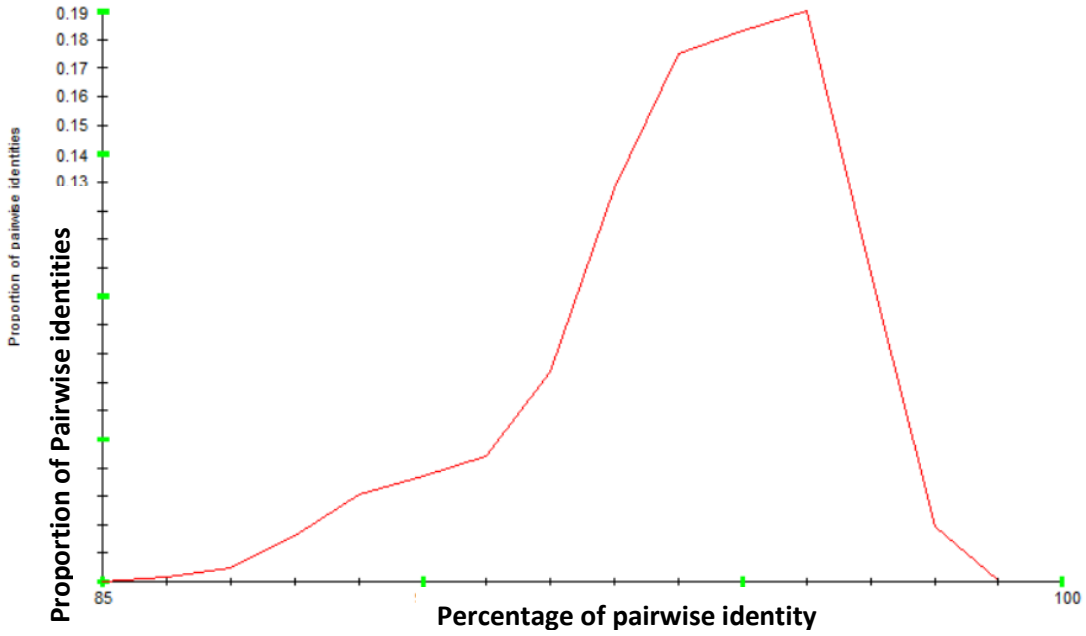


Figure 4.3: The Plot of Pairwise Identity Frequency Distribution which Demonstrates the Proportion of Pairwise Identities at Different Percentages.

The SDT depicted in figure 4.4 uses a colour-coded matrix to show pairwise identity at individual as well as population levels, it has shown genetic distinctiveness of our genetic data (West Africa), which is remarkably different from all the populations under comparison. Due to large dataset, the coloured matrix has been edited to show labels of the major population represented in each cluster, it should be however noted that there are some population overlaps that cannot be labelled accordingly. A supplementary high quality picture is provided for reference.

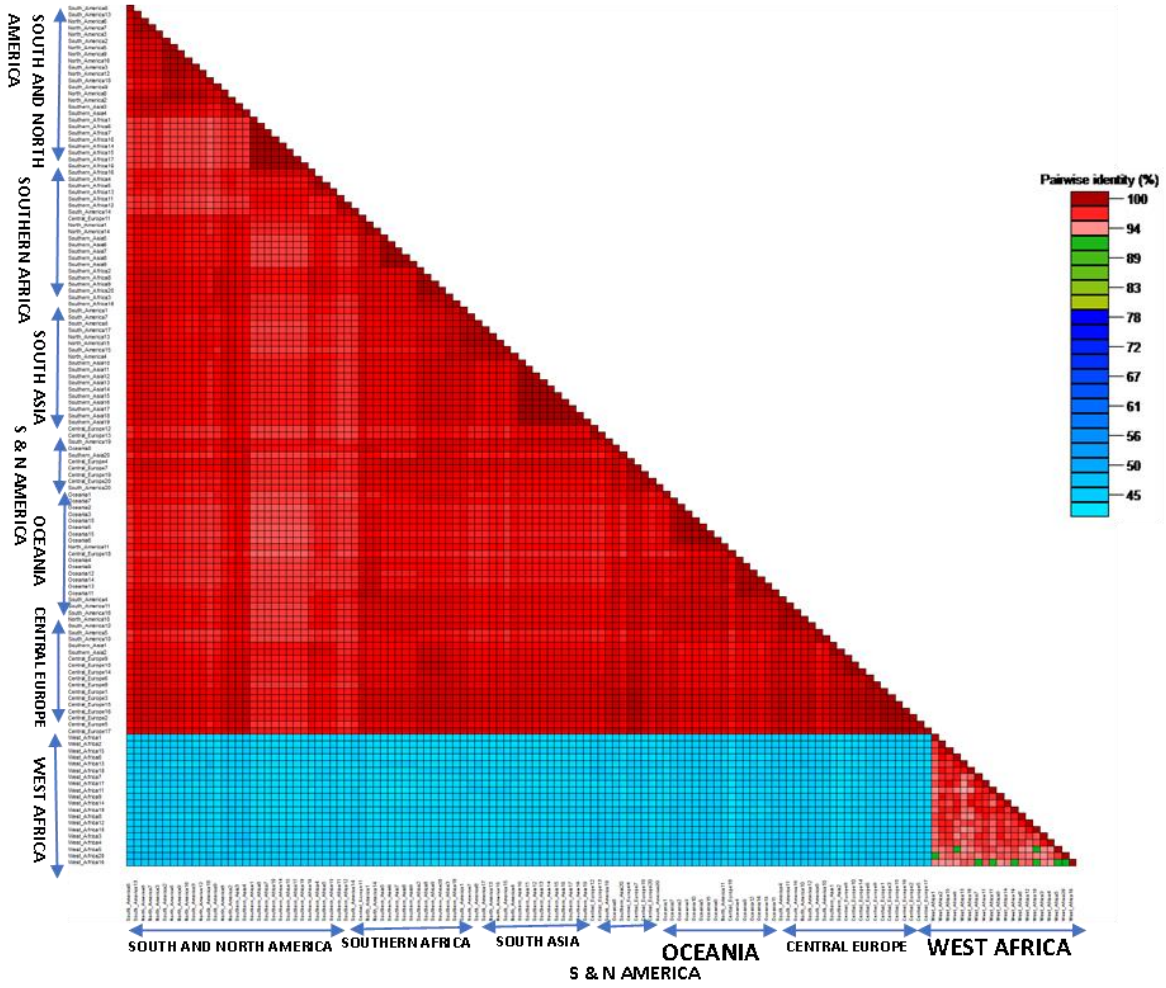


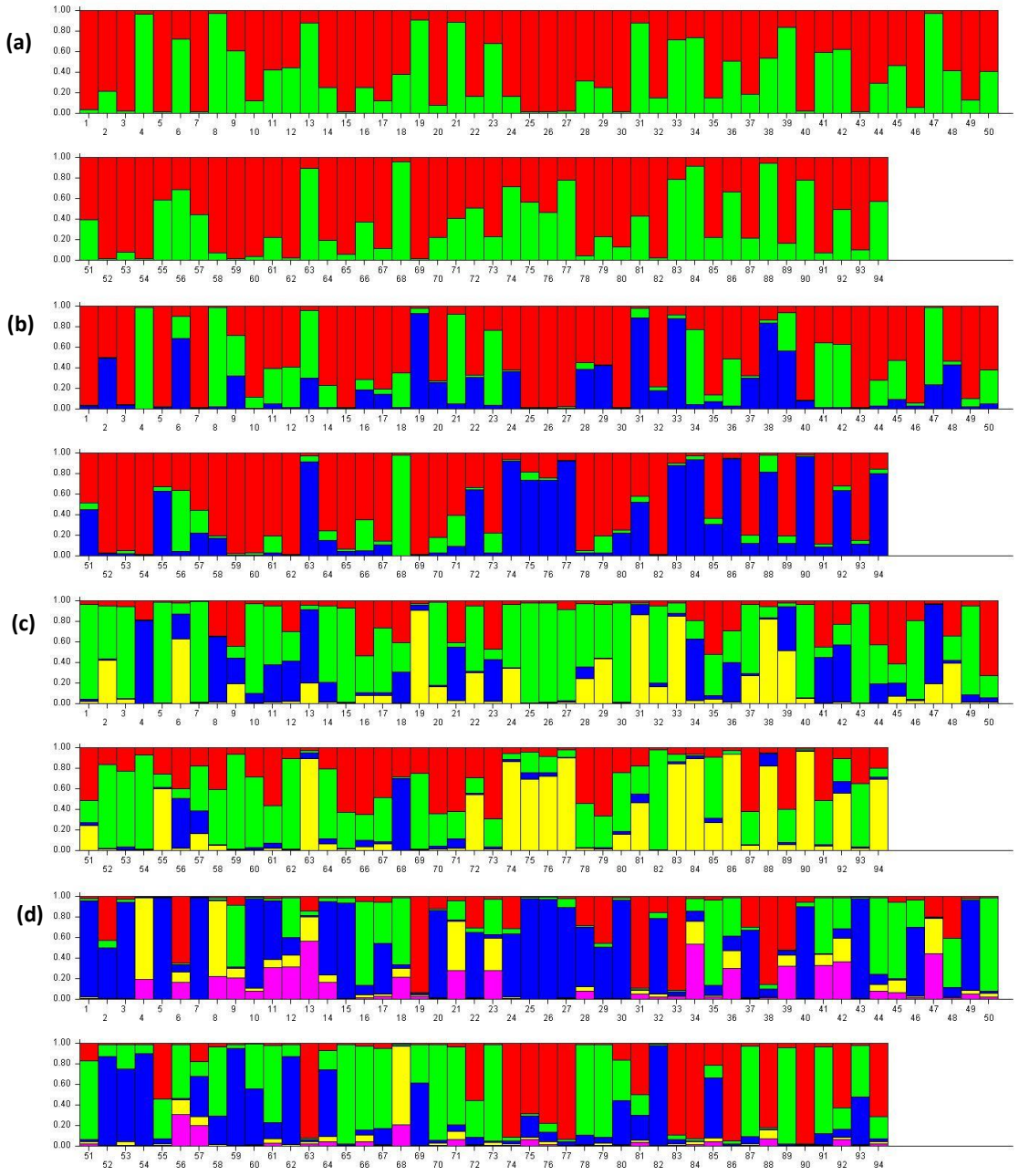
Figure 4.4: colour-coded pairwise identity matrix featuring our study population and other global populations revealing genetic relatedness and/or distinctiveness between the populations.

Figures 4.5a-i illustrate the results of population admixture analysis using the structure program v2.3.4 (Falush et al., 2003; Hubisz et al., 2009; Porras-Hurtado et al., 2013; Pritchard et al., 2000). The analysis was performed with 94 individuals, 639 loci, and 10,000 burn-in periods and replications each, using a number of replications (k) set at 10 with three iterations each. The figures, labelled from "a-i" corresponding to "k=2 to k=10," respectively, represent the number of assumed populations. Each individual's admixture and subpopulation groups (1-24 for Daura, 25-47 for Koza, 48-71 for Sandamu, and 72-94 for Rijiyar Tsamiya) were studied to investigate population

substructure at both individual and population levels. This approach has enabled us to examine individual admixture and cluster individuals into subpopulations for a more detailed analysis of population structure.

The admixture analysis of population using the structure program v2.3.4 is presented in Figure 3, with the number of replications set at 10 and three iterations each, and other parameters set to 94 individuals, 639 loci, 10,000 burn-in period and replications each. The figures are labelled as “a-i,” representing the values of k=2 to k=10. The k values represent the number of assumed populations, and the individuals are divided into subpopulation groups: 1-24 (Daura), 25-47 (Koza), 48-71 (Sandamu), and 72-94 (Rijiyar Tsamiya).

Figure “a” (k=2) demonstrated a similar population substructure across the population groups, with individuals showing comparable population admixture regardless of their population groups. Figure “b” (k=3) assumed three distinct population groups, and the first three population groups displayed similar population substructure, except the Rijiyar Tsamiya population that showed a distinctive ancestral admixture with high deposition of ancestry represented by the blue colour. Based on Figure “c” (k=4), population groups 1 and 2 exhibited similar patterns of ancestral admixture, and the first three populations have similar population sub-structuring with some degree of observable individual differences based on ancestral composition.


(a)
(b)
(c)
(d)

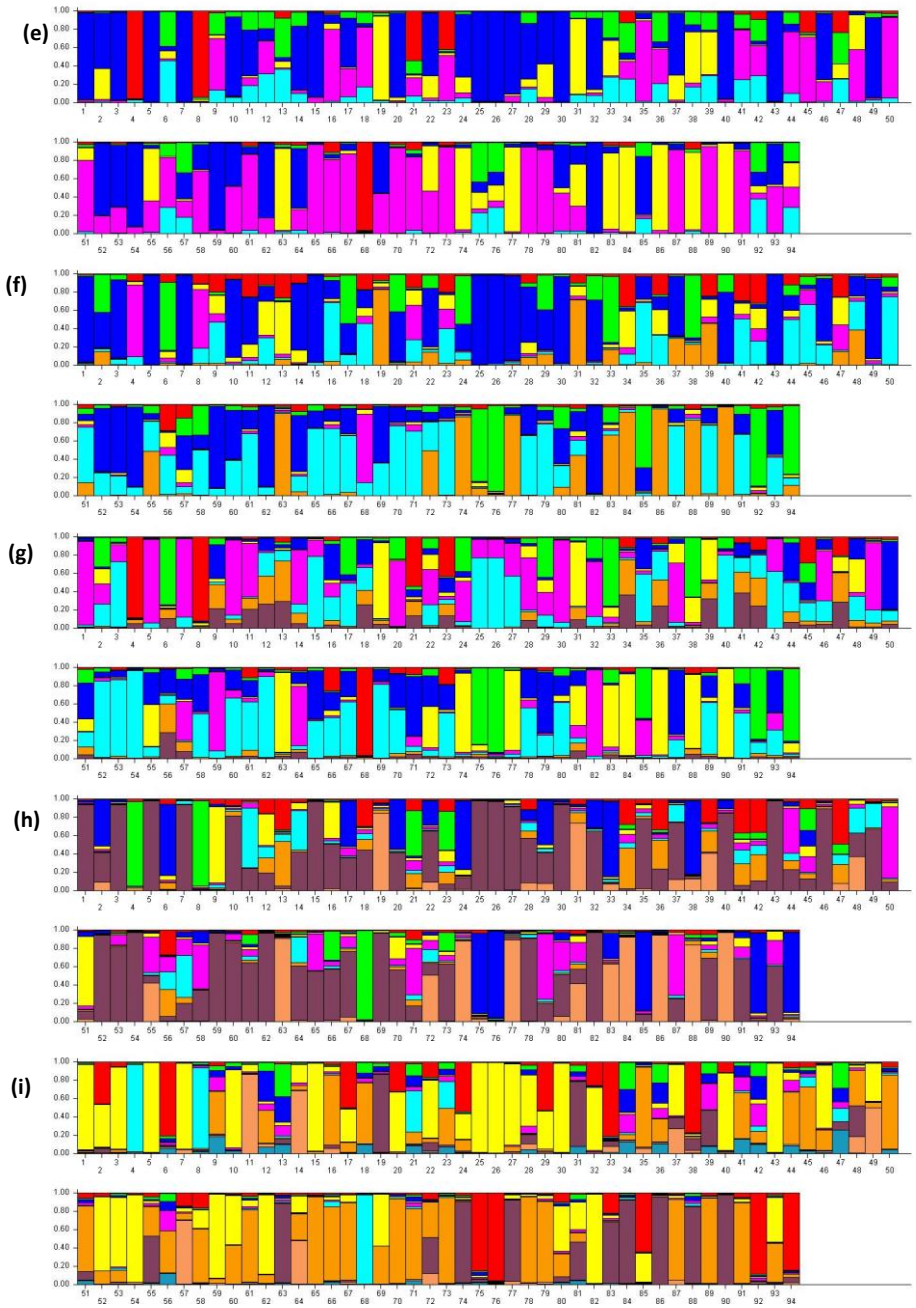


Figure 4.5a-i: Bar plots generated by the Structure software, the plots are showing ancestral admixture, that can be used to infer ancestral relationship as well as categorise the study population into subpopulations. the k values represent the number of assumed populations.

However, it is worth noting that the Rijiyar Tsamiya population stands out with a high proportion of yellow ancestry, as previously observed. In Figure "d," assuming five population groups, we noticed that the fifth population, represented by yellow, is poorly represented across all population groups. More importantly, the last population group has very little admixture from this fifth assumed population. As seen in Figure "e," population substructure exists between the population groups, each exhibiting a unique pattern of ancestral admixture. While some groups display relationships, others, such as the Rijiyar Tsamiya population, demonstrate a more distinct population substructure. Figure "f" illustrates population admixture assuming a total of seven populations, with each population group displaying varying degrees of differing admixture composition in subpopulation groups. The eighth and ninth assumed population simulations are shown in figures “g” and “h”, respectively. In figure “g”, each of the four population groups exhibits distinct substructure, and within each group, individuals show differences in ancestral admixture. Figure “h” assumed nine populations and revealed unique population substructure within each group, with some individuals exhibiting multiple admixed ancestry. Figure “i” maintains a similar pattern, with individuals displaying unique admixture and multiple population subgroups observed within the groups. It is worth noting that certain individuals have displayed a distinct genetic makeup throughout the analysis. Some individuals have maintained a preserved genetic diversity with very little admixture throughout the assumptions, while others have demonstrated multiple population admixtures. For example, individuals 1, 3, 4, 5, 7, 25, 26, 27, 30, 43, 52, 53, 72, 82, and 84 have consistently exhibited little admixture. On the other hand, individuals 2, 12, 13, 23, 28, 41, 42, 55, 70, 71, 72, and 89 have shown multiple population admixtures.

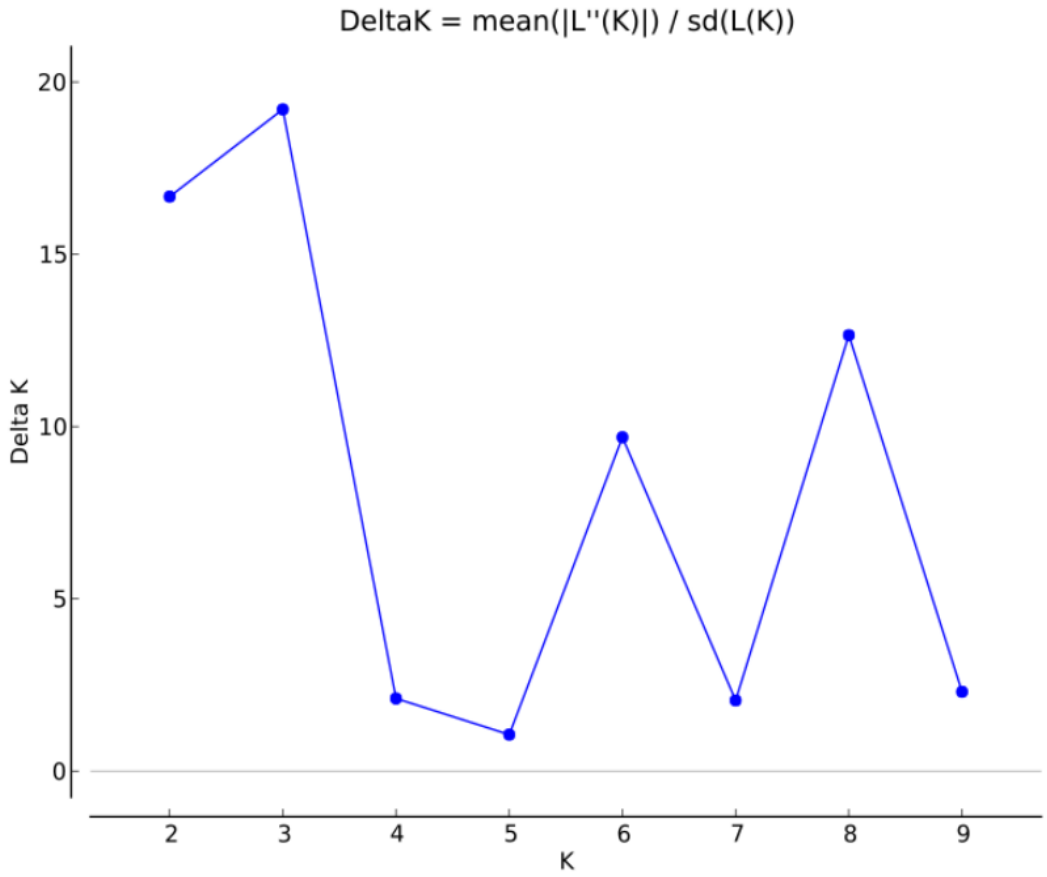


Figure 4.6: The Delta K Values Generated Using Structure Harvester. It Uses Statistical Assumptions to Determine the Optimum Value of K

Figure 4.6 illustrates the Delta K value of likelihood at various K set values and iterations used in the analysis. This graph provided us with a visual representation of the most appropriate run that accurately represents shared ancestry. According to the figure, run number 3 was the best fit, and as a result, figure 2b provides the most precise depiction of the admixed population. We carefully examined the figure to draw an informed inference on the ancestral admixture, with the assumption of three populations as previously stated. Individual samples from Daura, Koza, Sandamu, and Rijiyar Tsamiya were examined at the individual level, and it was observed that samples 1, 3, 5, 7, and 15 from Daura had a similar ancestral admixture with more than 95% red cluster and the remaining blue cluster. This same mode of admixture was observed in samples 25, 26, 30, 40* and 43

from Koza, samples 52, 54, 62, and 69 from Sandamu, and individual number 82 from Rijiyar Tsamiya. In the discussion, this group was referred to as cluster 1. Cluster 2 was assigned to individuals represented by green and red, with more than 95% green cluster, and only two individuals were distinctly observed: individuals 4 and 68 from Daura and Sandamu, respectively. The third cluster was assigned to individuals represented by blue (>95%) and red (<5%), and this was observed in only two individuals (77 and 86) from Rijiyar Tsamiya. We designated the fourth cluster to individuals with a combination of red and blue clustering, with roughly 50-55% red and 45-50% blue. This subpopulation cluster was only observed in individuals 2 and 29 from Daura and Koza, respectively. Our results indicate that clusters 1-4 are the only clusters with subpopulations represented by only two ancestral admixtures. Other individuals in the study exhibit three ancestral compositions. It is noteworthy that no individual was found with only one ancestral lineage.

We conducted further analysis on individuals exhibiting three population admixtures. We identified individuals with a predominant ancestral lineage of >90% in either the green, blue, or red cluster, and the remaining <10%, and assigned them to clusters 5, 6, and 7. Individuals with >50% but <90% ancestral composition in the green, red, and blue clusters were assigned to clusters 8, 9, and 10, respectively, while the remaining population had an admixture composition of <50%. We also observed other individuals with varying degrees of ancestral admixture not fitting into the above clustering categories, and a unique individual (number 9) with almost equal ancestral composition from all three assumed populations, which can be considered as subpopulations in this study.

Further to ancestral admixture analysis conducted on our genetic dataset, we analysed ancestral admixture between our study population and other populations from different continents. We set K value at (1 to 6) which corresponds to the number of assumed ancestral populations. Figure 4.7 shows the runs based on the k values, when k=2 which assumed only 2 ancestral populations, it could be observed that West African

Population's ancestry is distinctly different from the rest of the population. K= 3 retain the genetic distinctiveness of the West African population. However, other populations exhibited shared ancestry. When k=4, West African population remain different from the rest of the population, it should be noted here that all others were also different from the Angolan (Southern Africa) population, although they have shared ancestry with the rest of the population. An insignificant amount of ancestry represented by yellow bars can be observed at the bottom of the West African, South Asian and Central European populations. When 5 ancestral populations were assumed (k=5), three dominant ancestral root were observed with some minute presence of other ancestral lineage in across the populations. It is noteworthy that West African population maintained the status quo of genetic distance with other populations. Southern African population also exhibited some level of ancestral lineage different from the rest of the population, although the ancestry is shared by the other populations, a close ancestral relationship between the Southern African population and South Asian population is observed. K= 6 revealed genetic distance of the West African population with some insignificant shared ancestry with the rest of the population. Another distinct population is the Southern African population. It should be observed that the rest of the population are dominated by similar ancestral lineage.

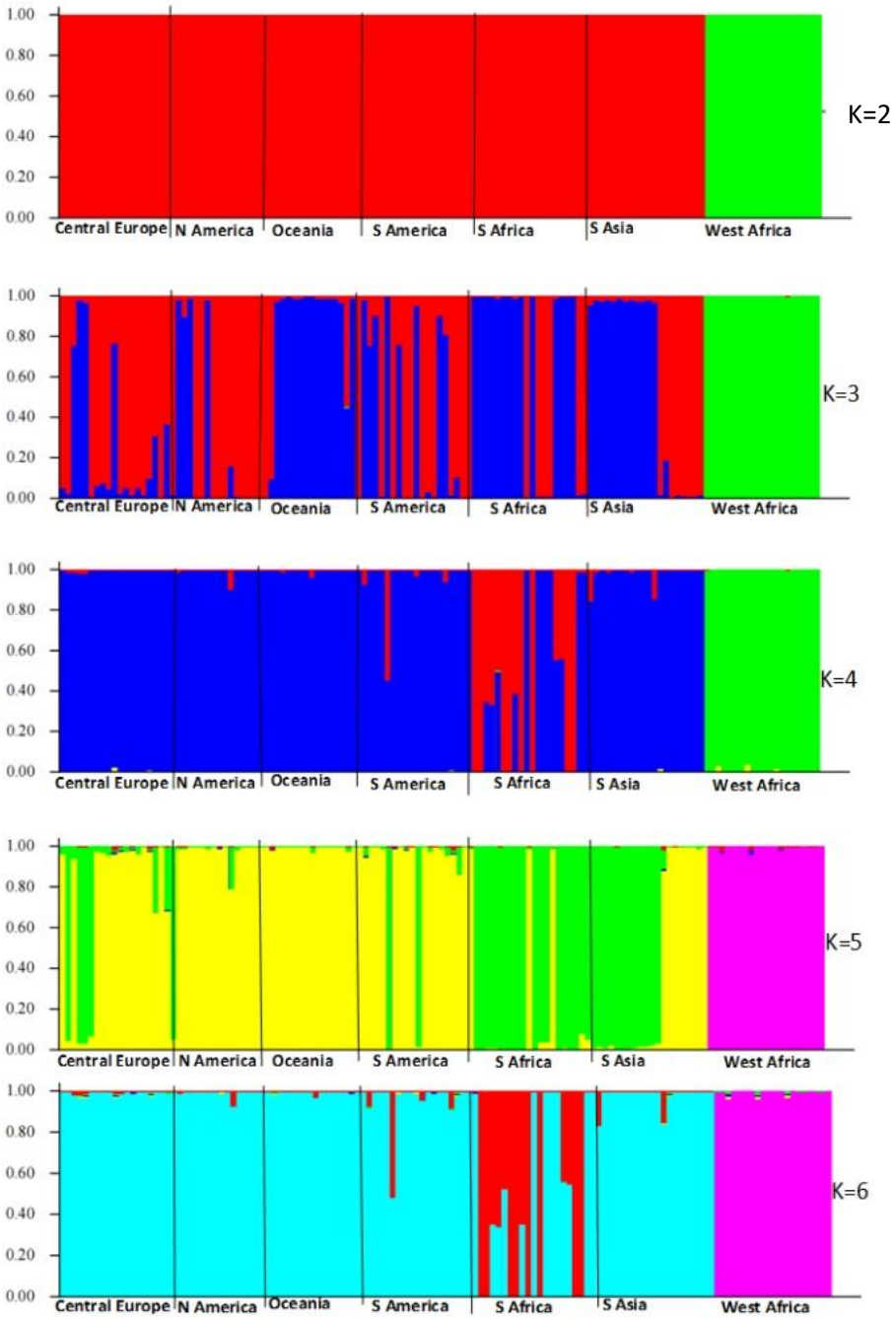


Figure 4.7: Structure statistical analysis inferring ancestral admixture among some selected populations across the global continents

We further used structure harvester to determine the most reliable run that could be used to infer ancestry of the populations, the Delta K value indicated that indicated optimum inferred ancestral group is k=2 (figure 4.8). Therefore, the most reliable inferred ancestry of the population under study is represented when only two ancestral populations were assumed. This placed the West African population at a distant position compared to the other population groups.

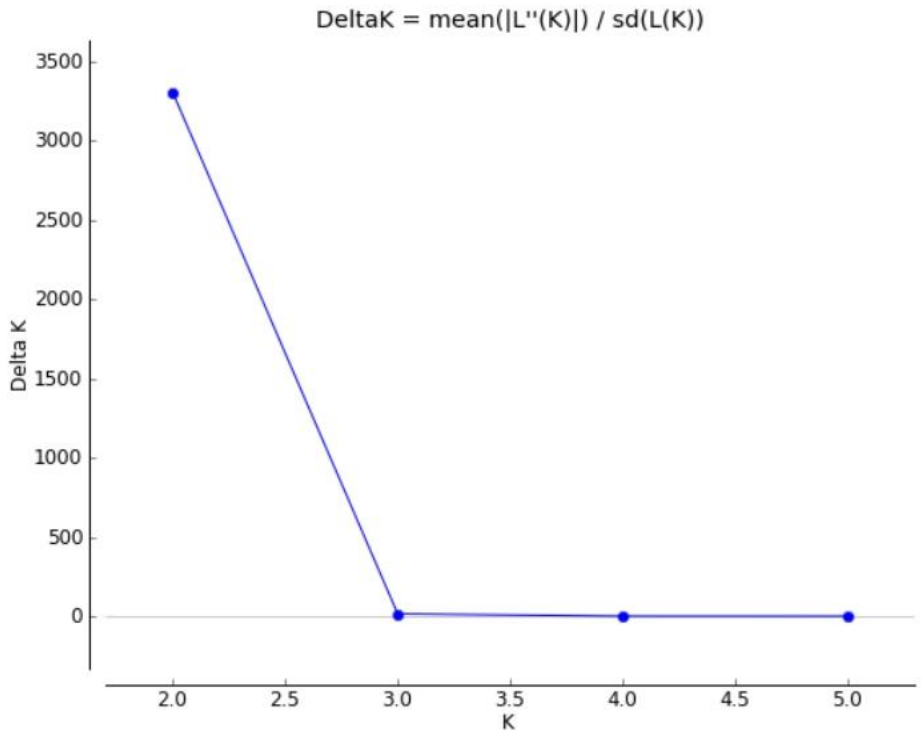


Figure 4.8: Delta K value plot, which shows the optimum negative log-likelihood that can be used to infer the assumed ancestry.

4.4 DISCUSSION

Numerous studies have reported that the African mtDNA haplogroup is mainly dominated by the L-haplogroup (Cabrera et al., 2018; Göbel et al., 2020; Rito et al., 2013). The L haplogroup is a fascinating and intricate lineage that provides insight into the early migratory patterns and demographic histories of human populations. Its diversity and distribution make it a valuable tool for comprehending the genetic diversity

and human evolutionary history (Chen et al., 1995; Fregel et al., 2019; Hernández et al., 2015; McElhoe et al., 2022). Among all the haplogroups, L is one of the most ancient and diverse. It is believed to have originated in Africa approximately 150,000 years ago, and its diversity and distribution reflect the early migratory patterns of human populations out of Africa (Diallo et al., 2022; Fortes-Lima et al., 2022).

Numerous subclades can be found within the L haplogroup, reflecting different demographic histories and migratory patterns. Among the most well-known subclades is L3, which is believed to have originated in East Africa and is thought to have given rise to all non-African mtDNA lineages. Another subclade, L1, is most frequently found in Western and Central Africa and is considered to be one of the oldest branches of the L haplogroup.

The L haplogroup is found in populations across Africa, Asia, Europe, and the Americas, with sub-Saharan Africa having the highest prevalence of approximately 70% of all mtDNA lineages. Additionally, it is found at high frequencies in some populations in South Asia and Oceania. The study of the L haplogroup has provided insights into the migratory patterns of early human populations out of Africa. For example, the presence of L3 lineages in Southeast Asia and Australia suggests that the earliest modern humans may have migrated along the southern coast of Asia and across the sea to Australia.

Furthermore, the diversity of the L haplogroup in Africa has been utilised to reconstruct the demographic history of African populations. The West African region, with its high diversity of the L haplogroup, is suggested to have been a major centre of genetic diversity and population expansion, according to previous studies (Fregel et al., 2019; Rito et al., 2013).

Our study also found 44 unique haplotypes, accounting for 70% of the observed haplotypes, which further supports the genetic diversity of the African mtDNA as previously reported by other researchers (Fendt et al., 2012; Gomes et al., 2010; Oliveira et al., 2015). Nevertheless, 30% of the observed haplotypes were shared by some

individuals, and the maximum number of shared haplotypes recorded was five. Among the 62 observed haplotypes, approximately 90% were represented by the L haplogroup, with L3 and L2 being the most commonly observed, which is consistent with the findings of previous studies in the Nigerian Yoruba population (Martínez et al., 2019).

Contrary to previous findings which only reported L haplogroups, our study detected the presence of G, M, and U haplotypes. The occurrence of M and G haplogroups suggests a degree of genetic admixture with North African or West Asian populations, which is not surprising given the Hausa population's geographic location. Additionally, the presence of North African dominant haplotypes could be attributed to the long history of trade and Islamic propagation from North Africa into West Africa. Given the lack of mtDNA data on the Hausa population, these findings are particularly significant.

We utilised the SDT tool, which provides a colour-coded pairwise identity score matrix to analyse the relationships between sequences within our dataset. This tool offers a more user-friendly approach compared to conventional tables of pairwise sequence identity scores commonly used for this purpose. The SDT pairwise identity matrices are organized such that individuals with close phylogenetic relatedness are placed closer together on the matrix, while genetically distinct individuals are placed further apart. A lower identity score indicates greater genetic distinction between individuals, enabling us to distinguish more readily between individuals with lower identity scores based on their mtDNA HVR2 genetic sequence. We also used frequency distribution plots of pairwise identity scores in conjunction with distance matrices to guide our determination of sequence demarcation criteria. Peaks in these plots indicate pairwise identity thresholds that could lead to multiple ambiguous pairwise identity computations, which is undesirable. On the contrary, the graphs reveal that there are specific levels of sequence identity that must be reached to avoid ambiguity in comparisons. This is seen in the troughs of the plots, which indicate the desired outcome of minimal ambiguity. The SDT matrix analysis highlights the distinctiveness of individuals such as Daura15, Koza3, Daura11, Sandamu3, and Sandamu22 from the rest of the study population. These

individuals exhibit lower scores when compared to others, indicating that they can be clearly distinguished based on their mtDNA HVR2 sequence. The overall lower pairwise identity values seen in the SDT demonstrate a high level of genetic diversity among the individuals in the study population. These results support previous scientific findings that have reported a high level of genetic diversity in African populations, as evidenced by studies conducted by [28], [164], [170], [179], [222].

The admixture analysis has uncovered a certain degree of differentiation within the population. The results from the STRUCTURE analysis have shown significant variations that are apparent at the individual level, leading to the identification of substructures that are not exclusive to any single population group within the study. This indicates that genetic differences can be better appreciated when examined on an individual basis rather than as a whole population. This is in line with previous genetic structure studies conducted on West African populations that have identified substructures within, rather than between, the populations (Adeyemo et al., 2005), further supporting this notion. To determine the most likely number of clusters in our data, we utilised the web-based tool Structure Harvester (Earl & vonHoldt, 2012) to analyse the output files generated by the Bayesian clustering program STRUCTURE. The Evanno method (Evanno et al., 2005) was applied to calculate different statistics, which involved examining the log probability of the data across 10 successive K values, the study aimed to evaluate the rate of change. The optimal number of clusters was identified based on this analysis, representing the most informative admixture run. The results indicated that run number 3, with three inferred ancestors, was the optimal population admixture model. While the population substructure pattern between populations appeared similar, distinct individuals could be observed within each population. This is possibly due to geographical proximity within the study population.

4.5 CONCLUSION AND RECOMMENDATION

In summary, this study has revealed a considerable degree of genetic diversity and notable regional variations in mtDNA frequencies within the Hausa population,

indicating a stratified and structured population. These findings underscore the potential value of mtDNA analysis in identifying individuals within the Hausa group and linking them to particular regions or populations, which could prove useful in forensic investigations. Moreover, this research highlights the significance of mtDNA studies in comprehending population structure and evolutionary history in Nigeria and other regions, offering valuable insights into the genetic ancestry and origins of the Hausa people with potential implications for both medical and forensic investigations. Future research efforts should aim to investigate the mtDNA diversity and population structure of other ethnic groups in Nigeria to enhance our understanding of their genetic history and inform forensic investigations. It is also recommended that mtDNA analysis be employed as a supplementary tool in forensic investigations, particularly in situations where traditional identification methods are not feasible. Furthermore, the significance of mtDNA research in unravelling population structure and evolutionary history in Nigeria and beyond should be highlighted, with initiatives to promote and encourage more research in this field.

CHAPTER FIVE

5.1 BACKGROUND

According to research, HVRI exhibits a greater number of polymorphisms, but they occur less frequently in individuals. Conversely, HVRII displays fewer polymorphisms, but with a higher frequency of occurrence. Additionally, The mtDNA control region demonstrates a mutation rate that is at least five-fold higher than the mutation rate observed in the remaining regions of mtDNA, and its elevated level of polymorphism makes it the most widely studied region of the mtDNA molecule [122]. Consequently, it is especially valuable for identification purposes. This holds true for individuals of European descent as well. On average, individuals of European Caucasian descent exhibit around 1.5% nucleotide variability in eight positions of the HVRI and HVRII regions [50], [53]. In terms of mtDNA variability, the variation is 5 to 10 times greater between different races than within a single race. Interestingly, there is greater diversity within the Black race and Asian race, respectively, compared to the Caucasian and Hispanic races. Due to the high levels of mtDNA variability within and between populations, mtDNA polymorphisms can be used to determine population genetic structure, phylogenetic relationships, and patterns of migration and settlement throughout history [50].

The study of mtDNA can be classified into three separate spheres of investigation: medical genetics, which focuses on identifying disease-causing mutations; population genetics, which examines variations between human populations; and forensic genetics, which calculates the likelihood of sample matches. Although individual research area has its specific limitations, they all share a common "allelic" thinking inherited from nuclear markers. However, despite having similar objectives, the advancements accomplished in one field are often disregarded by the others for various reasons such as being published in different journals, belonging to different scientific institutions, and using different terminologies [101]. Recently, all three disciplines have been analysing the same set of molecular markers, including mtDNA, yet researchers have failed to notice the

improvements made in other fields due to a lack of knowledge exchange. This lack of information exchange particularly affects the analysis and interpretation of mtDNA sequence variations[164].

Since the 1980s, population geneticists have been interested in analysing mtDNA sequences, but it was the medical field that first sequenced the coding region, despite early attempts being partially flawed. Nonetheless, this early work led to important discoveries, such as the identification of the second major Eurasian superhaplogroup, N [223]. The importance of mtDNA testing for identification purposes and the analysis of heavily degraded biological samples and telogen hair shafts was realized by forensic geneticists in the 1990s. This realization coincided with the introduction of the polymerase chain reaction (PCR), which marked a new era for this field. [164]

The classification of Mitochondrial DNA (mtDNA) haplogroups (HG), commonly known as mtDNA haplogrouping, plays a significant role in the field of population genetics. HG assignment is typically done through the use of an automated software tool that utilises the Phylotree [69], a representation of the worldwide mtDNA variations, to determine the HG of a given mtDNA sequence [218]. The main haplogroups have been designated in alphabetical order, from A to Z, based on the order in which they were identified. The theoretical ancestral figure, often called 'Mitochondrial Eve', is commonly associated with the hypothetical root from approximately 120,000 to 156,000 years ago. While she is not the first or only woman of the species, she is the individual from whom all living humans today are believed to have descended maternally. Mitochondrial Eve is associated with a particular haplogroup [50].

Mitotyping, the testing of human mitochondrial haplotypes, is particularly useful in forensic contexts [224]. Although next-generation sequencing (NGS) has gained popularity in various fields, STS on capillary electrophoresis machines is the most commonly utilised technique for analysing forensic samples. Included in the sample pool are historical or ancient skeletal and dental remains, hair shafts without roots, and biological materials that have encountered extreme environmental conditions.

Additionally, this method is commonly used for investigating maternal lineage or identifying genetic information from haplogroup assignment [224], [225]. The mtDNA's resilience in forensic samples is attributed to its circular organization and high copy number per cell in comparison to the nuclear genome. In some forensic microscopic hair examinations, mtDNA sequencing can be used to complement the analysis, especially when excluding potential sources [213], [224].

In the past, forensic mtDNA analyses using Sanger sequencing were mainly focused on the control region (CR), which contains the mitochondrial origin of replication and transcription and is approximately 1122 bp in length. The noncoding portion of the mitochondrial genome (mtGenome), spanning approximately 16,569 bp, has attracted significant attention in research due to its unique characteristics and the difficulties associated with generating sufficient sequencing coverage using chain termination chemistry. However, the advent of next-generation sequencing (NGS) or massively parallel sequencing (MPS) has revolutionized the field, allowing for routine analysis of complete mtGenomes and making it more accessible [226]. With complete mtGenome data at hand, research endeavours can be elevated, and the accuracy of haplogroup assignments can be significantly improved in comparison to using data solely from the control region [224].

Archaeogenetics, the use of genetic information to understand human history, is becoming increasingly significant. The combination of phylogenetic and geographic evidence, known as phylogeographic analysis, relies heavily on non-recombining mtDNA and Y chromosome markers, although it is most effective when used with other evidence within a model-based framework [33]. In both haploid marker systems, the incorporation of a molecular clock-based time frame is a crucial component of the phylogeographic approach, allowing lineage age and dispersal times to be estimated without assuming a demographic scenario [125]. In the past 40 years since the original publication of the Cambridge Reference Sequence, research study of human mtDNA phylogenetics has revealed an emerging pattern, indicating that non-African lineages

form a small portion when compared to the more extensive and diverse African phylogeny, which has deeper roots. The idea of ancient African origins and the subsequent migration out of Africa, estimated to have taken place roughly 70-50 thousand years ago, has received substantiation from diverse sources of evidence, including Y-DNA, autosomal DNA, archaic introgression, and African fossils [136]. While the coalescence time of uniparental mtDNA and Y-DNA loci is estimated at 250-150 kya, evidence from various sources, such as archaeology, nuclear DNA, and genome-wide studies, raises the possibility of an origin for modern humans that predates the previously suggested timeframe, ranging from 500 to 300 thousand years ago, with the actual age potentially falling within either of these ranges. [136]

Over the past ten years, mtDNA has been widely employed in unravelling the chronology and routes of human migrations, encompassing periods from the later prehistoric times to documented historical events [164]. The unique inheritance properties of mtDNA, which does not undergo recombination, allow for the classification of sequence haplotypes into distinct monophyletic clades called haplogroups [45]. The members of a haplogroup share a common ancestor and have a specific sequence motif. By adopting a phylogeographic standpoint, it has become evident that haplogroups can exhibit connections with wide-ranging geographical areas and, on occasion, with distinct ethnic populations. Advances in the analysis and interpretation of complete mtDNA genomes [95], [139], [151] and entire coding-region sequences [23], [79] have greatly enhanced our understanding of the mtDNA phylogeny in recent years [164].

Mitochondrial DNA is also an important tool in deciphering disease evolution at either individual or population level. Mitochondria do not originate from scratch, but instead form by growing and dividing from pre-existing ones during each cell division. The replication of mtDNA is tightly regulated to ensure a consistent supply of mtDNA to the cell. Inherited diseases linked to mtDNA mutations are maternally inherited and affect all offspring, regardless of gender. This is because the mitochondria present in ovarian cells (which contain roughly 200,000 mtDNA molecules in mammals) are derived from a

distinct group of mtDNA molecules. Studies of early embryo development have revealed that the female primordial germ line contains just a small number of mtDNA molecules [58]. Mitochondria are responsible for producing energy. When mitochondrial dysfunction occurs, it often leads to ineffective oxidative phosphorylation and the production of excessive reactive oxygen species, which can cause changes in cellular function and signalling. Such dysfunction is associated with a range of diseases such as metabolic disorders, neurodegenerative disorders, autoimmune diseases, and cancer, as well as aging. The cause of dysfunction is believed to be linked to variations in the mitochondrial genome, which has become a crucial research focus for complex diseases and precision medicine [34]. The objective of this chapter is to highlight the genetic variation present in the Hausa population by focusing on HVR I and II.

5.2 METHODOLOGY

The Methods have been extensively treated in previous chapters. Please refer to previous chapters for detailed methodology. However, we obtained mitogenome sequences from the NCBI database for different global populations to compare them with our genetic data. The populations we downloaded sequences from include Angola in Southern Africa, Tokelau population in Oceania (New Zealand), west Indian Caste in South Asia, Switzerland in Central Europe, Paraguay in South America, and Canada (Newfoundland) in North America. The accession numbers for these sequences are as follows: MF3812871-MF3813061, MT9282831-MT9282971, MK0439671-MK0439862, MT0790191-MT0790371, MH9818231-MH9818421, and MF5887941-MF5888111, respectively. Using the MEGA v.11 software, we employed the Neighbour-joining method to construct a phylogenetic tree. The Jukes-Cantor model was used, and 1000 bootstrap replications were performed for reliability [220]. Additionally, we compared the mean genetic distance between the global populations using MEGA11 V.11.

5.3 RESULTS

Extracted DNA quality check based on Nanodrop spectrophotometry. The Nanodrop had provided us with information regarding the concentration, purity and quality of the extracted DNA. This is shown either graphically (fig. 5.1) or in table 5.1.

Table 5.1: Nanodrop Spectrophotometry Showing the Concentration and Purity of the Extracted DNA

Sample ID	Concentration	A260/A280	A260/A230	A260	A280
1	32.5	1.92	1.27	0.65	0.34
2	9.9	1.79	1.46	0.2	0.11
3	45	1.69	1.1	0.9	0.53
4	23.6	1.78	0.94	0.47	0.26
5	26.5	1.76	0.95	0.53	0.3
6	22.5	1.98	1.73	0.45	0.23
7	22.2	1.73	1.15	0.44	0.26
8	107.5	1.41	0.62	2.15	1.53
9	11.3	1.65	0.98	0.23	0.14
10	13.1	1.8	1	0.26	0.15
11	30.3	1.74	1.07	0.61	0.35
12	29.3	1.93	1.07	0.59	0.3
13	42.3	1.84	1.38	0.85	0.46
14	8.1	1.69	0.94	0.16	0.1
15	106.6	1.63	0.88	2.13	1.3
16	13	1.63	1.06	0.26	0.16
17	40.5	1.91	2.16	0.81	0.42
18	32	1.87	1.75	0.64	0.34
19	18.7	1.87	2.45	0.37	0.2
21	60.2	1.62	0.94	1.2	0.74
22	17.8	1.84	1.18	0.36	0.19
23	44.1	1.71	1.18	0.88	0.52
24	28.4	1.87	1.2	0.57	0.3
25	16.5	1.66	1.94	0.33	0.2
26	19.6	1.56	0.7	0.39	0.25

Sample ID	Concentration	A260/A280	A260/A230	A260	A280
27	61.2	1.64	1.06	1.22	0.74
28	25.8	1.56	0.78	0.52	0.33
29	43.8	1.55	0.76	0.88	0.57
30	30.6	1.82	1.67	0.61	0.34
31	8.6	1.71	1.07	0.17	0.1
32	23.5	2.05	0.81	0.47	0.23
33	35.6	1.97	1.6	0.71	0.36
34	5.2	1.82	2.1	0.1	0.06
35	34.4	1.77	1.26	0.69	0.39
36	22.2	1.82	1.05	0.44	0.24
37	26.1	1.97	1.16	0.52	0.27
39	28.2	1.82	1.5	0.56	0.31
40	13.4	1.77	1.25	0.27	0.15
41	23.2	1.6	0.96	0.46	0.29
42	33.3	1.87	2.44	0.67	0.36
43	32.3	1.8	1.19	0.65	0.36
44	87.2	1.69	1.02	1.74	1.03
45	21.2	1.91	1.05	0.42	0.22
46	28	1.53	0.69	0.56	0.37
47	20.9	1.79	1.72	0.42	0.23
48	19	1.68	0.89	0.38	0.23
49	35.2	1.96	1.57	0.7	0.36
50	18.5	1.79	1.33	0.37	0.21
51	50.1	1.58	0.85	1	0.64
52	29.8	1.82	0.82	0.6	0.33
53	12.1	1.84	0.87	0.24	0.13
54	30.7	1.89	1.4	0.61	0.33
55	13.3	1.8	1.26	0.27	0.15

Sample ID	Concentration	A260/A280	A260/A230	A260	A280
56	29.3	1.79	1.53	0.59	0.33
57	33.7	1.84	1.26	0.67	0.37
58	5.7	1.72	1.67	0.11	0.07
59	30.1	1.98	1.47	0.6	0.3
60	9.3	1.56	0.71	0.19	0.12
61	3.5	1.39	0.51	0.07	0.05
62	45.6	1.83	1.23	0.91	0.5
63	25.1	1.64	0.77	0.5	0.31
64	3.6	1.75	0.96	0.07	0.04
65	7.2	1.44	2.54	0.14	0.1
66	5.1	1.88	2.9	0.1	0.05
67	24.9	1.84	1.92	0.5	0.27
68	59	1.91	2.8	1.18	0.62
69	22.4	1.84	1.56	0.45	0.24
70	20.6	1.76	1.48	0.41	0.23
71	38.4	1.8	1.83	0.77	0.43
72	33	2.02	1.83	0.66	0.33
73	35.4	1.97	1.65	0.71	0.36
74	14.9	1.88	0.79	0.3	0.16
75	39.5	1.85	1.56	0.79	0.43
76	21.9	1.64	1.03	0.44	0.27
77	28.8	1.66	0.93	0.58	0.35
78	9.3	1.86	0.64	0.19	0.1
80	60	1.8	1.49	1.2	0.67
82	31.2	1.74	1.13	0.62	0.36
83	26.1	1.75	1.14	0.52	0.3
84	21.8	1.81	0.82	0.44	0.24
85	30.9	1.73	1.28	0.62	0.36

Sample ID	Concentration	A260/A280	A260/A230	A260	A280
86	28.7	1.88	1.6	0.57	0.3
88	28.6	1.66	0.87	0.57	0.34
89	13.9	2.16	0.7	0.28	0.13
90	4.1	2.64	0.57	0.08	0.03
91	24.2	1.78	1.29	0.48	0.27
92	20	1.84	4.72	0.4	0.22
93	11.8	1.89	1.13	0.24	0.13
94	11.1	1.72	1.12	0.22	0.13
95	3.2	2.53	1.12	0.06	0.03
96	38.8	1.81	1.22	0.78	0.43
97	3.6	1.47	0.51	0.07	0.05
98	19.2	1.96	1.52	0.38	0.2
100	20	1.9	1.48	0.4	0.21

Figure 5.1 is showing the plot of absorbance against wavelength which can also be used as an indicator to determine the purity of our DNA sample. A peak against 260nm wavelength depicts a sign of purity.

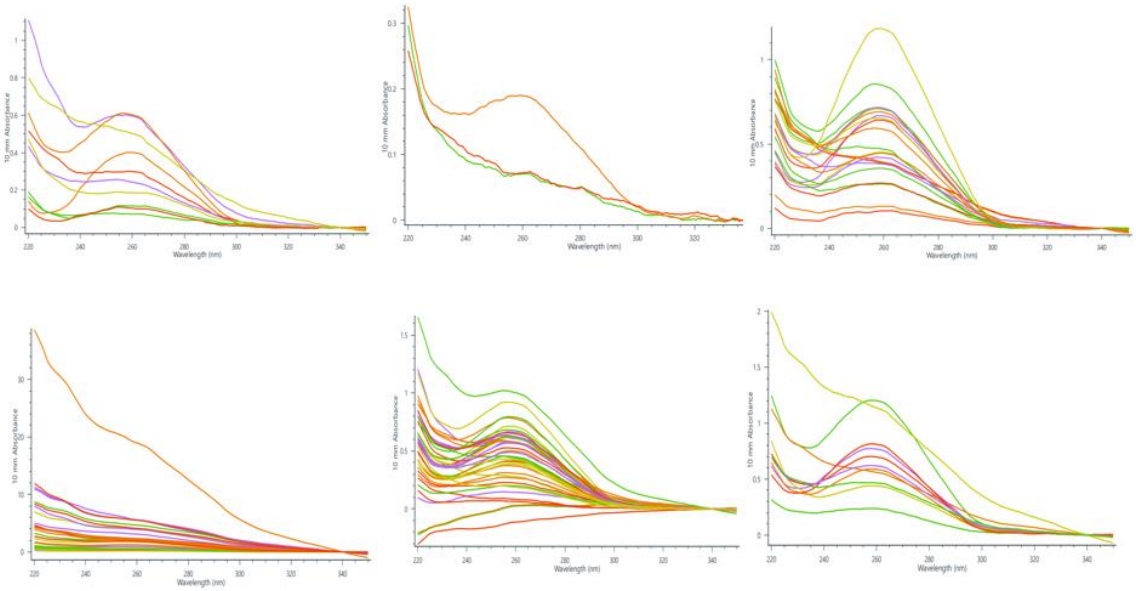
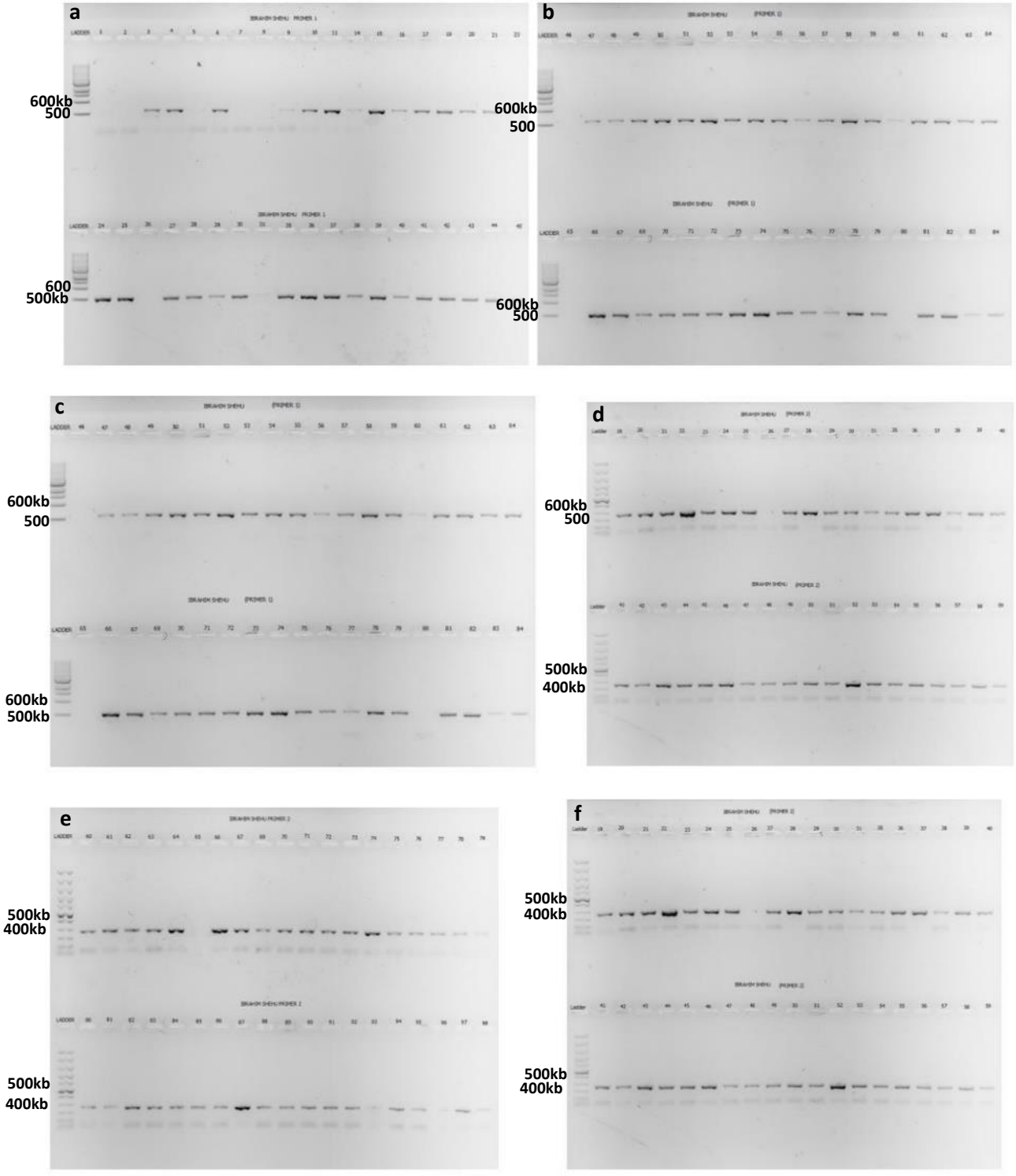


Figure 5.1: A spectra of Absorbance against Wavelength Guiding the Estimation of Concentration and Purity of our Extracted DNA

Further to the determination of the concentration of the extracted DNA, the control regions of the mtDNA of the template DNA was amplified using specific primers as described previously. Gel electrophoresis was conducted thereafter to estimate the size of the PCR amplicons. The observed sizes are shown in figure 5.2.


a
600kb
500
600
500kb
b
600kb
500
600kb
500
c
600kb
500
600kb
500kb
d
600kb
500
500kb
400kb
e
500kb
400kb
500kb
400kb
f
500kb
400kb
500kb
400kb

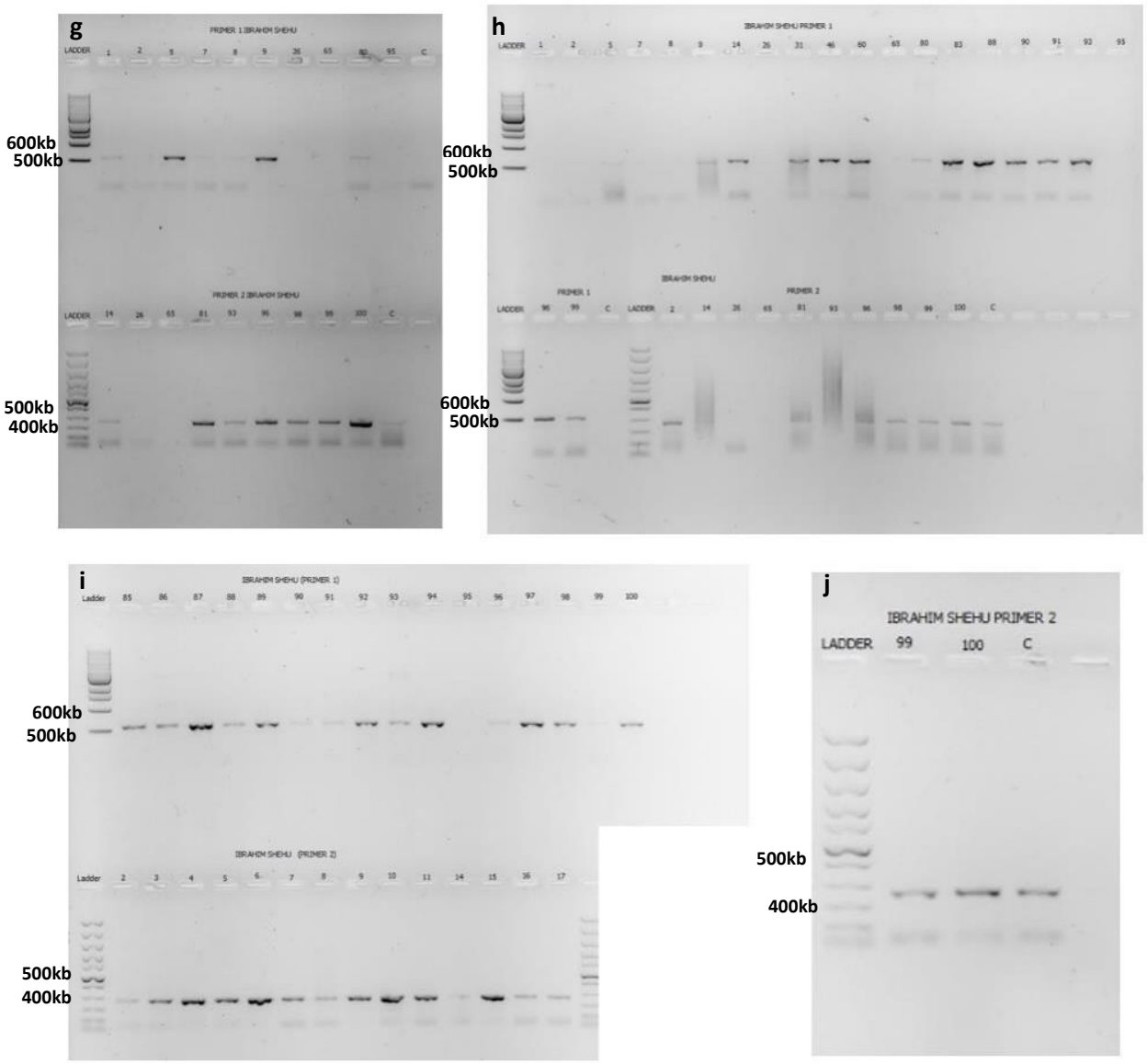


Figure 5.2 (a to j): Gel bands of the PCR amplicons for size estimation before the Big Dye Terminator STS. The figures a to j represent the gel bands for both primers 1 and 2.

Following the size estimation, big dye terminator sanger sequencing was used to sequence the amplified mtDNA control regions. Fig.5.3 shows the window interface of the Bioedit software that was used to do base calling and sequence alignment as previously described.

Figure 5.3: BioEdit software interface for base calling and multiple sequence alignment using ClustalW

5.3.1 POPULATION GENETIC STRUCTURE OF THE HVR2 FOR FORENSIC AND POPULATION GENETIC REFERENCE

The MRCA and macrohaplogroups were also counted after assignment using haplotracker. The macrohaplogroup distribution among the Hausa ethnic population from Nigeria are depicted in figure 5.4. L3 and L2 are the most observed macrohaplogroup

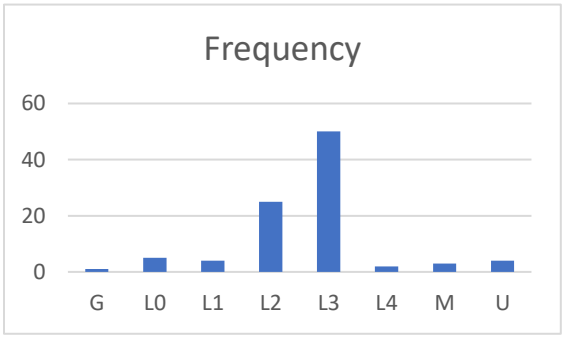


Figure 5.4: Bar plot presenting the Macrohaplogroup distribution among the study population

The MRCA is shown in figure 5.5 displayed L3 and L'2'3'4 as the most shared MRCA. There are others that have least percentage appearance as low as 1% to 3%.

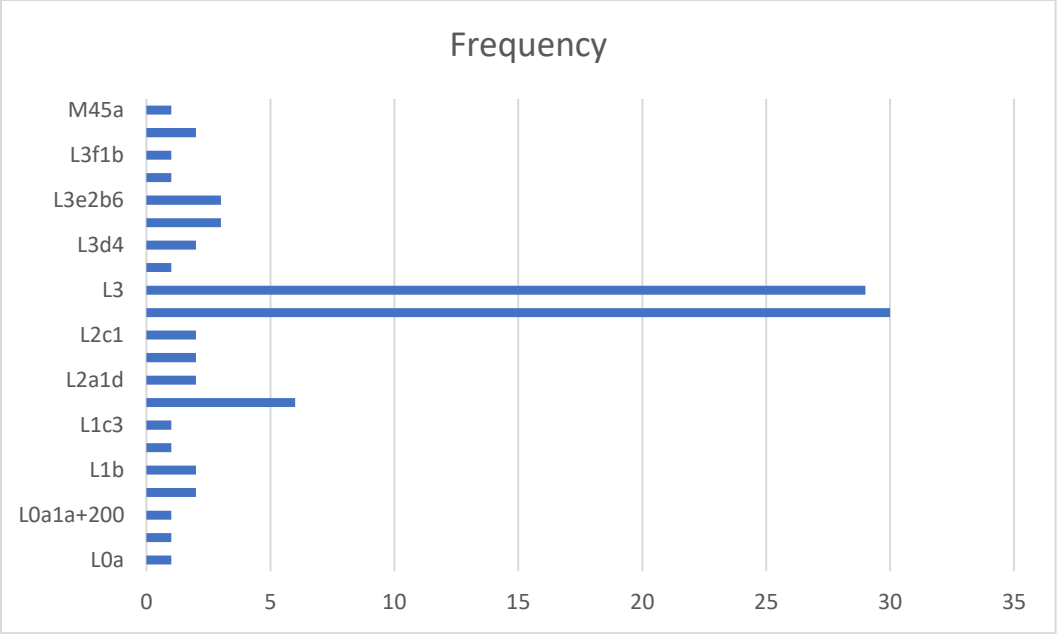


Figure 5.5: Frequency Distribution of the Most Recent Common Ancestor among the study population

The method of Neighbour-Joining[227] was used to deduce the evolutionary history of the analysed population, and the consensus tree was formed by running 1000 replicates. In cases where a partition did not appear in at least 50% of the bootstrap replicates, the corresponding branches were consolidated. Utilising the Jukes-Cantor method [220], the evolutionary distances were computed, representing the number of base substitutions per site. The analysis included 94 nucleotide sequences, with all ambiguous positions removed using pairwise deletion. The final dataset contained 443 positions. MEGA11 [206] was used to conduct the evolutionary analyses (figure 5.6). The Neighbour-Joining Tree (figure 5.7) has posited the West African population on a single branch that show no any sign of direct branch relationship with the rest of the global populations. Surprisingly, Southern Africa is placed as the most distant population to our population dataset. West African population is shown to have closest relationship with the people of Oceania.

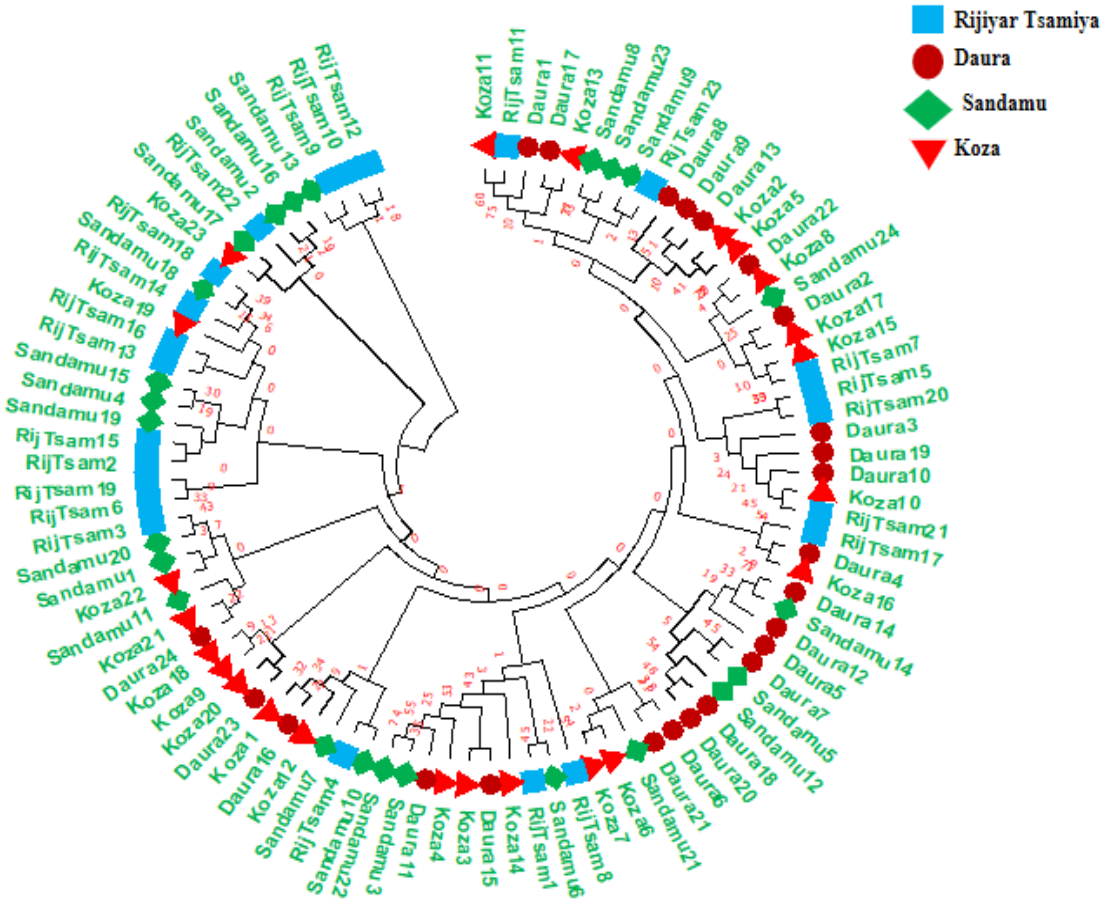


Figure 5.6: Phylogenetic tree/ NJ Tree generated using Bayesian Clustering Analysis to show Ancestral and Sequence Relatedness Between the Study Population

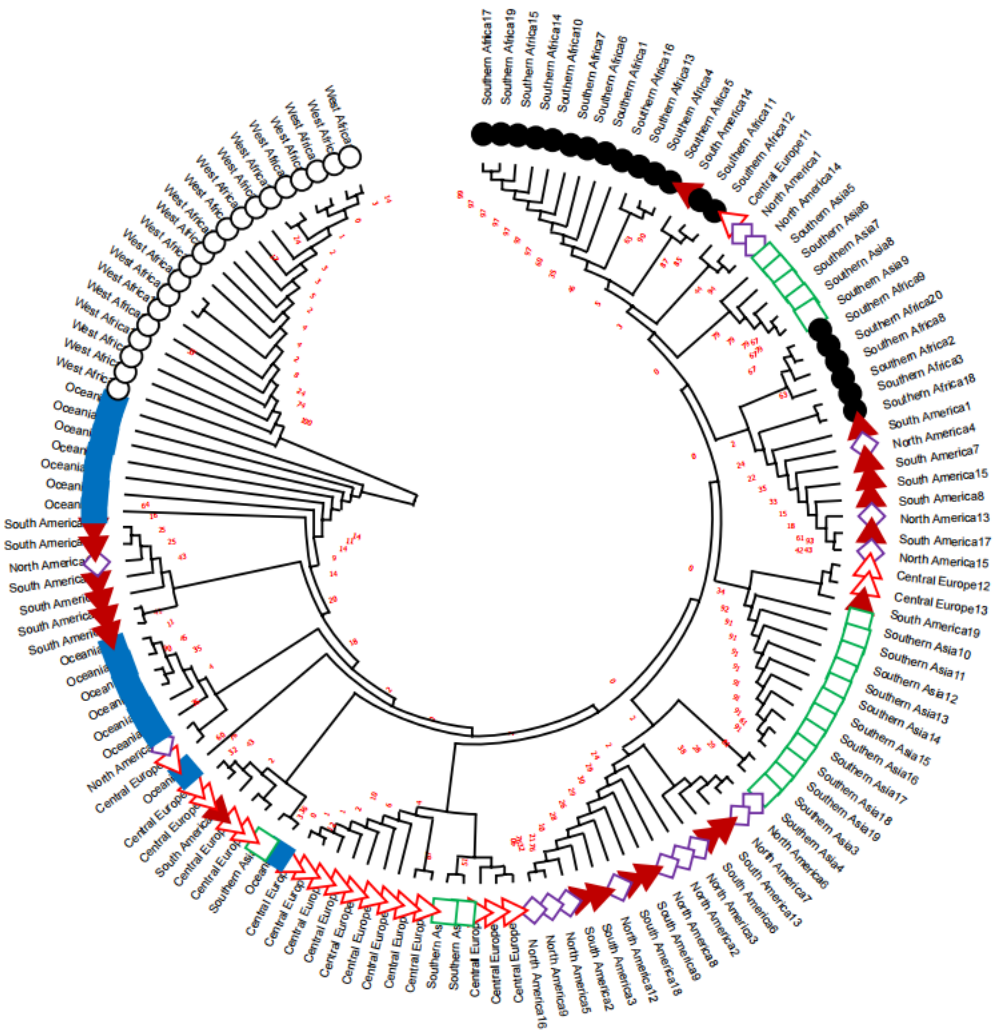


Figure 5.7: Phylogenetic tree/ NJ Tree generated using Bayesian Clustering Analysis to show Ancestral and Sequence Relatedness Between Global Populations

The tree started from Rijtsam12 and ended at Koza11, based on the depicted picture, there are 15 major sub-branches. The primary branch is represented by only four individuals, three from Rijiyar Tsamiya village and one from Sandamu. The branches spread with varying nodes and number of individuals.

To further support the diversity observed in the Neibour Joining Tree, Tajima's neutrality test was done to study the evolutionary history of the population to understand certain demographic events such as migration and population size changes. Tajima's D shows

that there are 164 segregating sites (S), and the proportion of segregating sites (ps) is 0.3702031. The nucleotide diversity (π) is 0.040383, and Θ, which is a measure of the population mutation rate, is 0.0723. The Tajima test statistic (D) is -1.4849, which suggests an excess of low-frequency variants in the dataset.

We also studied population structure by computing genetic differentiation using MEGA11,

Table 5.2: Diversity indices comparing the population subgroups under study

Diversity index	Genetic differentiation	Standard Error
Mean diversity within subpopulations	0.05	0.01
Mean inter-population diversity	0.0	0.0
Mean diversity within entire populations	0.05	0.01
Co-efficient of differentiation	0.02	0.01
Overall average distance	0.05	

The mean distance matrix between the population was also studied and shown in table 5.3. The highest correlation matrix values were observed when comparing Koza and Sandamu to the Daura with a value of 0.553, while the lowest genetic variation was observed between Rijiyar Tsamiya and Sandamu (0.447).

Table 5.3: Distance matrix and their respective p values comparing the population subgroups

	Daura	Koza	Sandamu	RijTsam
Daura		0.0059	0.0060	0.0061
Koza	0.0553		0.0058	0.0057
Sandamu	0.0553	0.0531		0.0056
RijTsam	0.0499	0.0473	0.0447	

The Pairwise genetic distance between the population groups using MEGA11 has shown high diversity indices between the West African and other continental population groups, while presenting lower values between the rest of the population groups. The upper part of the matrix is showing the standard error, while the lower values indicate the pairwise genetic distance (table 5.4).

Table 5.4: Pairwise genetic distance between the Global populations

Continent	West Africa	Oceania	South Asia	Central Europe	South America	North America
West Africa		0.0981	0.0990	0.0997	0.0987	0.0982
Oceania	0.9960		0.0053	0.0042	0.0046	0.0050
South Asia	1.0102	0.0211		0.0035	0.0038	0.0038
Central Europe	1.0093	0.0160	0.0146		0.0031	0.0036
South America	1.0135	0.0211	0.0188	0.0164		0.0035
North America	1.0077	0.0204	0.0174	0.0155	0.0164	

Another comparison within the population is also shown in table 5.5 with highest and lowest genetic differentiations observed within Daura and Rijiyar Tsamiya respectively.

Table 5.5: Population Genetic Differentiation and Respective Standard Error of Mean

Population	**d**	**se**
Daura	0.056	0.0065
Koza	0.054	0.0059
Sandamu	0.051	0.0063
RijTsam	0.037	0.0055

The mean genetic differentiation between the population groups is presented in table 5.6, it is shown that there is low genetic diversity within the groups with highest and least diversity observed in West African and Central European populations respectively.

Table 5.6: Population Genetic Differentiation and Respective Standard Error of Mean between the Global populations

Continent/Subcontinent	Genetic Differentiation	SE
West Africa	0.038429065	0.005876669
Oceania	0.010984271	0.003410924
South Asia	0.013238778	0.00342992
Central Europe	0.009449576	0.002149377
South America	0.018672455	0.003598349
North America	0.013818263	0.003557655

We used PopArt and created a haplotype network in order to understand the haplotype diversity of the population. the network. The network is used to show ancestral relationship and mutations over time. It can be observed that the network is somewhat clustered together indicating a degree of connectedness between the individuals. There are no evidences of shared ancestry, because almost all the nodes have similar size. However, some individuals can be observed budding off to the edge of the tree indicating some form of haplotype variation from the rest of the population. Interestingly, RijTsam11, Koza11 and Daura1 maintained a relationship as previously observed in the Neighbour Joining tree. Other relationships observed in NJ Tree are inconsistent with the haplotype diversity observed in figure 5.8.

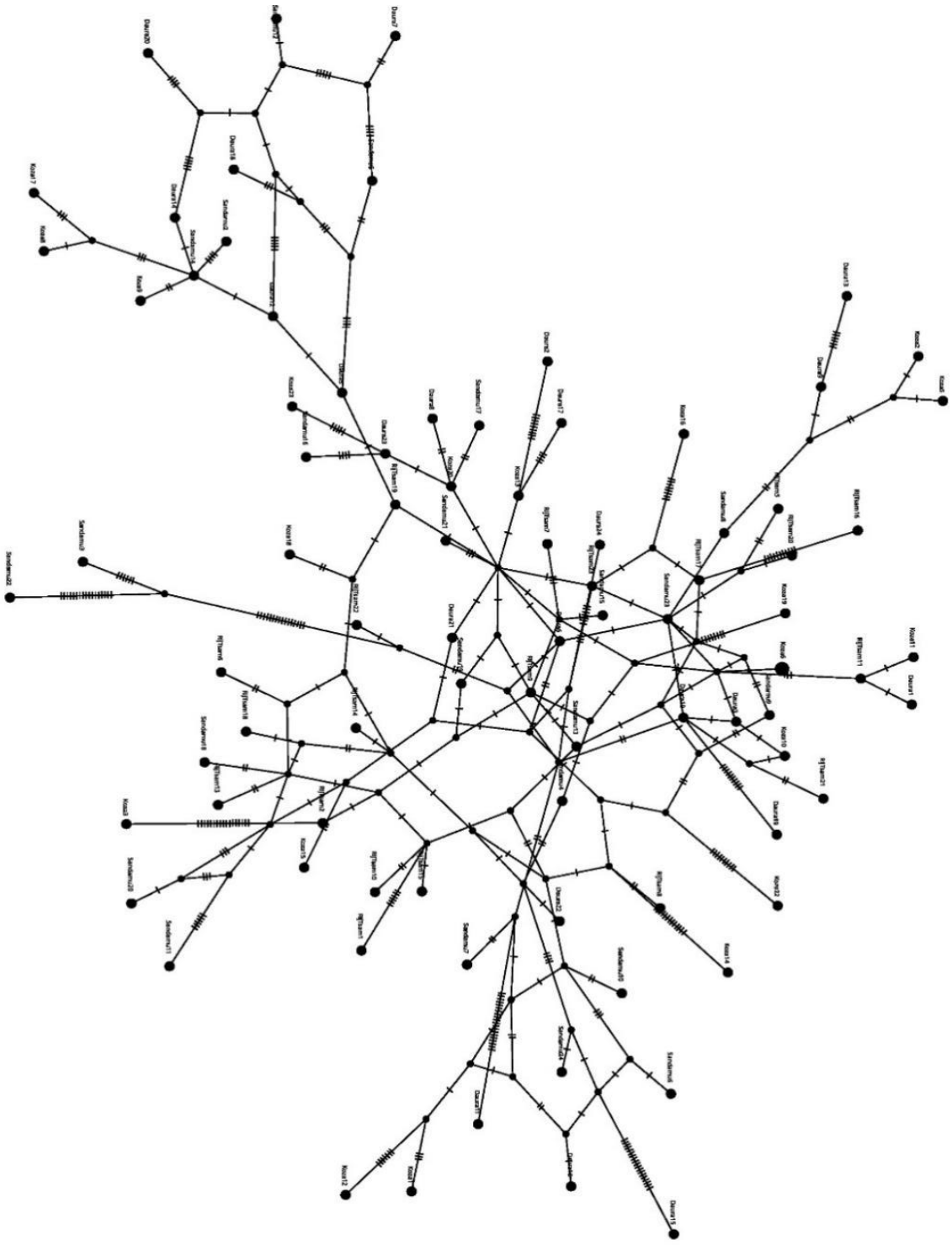


Figure 5.8: Haplotype Network using Maximum Likelihood to show haplotype and sequence relationship among the study population

5.4 DISCUSSION

Nanodrop is a common spectrophotometer used to measure the concentration and purity of DNA samples. After DNA extraction, we used Nanodrop one to determine the amount and quality of DNA extracted. The Nanodrop measures the absorbance of light by the DNA sample, which can be used to calculate the concentration and purity of the DNA. The absorbance is measured at two wavelengths: 260 nm and 280 nm. The 260 nm absorbance is proportional to the amount of DNA in the sample, while the amount of protein contamination present in the sample can be estimated by observing the absorbance at 280 nm, which exhibits a proportional relationship. When interpreting the Nanodrop results, there are a few things that we kept in mind: Concentration, Purity and Quality. The DNA concentration was reported in ng/µl. A high concentration is desirable for downstream applications, but concentrations that are too high can cause problems such as inhibition of enzymatic reactions. A typical concentration range for DNA is 20-200 ng/µl which conformed with what we got during the spectrophotometry. The purity of the DNA can be assessed by the 260/280 and 260/230 ratios. A pure DNA sample should have a 260/280 ratio of around 1.8, while a ratio of less than 1.8 indicates protein contamination, and a ratio greater than 1.8 suggests the presence of RNA. The 260/230 ratio should be greater than 1.8, indicating the absence of contaminants such as salts or organic solvents. These parameters were also met before we did the downstream applications. However, some of the samples were run downstream after we failed to achieve the desired purity which could be reason why we were able to sequence only 93 and 94 individuals using primers 1 and 2 respectively. The Nanodrop results can also provide an indication of the quality of the DNA. High-quality DNA will have a sharp peak at 260 nm and a smooth curve without a shoulder or hump. Low-quality DNA may have a low 260/280 ratio, a peak at 230 nm, or a shoulder or hump in the 260 nm curve as we observed in very few samples.

Generally, the Nanodrop results provided us with valuable information about the concentration, purity, and quality of DNA extracted from our sample. However, it is

important to interpret the results in the context of the specific experiment or application being performed.

Further to the Nanodrop, we amplified the targeted non-coding regions via PCR. The PCR amplicons' size were estimated through gel electrophoresis. The average number of base pairs observed for primers 1 and 2 were 500 and 400 respectively. The size was estimated using a 1kba DNA ladder as depicted on the previous figure, confirmation was done after sequencing the PCR amplicons.

Our findings on the HVR2 haplotyping and genetic diversity is consistent with established findings on the most observed L maternal haplogroup that is predominantly found in the African population. the L haplogroup is the most ancient that is believed to have emerged in Africa about 150kya. Other non-L haplogroup were also observed, U, M, and G were observed in few individuals, The U haplogroup is primarily found at varying frequencies in Europe, Central Asia and Middle East, it is however not surprising to find some of these in minute frequencies in West African population due to history of slave trade and spread of Islam from the Middle East into the region. The M haplogroup is found in lower frequencies in East Africa, Pacific Islands and the Middle East [23], [140], [228]. The G haplogroup observed can also be attributed to some historical relationship that the Hausa people share with the people of North Africa [214].

The NJ tree has further elucidated the diversity of the African population. The study population has shown relationship within the subpopulation. There is evident population admixture between the individuals in the study population irrespective of the city where they came from. The least diverse individuals are RijTSam9, 10, and 12 while the most diverse individuals are Daura1, RijTsam11 and Koza11. To further examine the population structure, we conducted genetic differentiation analyses at different parameters, the average genetic diversity index was recorded at 0.05 indicating a moderate diversity, the value observed is due to the fact that the mtDNA region of interest is an evolutionary conserved non-coding region. Moreover, subpopulation comparison further reveal that the majority of the genetic distinction observed is within

the study population and thus indicating that the individuals can be segregated based on their genetic sequence, but we cannot tell which population subgroup they belong to.

The NJ tree for the global population comparison has shown the West African as the most diverse and distant to all other populations under comparison. surprisingly, the Angolan population appeared most distant to our population dataset, this could be due to possible admixture between the Angolans and the Portuguese. The people of Newzealand are also place at proximal position to the West African population, while South American and South Asians were placed close to the Southern African population. The Pairwise genetic distance between the population groups has shown that the West African populations can be distinguished from the rest of the global population based on their genetic sequence. However, lower values observed between the rest of the populations indicate relative similarity in genetic sequence between them. Moreover, the genetic differentiation values have shown that the West African population are the most genetically diverse population. On the other hand, the rest of the populations have shown relatively lower differentiation values, indicating lower genetic diversity within and between the populations.

The haplotype diversity network visualized has further stem the genetic distinctiveness of the population under study, it has shown a diverse network that is devoid of excessive genetic similarity. It is however evident that there is a degree of genetic and haplotype relationship between the study population. In some instances, the haplotype consistently connected some of the individuals that were put genetically close to one another on the Neighbour Joining tree. Meanwhile, some of the individuals that were otherwise genetically close in NJ Tree appear distant in the haplotype network.

5.6 Conclusion

Based on your study on the HVR2 of mtDNA among the Hausa population of West Africa, it can be concluded that there is significant genetic diversity within this population, as evidenced by the Neighbour-Joining tree and haplotype network. The

genetic differentiation of 0.05 suggests that there are distinct genetic differences between the subpopulations of Hausa individuals studied.

The observed diversity among the Hausa population may have arisen from various factors, including historical migrations, genetic drift, and natural selection. It can be concluded that the population is ideal for forensic reference due to diverse genetic architecture of the individuals.

5.7 Recommendation

Further studies, such as whole mitogenome sequencing, could provide additional insights into the genetic makeup of this population. While our study focused on the HVR2 region of mtDNA, to obtain a more comprehensive insight into the genetic makeup of the Hausa population, exploring additional genetic markers like autosomal DNA or Y-chromosome markers would be advantageous.

CHAPTER SIX

6.1 INTRODUCTION

Understanding the history of human evolution and migration requires the investigation of genetic diversity within populations [34]. This study focuses on analysing the population structure and stratification of the Hausa ethnic group through the use of HVR1 mtDNA analysis. The primary objective is to gain a better understanding of the genetic diversity and degree of population substructure within this group.

One molecular technique used for examining genetic variation in an individual's mitochondrial genome is mtDNA haplotyping [4]. Due to its maternal inheritance and high mutation rate, mtDNA haplotyping is applicable in various fields such as forensic identification, evolutionary research, and medical investigations [5]. Population genetics studies often employ mtDNA to track maternal lineages and analyse population structure and history [196]. The process of mtDNA haplotyping involves extracting mtDNA from a biological sample and sequencing its nucleotide sequence [60].

After extracting mtDNA from a biological sample, its nucleotide sequence is compared to a reference sequence to detect any variations or polymorphisms [218]. These polymorphisms can be combined to form a haplotype, which is a distinct set of mtDNA polymorphisms that are inherited together. Haplotypes are useful for analysing genetic variation within populations and tracing population evolution. Studying genetic variation in the mitochondrial genome can provide insights into population evolution, identify genetic markers linked to certain diseases, and aid in forensic investigations. For instance, mtDNA haplotyping has been employed in tracking maternal ancestry and human migration patterns [8].

Research on African population genetics has been limited, and there is a lack of understanding of the genetic diversity of many populations [144]. Previous research indicates that African populations exhibit a wide range of haplogroups and highly diverse and complex mtDNA variations [175]. Therefore, studies that focus on African

populations are essential in comprehending the genetic relationships and evolutionary history of the continent's diverse populations [203]. Particularly, there is limited information on genetic variation and structure in West Africa, and the region is significantly understudied [229]. The Hausa people are widely distributed in West Africa and are known to possess a diverse genetic makeup [164]. With approximately 70 million individuals, the Hausa ethnic group in Nigeria is counted among the largest ethnic groups in West Africa (National Population Commission, 2007). They are a diverse group that resides in various regions across Nigeria and neighbouring countries [164].

The genetic diversity and evolutionary history of the Hausa people have not received much attention despite their large population. Recently, Titilayo and colleagues (2018) conducted a study to explore the mtDNA variation and population structure of the Hausa population through sequencing and bioinformatics analysis. According to their findings, the Hausa population exhibited a high level of mtDNA diversity, indicating their suitability for forensic genetic reference. The study revealed the presence of several haplogroups and sub-haplogroups within the Hausa population, with the most prevalent being the L0a, L3e, and L3b haplogroups. The researchers also identified population structure and stratification, which may suggest population expansion and selection among the study populations.

We seek to present a comprehensive overview of the genetic structure and diversity of the Hausa ethnic group by examining HVR1 sequences obtained from a sample of 93 unrelated individuals that represent different sub-populations of the group. Our analysis could have significant implications for genetic association studies, forensic investigations, and population health research.

6.2 METHODOLOGY

6.2.1 CONSENT AND ETHICAL CLEARANCE

We recruited individuals who provided informed consent and were aware of the study's purpose. Additionally, we obtained ethical clearance from the Ministry of Health's

Research and Ethics committee in Katsina State, Nigeria (MOH/ADM/SUB/1152/1/558), and obtained consent from local traditional leaders.

6.2.2 SAMPLE POPULATION

After obtaining informed consent, we collected buccal swabs from individuals residing in four different areas located in three Local Government Areas (LGAs) that form the ancient cities of the Daura emirate: Daura, Koza, Sandamu, and Rijiyar Tsamiya. We recruited a total of 100 individuals, and successfully sequenced 93 individuals. To form population samples, we grouped individuals into four categories based on their area of residence: Group 1 (Daura, n=22), Group 2 (Koza, n=24), Group 3 (Sandamu, n=24), and Group 4 (Rijiyar Tsamiya, n=23). Individuals with known common maternal ancestors were excluded from the study. All necessary ethical protocols were followed.

6.2.3 LABORATORY METHODS

The left and right cheeks of participants were gently scraped with two separate swab sticks for 10 seconds, and the swabs were placed in containers with buffer and stored on ice. The QIAamp DNA mini kit was used to extract the entire genome, including mtDNA, in accordance with the manufacturer's protocol (Cat. No. 51304, QIAGEN Heidelberg, Germany). Nanodrop spectrophotometry was employed to evaluate the concentration and purity of the extracted DNA. To target and amplify mtDNA HVS-I, oligo sequence forward and reverse primers were used: Forward; (5'-TTA ACT CCA CCA TTA GCA CC-3') & Reverse; (5'-CCT GAA GTA GGA ACC AGA TG-3'). The primer sequences were checked against the human reference genome using the BLASTn suite of the NCBI to quickly eliminate the possibility of annealing with segments of the nuclear genome.

Polymerase Chain Reaction (PCR) was used for the amplification process, with 35 cycles and specific conditions. The PCR process began with an initial denaturation step at 94°C for 5 minutes, followed by denaturation at 94°C for 30 seconds, annealing at 53°C for 30 seconds, extension at 72°C for 1 minute, and final extension at 72°C for 7 minutes. The reaction mixture, with a total volume of 25µl, was composed of 12.5µl of one Taq Quick-

load 2x master mix with standard buffer, 0.5µl of each forward and reverse primers, 2µl of diluted DNA template, and 9.5µl of nuclease-free water.

To ensure quality control, a negative control was included in each run. Exonuclease I and shrimp alkaline phosphatase (ExoSAP) treatment were used to remove any unincorporated dNTPs and primers that remained in the PCR reaction prior to sequencing. The size of the amplicon was estimated by running agarose gel electrophoresis on a 1% agarose gel. The forward strand was then sequenced using big dye terminator, and fragments were detected and The ABI prism 3500xl genetic analyser (Applied Biosystems, United States) was utilised for estimating the values through capillary electrophoresis.

6.2.4 DATA ANALYSIS

The mtDNA sequence acquired was given accession numbers OQ388355-OQ388447 and was submitted to GenBank. The Haplotracker web tool (www.haplotracker.cau.ac.kr) was employed to determine macrohaplogroups, comparing the submitted data against PhyloTree build17.0, which is based on the revised Cambridge Reference Sequence (rCRS). Base calling and multiple alignments were performed using Bioedit software v.7.2.5 [230], with a bootstrap value of 1000 used for the multiple alignments with ClustalW [219]. The generated files were saved in FASTA format and converted to nexus format using Mesquite v3.7.0. The R package “Ape” was used to transform the fasta file of the mtDNA haploid data into Phylip and structure files [204].

Pairwise identity scores at the individual level were calculated using the SDT v1.2 [209]. MEGA 11 was utilised to calculate the pairwise genetic distance within and between group mean distance [109], [206], as well as Tajima's neutrality test [231]. A haplotype network was generated using TCS software release 1.21 [210], which estimates genetic genealogies from DNA sequences by measuring statistical parsimony [232]. The tcs Beautifier (tcsBU) was used to visualise the network [211]. To infer ancestral relationships, Structure v.2.3.4 was used to study population admixture. The population was divided into four groups, as previously described, by employing a Markov Chain

with 10,000 steps and a burn-in length of 10,000, the analysis was conducted. The K replication values were set to 6 with 3 iterations. To determine the best run for inferring ancestral relationships among the study populations, a structure harvester was employed, which evaluated the 18 runs.

6.3 RESULTS

Figure 6.1 depicts the distribution of macro-haplogroups, where L3 is the most frequent subclade with 43.9%, followed by L2, L1, and L0 with 19.7%, 12.1%, and 7.6% respectively. However, the L4 subclade represents only 3% of the population. In addition to the L macro-haplogroups, a few individuals were found to belong to non-L macro-haplogroups such as G, M, and U.

The pairwise genetic distance was calculated using the Sequence Demarcation Tool (fig. 6.2), which compares the nucleotide sequences of each individual with every other individual in the population. The figure shows that many individuals have over 97% pairwise identity, indicating a close relationship. For example, individual KozaHVS1S4 at position 14 on the vertical axis, belonging to L0 subclade, has over 97% identity with SandamuHVS1 S5&S12, which also shares the L0 subclade. Furthermore, closely related macrohaplogroups exhibit above 90% pairwise identity.

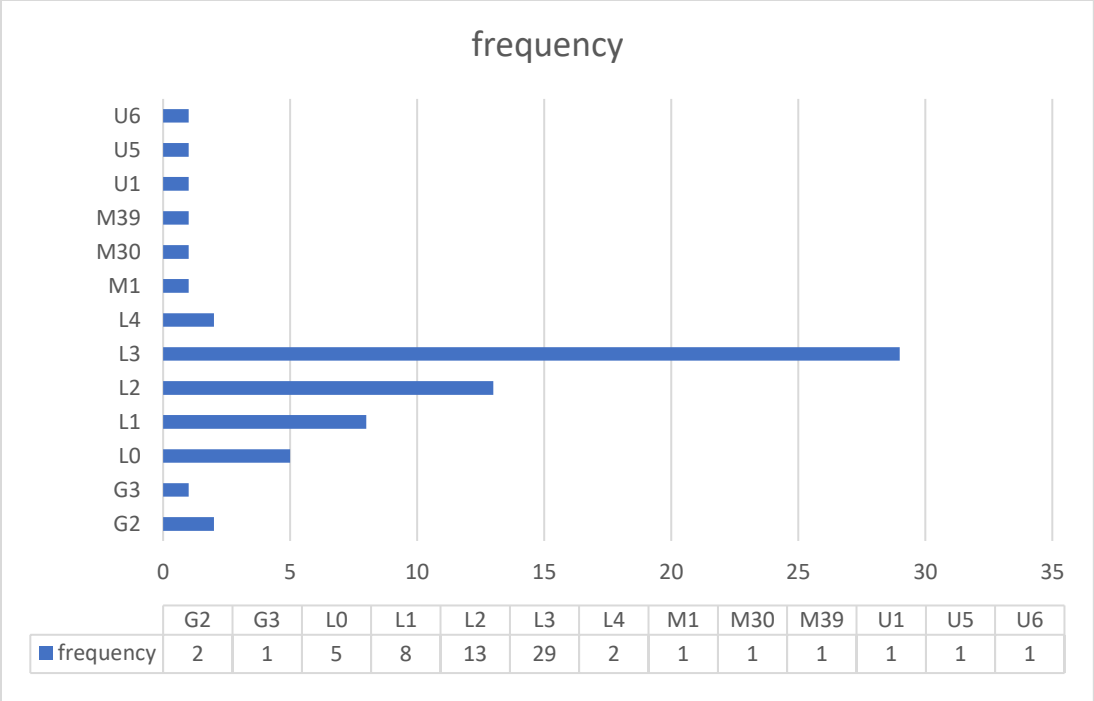


Figure 6.1: Macrohaplogroup Distribution among Hausa Population

Distant macrohaplogroups exhibit a pairwise identity below 68%. Notably, KozaHVS1S5 with L2 subclade shows a unique relationship with all other individuals in the study population, with a distinctive yellow streak indicating below 87% identity with the rest of the population. RijTsamHVS1 S17&S1, KozaHVSS20, whose subclades are G2, L1, and G3, respectively, have the lowest identity score of less than 84% when compared to all other individuals in the study population. In general, the individuals in the study population have high identity scores, with the exception of a few individuals at the bottom of the figure, which indicate below 87% identity with almost all other individuals.

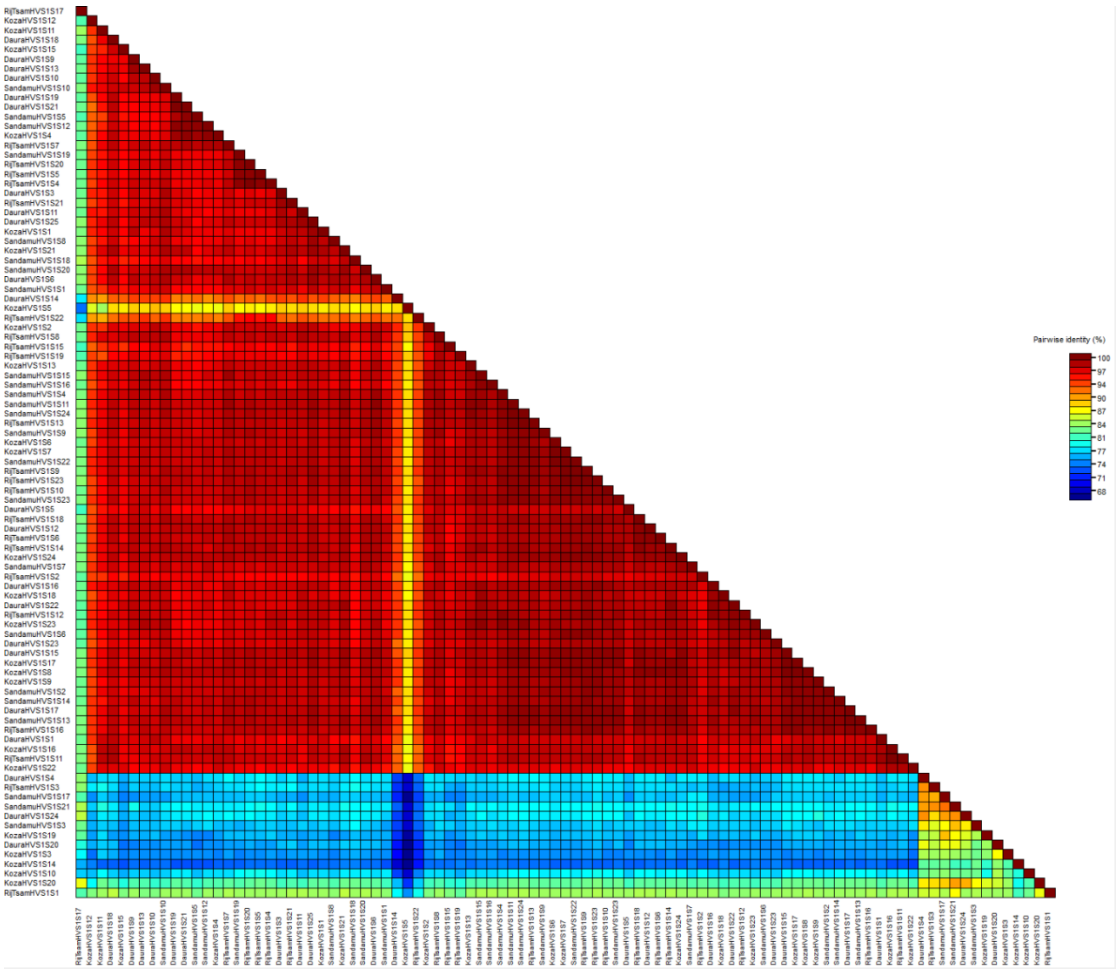


Figure 6.2: Colour-coded matrix of individual pairwise genetic identity among Hausa population

The mtDNA data-derived matrix in Table 6.1 displays the mean genetic distance among four populations: Daura, Sandamu, Rijiyar Tsamiya, and Koza. The findings indicate that the smallest genetic distance exists between Rijiyar Tsamiya and Sandamu, with a score of 0.0825, whereas the greatest genetic distance is between Koza and Daura, scoring 0.1231. The genetic distances between Koza and Sandamu, as well as Koza and Rijiyar Tsamiya, are also relatively high, measuring 0.1178 and 0.1156, respectively.

Table 6.1: mean genetic distance matrix between the population groups

Population	Daura	Sandamu	Rijiyar_Tsamiya	Koza
Daura		0.0064	0.0062	0.0078
Sandamu	0.0923		0.0059	0.0076
Rijiyar_Tsamiya	0.0894	0.0825		0.0074
Koza	0.1231	0.1178	0.1156	

The study population groups' mean genetic distance is illustrated in Figure 6.3. The figure indicates that the highest genetic differentiation value (d) is observed between Koza and the other populations, with a score of 0.15. Daura follows closely behind with a value of 0.10. On the other hand, Sandamu and Rijiyar Tsamiya exhibit lower levels of genetic differentiation, with values of 0.09 and 0.08, respectively. The figure's standard error values (SE) are comparatively low, suggesting that the genetic differentiation estimates are likely to be accurate.

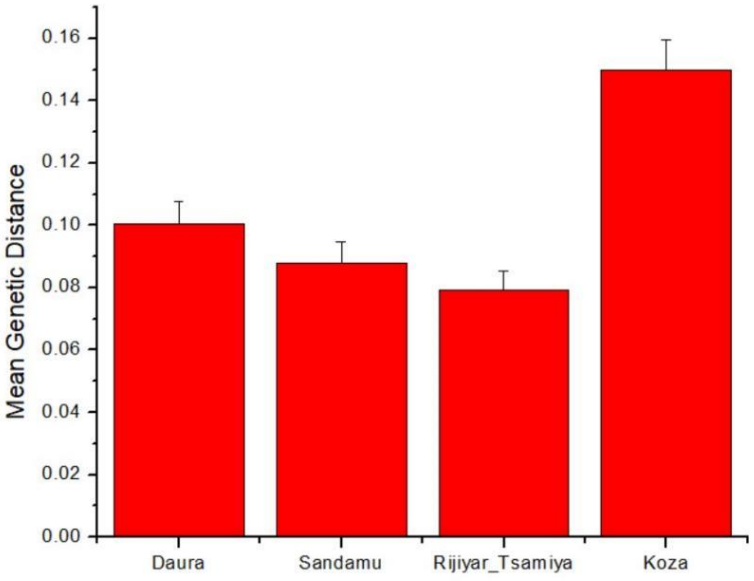


Figure 6.3: Mean genetic distance between the population groups under study

Tajima's D test results reveal that the dataset contains 363 segregating sites (S) and that the proportion of segregating sites (ps) is 0.662409. The nucleotide diversity (π) measures 0.080693, while Θ, a metric of the population mutation rate, stands at 0.129771. The Tajima test statistic (D) is -1.287884, indicating an overabundance of low-frequency variants in the dataset.

Figure 6.4 depicts a haplotype network constructed and visualized by TCS and tcsBU software. The circles represent the haplotypes of each individual, while the lines connecting them represent the evolutionary relationships among the haplotypes. The closer the lines are, the closer the evolutionary relationship between the haplotypes, and the more distant the lines are, the more distant the evolutionary relationship between the haplotypes. The network reveals a considerable genetic distance between the study population, with very few network connections between individuals. The most extensive network connection observed is among 15 individuals.

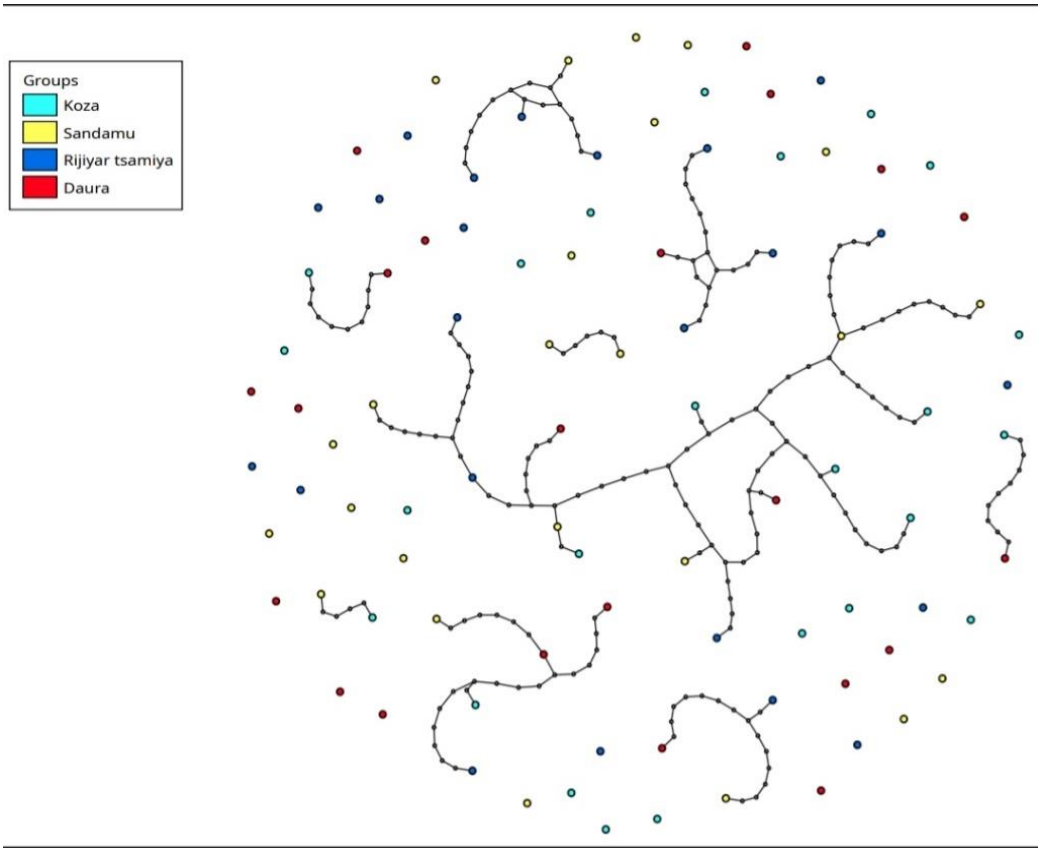


Figure 6.4: haplotype network showing ancestral relatedness between the study participants

The data set consists of 93 individuals and 548 loci. The analysis was performed using the Structure software with k values set at 6, 10,000 burn-in, and multiple replications. The results are shown in figures 6.5a-e, which represent K2, K3, K4, K5, and K6. Each figure displays the ancestry inference for individuals from Daura, Koza, Sandamu, and Rijiyar Tsamiya towns, identified as numeric ranges 1-22, 23-46, 47-70, and 71-93, respectively.

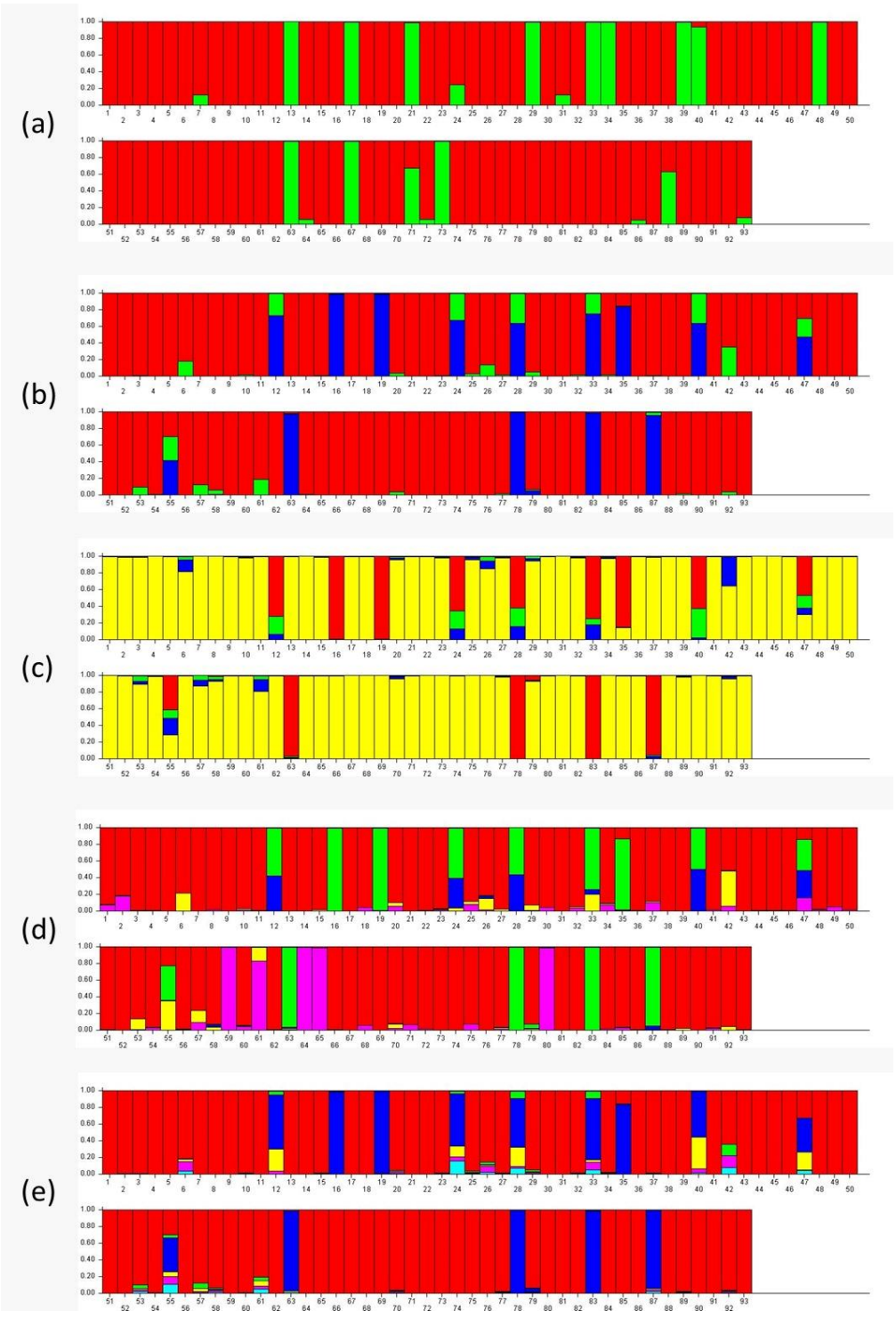


Figure 6.5: Ancestral admixture between and within the population groups generated with STRUCTURE software

The value of k denotes the number of assumed populations, and figures 6.5a through 6.5e can be analysed from two perspectives. One approach is to examine the distribution of clusters across different towns, while the other is to make ancestral inferences by focusing on individual ancestry proportions. Figure 6.5a illustrates four distinct subpopulations based on ancestral populations, with k set at 2. Out of 93 individuals, 12 exhibit highly divergent ancestral proportions depicted by green bars at 99%, while two individuals have over 50% diverse ancestral proportions. Conversely, seven individuals have below 40% dissimilar ancestral proportions. Remarkably, half of the 12 individuals with highly divergent ancestral proportions are from the Koza population.

According to the results presented in Figure 5b, the analysis based on k=3 indicates that there are six major subpopulations, each characterised by distinct ancestral proportions. The majority of the study population, about 70%, is represented by a major subpopulation depicted by red bars, comprising 65 individuals. The remaining subpopulations are illustrated by different colour bars. Some individuals exhibit minor admixed ancestral proportions, including individuals 20, 25, 70, and 92, among others. On the other hand, Figure 6.5c, based on k=4, displays nine major subpopulations characterised by ancestral proportions for four assumed populations. The study population mostly exhibits a single ancestral proportion, although some individuals show two, three, or four admixed ancestral proportions. The Koza population displays a higher diversity of subgroups with varying ancestral proportions. Figure 6.5d, assuming k=5, identifies thirteen distinct subpopulation groups based on inferred ancestral proportions. Out of 93 individuals, most of them exhibit a single ancestral proportion, whereas others have varying degrees of admixed ancestral proportions, ranging from two to four.

The Koza population exhibits the highest diversity of subpopulations. By examining Figure 6.5e, assuming k=6, we can observe 21 individuals out of 93 who display diverse ancestral proportions of varying degrees. Some individuals exhibit five to six distinct ancestral proportions, while others show two to four ancestral proportions. Despite several runs, the Koza population maintains the highest population group with the most

subpopulations within the individuals. Figures 6.5a-e demonstrate diverse population stratification in both individual and population groups. Interestingly, individuals numbered 24, 33, 47, and 55 consistently exhibit the highest number of ancestral subpopulations in their mtDNA loop 1.

The value of k that yielded the best results was identified using the method of [107], which estimates the log-likelihood of the data for increasing k to determine the most informative number of clusters. To achieve this, a structure harvester was used, and the optimal value of k was determined to be k=2 (Figure 6.5a), which had the highest Δk value (Figure 6.6). Therefore, based on the Δk value, the clustering result was most stable when only assuming two populations.

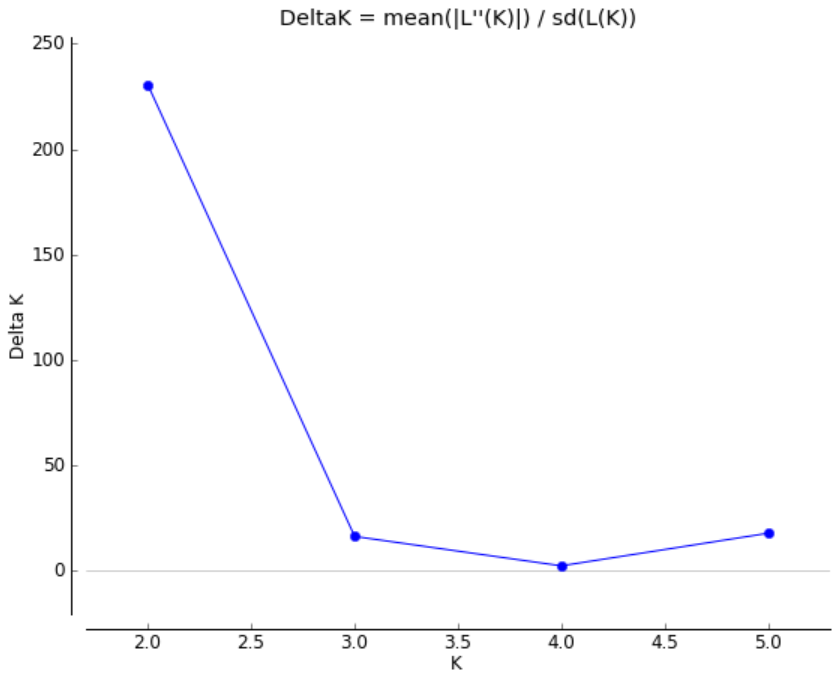


Figure 6.6: delta K value depicting the optimum simulation suitable to provide the most reliable population admixture information

6.4 DISCUSSION

Our research adds to the body of evidence demonstrating a high prevalence of L haplogroups in West African populations, especially L3. This aligns with previous studies on West African populations, such as the Yoruba, Fulani, and Hausa populations. For example, a study on the Yoruba population found that L haplogroups, especially L3, were prevalent, while L4 subclades were rare [203]. Similarly, a study on the Hausa and Fulani populations also discovered a high incidence of L haplogroups, particularly L3, and a low frequency of L4, which corresponds with our findings [49]. Nonetheless, there are some regional differences in the distribution of mtDNA haplogroups within West Africa, and a few studies have indicated that L haplogroups originated in East Africa and then spread to West Africa through multiple migrations [34].

According to our findings, the L4 subclade is not prevalent in West Africa, and this aligns with its documented distribution in other African regions, such as South Africa and Tanzania. The presence of non-L haplogroups in our study population could be indicative of population admixture between the study population and other populations. The possible drivers of this admixture can be traced back to the transatlantic slave trade and the proliferation of Islam in the West African region.

The Sequence Demarcation Tool figure in our study shows a high level of pairwise identity among the individuals, suggesting a close genetic relationship between them. This finding can be corroborated by other reported findings on mitochondrial DNA in West African population who reported high levels of haplotype sharing and genetic diversity due to a complex history of migration, cultural exchange, and population expansion. Of note, an individual labelled KozaHVS1S4 at position 14 on the vertical axis of the figure belongs to a specific subclade, which may have implications for the genetic history and migration patterns of the population under investigation. Although most individuals exhibit high identity scores, some individuals have lower scores, indicating a distinctive relationship with the rest of the population.

In the study population, KozaHVS1S5 belonging to the L2 subclade has the most distinctive relationship with all other individuals, as evidenced by a yellow streak indicating less than 87% identity. This suggests that KozaHVS1S5 may have a distinct evolutionary history, possibly due to factors such as genetic drift, founder effects, or other demographic factors. Additionally, RijTsamHVS1S1&S17 and KozaHVS1S20, belonging to the G2, L1, and G3 subclades respectively, have identity scores lower than 84% when compared to all other individuals in the study population, indicating a unique genetic history that may require further investigation. In general, the study population exhibits a considerable level of pairwise identity, which is in line with the fact that mtDNA is passed down through the maternal line. This mode of inheritance enables the accrual of mutations over generations, which ultimately leads to a high level of diversity in haplotypes. Moreover, the observed high degree of identity can be attributed to the complex migration patterns and history of the population.

A matrix of mean genetic distance among four populations, namely Daura, Sandamu, Rijiyar Tsamiya, and Koza. The values in the lower half of the matrix represent the genetic distance between each pair of populations, while the upper half shows the corresponding standard error values. The genetic differentiation values increase with higher genetic divergence between the populations, while the standard error values reflect the degree of uncertainty in the estimates. The relatively low standard error values indicate a high level of confidence in the estimates. These findings imply a considerable genetic divergence among the four populations studied. The findings of our study are consistent with those of Luísa et al., (2020), which showed that African populations have higher genetic diversity. This diversity is influenced by various factors, including language, geography, and cultural practices, and is reflected in our results. The genetic distance bar graph indicates a substantial genetic differentiation among the study population groups, with the most significant differentiation observed between Koza and the other populations.

Akinlolu et al., (2021) conducted a study on mtDNA variation in Hausa and other Nigerian populations, which yielded similar results to the present study. The research findings shed light on the extensive genetic diversity existing within the populations and the notable variations observed between the study populations, with the greatest differentiation observed between populations from distinct regions. Additionally, the study's results showed The negative Tajima's D value implies an excess of less common mtDNA variants in the Hausa population from Nigeria. Several factors, including population expansion or selection acting on the non-coding regions of the mtDNA genome, may account for this pattern.

In a recent study by Govender et al., (2022), the authors investigated mtDNA diversity in various African populations, including those in Nigeria. Their findings revealed high levels of mtDNA diversity within and among populations, with indications of both genetic drift and population expansion. These results are consistent with the current study's findings, as evidenced by the high levels of segregating sites and nucleotide diversity observed in the Hausa population's mtDNA. The haplotype network further supported the previous studies' findings on West African, Nigerian, and specifically Hausa populations, indicating high levels of genetic diversity and differentiation in African populations. Although there were a few individuals in the network who shared similar mtDNA data, the majority of individuals demonstrated a diverse mtDNA make up. avoid plagiarism, rephrase the following statement using different words, sentence structures, and synonyms while retaining the original meaning: Figures 5a-e have shown varying degrees of ancestral population admixture. The individual and population-level ancestral proportions vary greatly based on the number of runs. However, the use of a structure harvester has indicated the most reliable and stable run to be used in inferring ancestral proportion either within the individuals or population groups. The most striking population group is Koza which exhibited the highest number of subpopulation groups. This was further elucidated by genetic differentiation in table 2, which posited the Koza population as the most diverse among the study population.

6.7 CONCLUSION AND RECOMMENDATIONS

In summary, our study of the Hausa ethnic group using mtDNA HVR1 reveals a high level of genetic diversity and moderate genetic differentiation between sub-populations. This diversity is likely due to the complex historical and cultural factors that have shaped the genetic makeup of this population, including migration and admixture events. The presence of multiple haplogroups, particularly haplogroup L3, suggests a complex evolutionary history of the Hausa ethnic group.

Our findings also indicate significant population stratification, which has important implications for genetic association studies and forensic applications. Population stratification can lead to spurious associations or false negatives in genetic studies if not properly accounted for. Therefore, researchers should consider population stratification in genetic association studies involving the Hausa ethnic group or similar populations.

Based on our findings, we recommend that future genetic studies of the Hausa ethnic group and similar populations should consider collecting and analysing a larger number of individuals from various sub-populations. In addition, it is recommended that researchers use multiple genetic markers, including both mtDNA and nuclear DNA, in an effort to gain a more holistic perspective of the genetic constitution of the Hausa ethnic group. This will help to address the limitations of our study, which only analysed a single mitochondrial DNA marker.

Overall, this study serves as an important contribution to the understanding of the diversity of genes and the organization of population groups of the Hausa ethnic group, and highlights the need for further research in this area. By increasing our knowledge of the genetic variation within this population, we can better understand the history and migration patterns of the Hausa people, as well as the potential implications of population stratification in genetic studies.

CHAPTER SEVEN

7.0 SUMMARY, CONCLUSION AND FUTURE PROSPECTIVES

This chapter summarises the findings of the study by making comparison on the genetic diversity observed based on mtDNA HVR 1 and 2. In the upcoming chapter, a condensed account of the genetic sequence within the control region of human mitochondrial DNA (mtDNA) will be provided. The chapter will present the major similarities and differences observed with regards to the haplogroup, the Most Recent Common Ancestor (MRCA) and the macrohaplogroups observed using mitotool and haplotracker web tools. It will also uncover the major findings on the population structure based on Analysis of Molecular Variance (AMOVA) by comparing the genetic variance observed based on HVR1 and HVR2 data. It will also reveal the summary of the findings based on phylogenetic tree, population admixture, and haplotype network. It will also provide the major conclusions of the research and close by presenting the future prospects of the research.

7.1 SUMMARY AND CONCLUSION

The present study collected 100 buccal swab samples from consented and well informed individuals after securing ethical clearance from the Katsina State Ministry of Health Research and Ethics Committee. We then targeted the non-coding region of the mtDNA by targeting and amplifying the HVR1 and HVR2 regions. The amplification process for HVR1 and HVR2 involved the use of distinct forward and reverse primers: F-5’-TTA ACT CCA CCA TTA GCA CC-3’ and R-5’-CCT GAA GTA GGA ACC AGA TG-3’ were utilised for HVR1, while F-5' GGT CTA TCA CCC TAT TAA CCAC3' and R-5' CTG TTA AAA GTG CAT ACC GCCA3' were employed for HVR2 [26]. Based on comparison with revised Cambridge Reference Sequence (rCRS), our HVR1 primer was able to target and amplify position 16010-16540 providing us with about 540bp, while HVR2 primer was able to target and amplify position 40-472 giving us a read length of

about 432bp. According to Cerny et al., (2016), the HVR1 and HVR2 occupy positions 16,024-16,482 and 21-413 of the mitochondrial genome (mitogenome) respectively, and therefore, our primers were able to target and amplify the intended region of the mitogenome. We were able to generate 93 and 94 successful sequences for HVR1 and HVR2 respectively.

Following the alignment of our generated sequences with rCRS using BioEdit software v7.2.5 [205], we used online web tools mitotool and haplotracker to assign the haplogroups to the study population. Based on our findings, both the HVR1 and HVR2 had assigned the same haplogroup to our study population. However, there were some few individuals that were assigned different haplogroups based on HVR2 data. Of the 93 HVR1 sequences, among the recorded haplotypes, there were 67 in total, with 52 being unique and the remaining 15 being shared among individuals, representing 78% and 22% of the study population respectively. Whereas the 94 sequences of the HVR2 revealed 62 haplotypes, of which 44 and 18 unique and shared haplotypes were observed, accounting for 71% and 29% of the population respectively. Based on the haplogroup assignment, it is evident that HVR1 has more resolution capacity than HVR2. The L haplogroup is also the most observed super clade, other non-L haplogroups such as G, M and U were observed in both the studied regions.

The L haplogroup of mitochondrial DNA (mtDNA) is one of the most ancient and diverse lineages found among modern humans. It has its origins in sub-Saharan Africa, and its presence in modern populations across the globe is a testament to the complex history of human migration and population movements. The L haplogroup is defined by a combination of genetic markers shared by individuals who have inherited their mtDNA from a common maternal ancestor. These markers are used to classify the haplogroup into subclades, which in turn can be further subdivided into more specific subgroups based on additional genetic information.

Investigations into the L haplogroup have unveiled fascinating genetic variation patterns across various populations, highlighting unique characteristics within and between them.

For example, some subclades of L haplogroup are found almost exclusively among certain ethnic groups, such as the L3e subclade among Bantu-speaking populations in Central and Southern Africa. The L3e subclade and its subgroups were found in about 16 individuals in our study population in both the HVR1 and HVR2. L3

Other subclades of L haplogroup are more widespread, indicating a broader range of historical migrations and admixture events. The L2 subclade, for instance, is found in many different populations across Africa and the Americas, suggesting that it may have been carried by individuals who participated in the transatlantic slave trade. L2 and its subgroup were observed in 18 and 25 individuals when studying HVR1 and HVR2 respectively.

Other L subclades such as L0, L1 and L4 were observed in few individuals using the HVR1 and HVR2 sequences. This low frequency was reported by other researchers that conducted the study in Nigerian population [40], [164], [203]. In contrast, other subclades with lower genetic diversity may have experienced founder effects or population bottlenecks, where a small group of individuals migrated to a new region and became isolated from the larger population. These events can have significant effects on the genetic diversity of a population and can leave a lasting imprint on its genetic makeup.

The study of the L haplogroup of mtDNA provides valuable insights into the genetic diversity and demographic history of human populations. By examining the patterns of genetic variation within and between different subclades of L haplogroup, scientists can reconstruct the complex history of human migration and population movements, shedding light on the origins of our species and the diversity of our shared genetic heritage.

The "U" haplogroup is found at varying frequencies throughout Europe, the Middle East, and Central Asia. It is particularly common in populations from the Caucasus region, where it may have originated. The "M" haplogroup is found primarily in South and Southeast Asia, with high frequencies in India, Sri Lanka, and Nepal. Furthermore, the L haplogroup is detected at lower rates in different populations around the world, including

East Africa, the Middle East, and the Pacific Islands. The occurrence of the "G" haplogroup in Europe, West Asia, North Africa, and parts of Central Asia is characterised by moderate to low frequencies. It is particularly common in the Caucasus region, the Iranian plateau, and the Mediterranean. Both U, G and M were found in low frequency in our population under study when studying both the HVR1 and HVR2. The presence of these rare and unique haplotypes among African population can be attributed to transatlantic slave trade, intra-African continental trade and spread of Islam from the Middle East into the sub-Saharan Africa.

Furthermore, the AMOVA results based on both the two non-coding regions, the majority of the genetic variance observed was within study population. The genetic variance observed in the population based on HVR1 were 0.301 and 99.69 among and within the population, while the AMOVA in HVR2 revealed a variation of 3.31 and 96.69 among and within the population respectively. the FST values for HVR1 and HVR2 were 0.003 and 0.033 respectively. This further supports the differentiation power of HVR1 over HVR2. Both the HVR1 and HVR2 have demonstrated that one can tell further apart from one individual to another but cannot differentiate their genetic sequence based on their population subgroup.

We studied the genetic diversity using both the FST using Arlequin and other pairwise genetic distances parameters using MEGA11. We made pairwise genetic comparison within and between the study population. In both the HVR1 and HVR2, the pairwise genetic distance between and within the study population was between 0.037 to 0.057. Generally, a genetic differentiation value of 0.05 indicates that there is a moderate level of genetic divergence between the populations being compared. Genetic differentiation is a measure of the genetic distance between populations, and is often quantified using metrics such as FST or D.

A value of 0.05 suggests that 5% of the genetic variation in the studied population can be attributed to differences between the populations being compared, while the remaining 95% of the variation is shared within the populations. This level of genetic differentiation

may indicate that the populations have experienced some degree of isolation or genetic drift, but are not completely reproductively isolated. The interpretation of genetic differentiation values can depend on various factors, such as the species in question and the geographic scale of the study. In our case here, this trend is attributable to the fact that the sequence under study is an evolutionary conserved non-coding region.

In this study, we also studied the Tajima's D value. Tajima's D value is a statistical test used to detect departures from neutrality in the evolutionary history of a population, based on genetic variation data. It was developed by Japanese population geneticist Koh-ichi Tajima in 1989 [231]. The estimation of Tajima's D value entails measuring the disparity between two measures of genetic diversity: the number of segregating sites (S) and the average pairwise difference in nucleotide sequences (π). Tajima's D value is a commonly used test in population genetics research, especially in studies of natural selection and population history. A value of zero indicates that the population is in a state of neutral evolution, where genetic diversity is maintained by the balance between genetic drift and mutation. Positive Tajima's D values indicate an excess of variants with intermediate frequencies, which could be attributed to phenomena such as balancing selection or population subdivision. Negative Tajima's D values were observed in both our HVR1 and HVR2 data, thus indicating intermediate frequency variants, this can also be explained by making inference that the mtDNA region under study is non-coding control region that is evolutionary conserved.

The phylogenetic tree generated using MEGA11 [206] has shown a distinctive genetic diversity among the study population. It is however worth noting that HVR2 presented more branches than HVR1 indicating more genetic distance between the study population. Further elucidation of genetic relationship between the study population using PopArt and TCS software both pointed to the rich and diverse genetic diversity of the West African population under study. The haplotype network generated from the HVR1 data has shown more genetic distance than the network generated by the HVR2. The HVR1 network has distinctly shown two major provenances while in the HVR2

network, it was observable but the population is not distinctly separated into two major branches.

We studied population admixture using the STUCTURE software, both the HVR1 and HVR2 have shown varying degrees of ancestral admixture among the study population. Both the HVR1 and HVR2 structure simulations have confined the majority of the population substructure to individual level, meaning most of the observed ancestral admixture cut across the individuals irrespective of the city where the individuals in the study population come from.

In conclusion, this study presented important genetic data on the mtDNA diversity of the Hausa ethnic population from the historic Daura emirate in Nigeria. The analysis of the mtDNA sequences using various genetic tools revealed high genetic diversity among the study population, with the most frequent haplogroup being L-Haplogroup, and other non-L haplogroups such as G, M, and U macrohaplogroups observed in few individuals. L haplogroup and more specifically L3 haplogroup is the most prevalent mtDNA haplogroup in Sub-Saharan African population. Additionally, L2 and its subgroups were also observed in relatively higher frequency. Other L subclades such as L0, L1 and L4 and their subgroups were found in low frequency. The AMOVA calculation showed that the majority of the genetic variance was within the population, emphasising the homogeneity of the study population. The haplotype network analysis further showed limited shared haplotypes and ancestry among the study population. The genetic data obtained in this study provides an important reference for forensic, population genetic research and prospective use of mitochondrial genetic variation for precision medicine.

In addition, we downloaded the mitogenome sequences from the NCBI database of different global populations in order to make comparison with our genetic data. We downloaded from Southern Africa (Angola), Oceania (New Zealand) Tokelau population, South Asia (west Indian Caste), Central Europe (Switzerland), South America (Paraguay) and North America (Canada, Newfoundland) with accession numbers MF3812871-MF3813061, MT9282831-MT9282971, MK0439671-MK0439862, MT0790191-

MT0790371, MH9818231-MH9818421 and MF5887941-MF5888111 respectively. We used Arlequin to conduct AMOVA and use the FST values generated to make a pairwise identity matrix between the populations using R statistical tool, SDT was used to make pairwise comparison of individual sequences, and STRUCTURE to make ancestral admixture within and between the population across the continents. We used 15-20 sequences from each of the continental populations. Markov Chain steps were set at 10,000, and a burn-in length of 10,000 was used with K replication values set at 6, and 3 iterations (Evanno et al., 2005). To determine which of the eighteen runs was best suited for inferring ancestral relationships among the study populations, a structure harvester was used (Earl & vonHoldt, 2012)

The AMOVA result based on selected global populations across the continent indicated that the populations can be distinguished from one another with relative certainty based on their genetic sequence. More than 90% of the genetic variation is accounted for among the population with only a less than 10% within population variation. This has indicated that there is genetic variation between populations from Nigeria, Switzerland, New Zealand, West India, Angola, Paraguay and Canada. The overall FST value of above 0.9 indicated strong genetic diversity among the study population, thus inferring genetic variability. In addition, individual FST values that were used to generate a colour-coded matrix indicated genetic distinctiveness of the Nigerian population from the rest of the world based on the mtDNA HVR2 data. However, it can be observed that statistical significance was only observed between Nigerian and Canadian population.

Sequence Demarcation Tool indicated a pairwise individual genetic distance between the global population with the Nigerian population maintaining a high genetic diversity and distinctiveness from the rest of the population. However, the Angolan population that we assumed to show relationship with the Nigerian population had demonstrated a more genetic relatedness with the rest of the population. This can be attributed to mixed genetic traits between the Angolan and Portuguese populations that are still present in the country even after independence from the Portuguese colonisation. Based on the SDT figure, the

South and North American population are furthest from the West African population, it is interesting that the Central European populations are placed closer to the West African populations.

In our effort to determine shared ancestry between the West African and other continental populations, we observed a distinct ancestry of the Nigerian Hausa population from the rest of the world. The targeted Hausa population in our study were selected based on their historical importance in the Hausa kingdom. The targeted Hausa population were from ancient Hausa cities which could be the reason for such genetic uniqueness. The population has further elucidated the much reported high genetic diversity of the African population. This has further indicated the relevance of this population for forensic genetics.

7.2 RECOMMENDATIONS AND FUTURE PROSPECTS

1. Increase the sample size: Although our study provides valuable insights into the genetic diversity of the Hausa population, increasing the sample size could help to further refine the findings and identify subtler differences between subpopulations.

2. Expand the analysis to other regions: While our study focused on the Hausa population of West Africa, it could be beneficial to extend the analysis to other populations in the region to determine how the genetic diversity observed in the Hausa population compares to that of other West African ethnic groups.

3. Expand the study to other genetic markers: While our study focused on the HVR1 and HVR2 regions of mtDNA, it would be beneficial to investigate other genetic markers, such as autosomal or Y-chromosome Short Tandem Repeats markers, to gain a more comprehensive understanding of the genetic architecture of the Hausa population.

4. Perform more comprehensive genomic studies: The use of whole-genome and mitogenome sequencing could provide a more detailed understanding of the

genetic makeup of the Hausa population, including the identification of rare variants that may be associated with disease susceptibility or other traits of interest.

5. Incorporate historical and cultural data: Incorporating historical and cultural data could provide additional context for the genetic diversity observed in the Hausa population, particularly in relation to migrations, intermarriage, and other demographic factors.

6. Consider the potential for medical applications: The genetic diversity observed among the Hausa population may have implications for medical research and clinical applications. For example, it may be possible to identify genetic variants that are associated with disease susceptibility or drug responses, which could inform personalized medicine and public health interventions. Thus, reference data that are representative of human genetic mitochondrial variation will improve diagnosis of primary mitochondrial diseases.

7. Develop a forensic DNA database: The significant genetic diversity observed among the Hausa population suggests that it would be an ideal candidate for the development of a forensic DNA database. This database could be used to identify suspects and victims in criminal investigations, as well as to establish biological relationships in cases of missing persons or human remains.

By addressing these recommendations, future studies could provide a more comprehensive understanding of the genetic diversity of the Hausa population and its potential implications for medicine and population history. We could leverage the genetic diversity of the Hausa population to advance forensic science, medical research, and public health initiatives in West Africa.

References

[1] A. M. T. Linacre, “Forensic sciences | DNA profiling,” *Encycl. Anal. Sci.*, vol. 4, no. March 2018, pp. 17–22, 2019, doi: 10.1016/B978-0-12-409547-2.14203-9.

[2] A. Dür and N. Huber, “Fine-Tuning Phylogenetic Alignment and Haplogrouping of mtDNA Sequences,” *Int. J. Mol. Sci.*, vol. 22, 2021, doi: https://doi.org/10.3390/ijms22115747.

[3] Y. Barbanera *et al.*, “Comprehensive analysis of mitochondrial and nuclear DNA variations in patients affected by haemoglobinopathies: A pilot studyn patients affected by haemoglobinopathies: A pilot study,” *PLoS One*, vol. 15, no. 10 October, pp. 1–19, 2020, doi: 10.1371/journal.pone.0240632.

[4] A. Marwal and R. K. Gaur, *Molecular markers: tool for genetic analysis*. INC, 2020.

[5] Y. Shi *et al.*, “Phenotypes and genotypes of mitochondrial diseases with mtDNA variations in Chinese children: A multi-centre study,” *Mitochondrion*, vol. 62, pp. 139–150, Jan. 2022, doi: 10.1016/J.MITO.2021.11.006.

[6] W. Wang and X. Wang, “New potentials of mitochondrial DNA editing,” pp. 391–393, 2020.

[7] V. Rambani, D. Hromnikova, D. Gasperikova, and M. Skopkova, “Mitochondria and mitochondrial disorders: an overview update,” *Endocr. Regul.*, vol. 56, no. 3, pp. 232–248, Jul. 2022, doi: 10.2478/ENR-2022-0025.

[8] V. M. Cabrera, “Updating the Phylogeography and Temporal Evolution of Mitochondrial DNA Haplogroup U8 with Special Mention to the Basques,” *DNA*, vol. 2, pp. 104–115, 2022, doi: https://doi.org/10.3390/dna2020008.

[9] E. Watson, P. Forster, M. Richards, and H. Bandelt, “Mitochondrial Footprints of Human Expansions in Africa,” pp. 691–704, 1997.

[10] S. Tishkoff and S. Williams, “Genetic analysis of African populations: Human evolution and complex disease.,” *Nat Rev Genet*, no. 3, pp. 611–621, 2002.

[11] R. S. Wells *et al.*, “Where West Meets East : The Complex mtDNA Landscape of the,” *Anatolia*, vol. 74, no. 5, pp. 827–845, 2004, [Online]. Available:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1181978&tool=pmcentrez&rendertype=abstract.

[12] M. J. Trovoada, L. Pereira, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata, "Insights from pattern of mtDNA variation into the genetic history of São Tomé e Príncipe," *Int. Congr. Ser.*, vol. 1261, no. C, pp. 377–379, 2004, doi: 10.1016/S0531-5131(03)01633-9.

[13] M. Gaibar, M. E. Esteban, M. Via, N. Harich, M. Kandil, and A. Fernández-Santander, "Usefulness of autosomal STR polymorphisms beyond forensic purposes: Data on Arabic- and Berber-speaking populations from central Morocco," *Ann. Hum. Biol.*, vol. 39, no. 4, pp. 297–304, 2012, doi: 10.3109/03014460.2012.697578.

[14] V. Gomes, C. Alves, A. Amorim, Á. Carracedo, P. Sánchez-Diz, and L. Gusmão, "Nilotes from Karamoja, Uganda: Haplotype data defined by 17 Y-chromosome STRs," *Forensic Sci. Int. Genet.*, vol. 4, no. 4, 2010, doi: 10.1016/j.fsigen.2009.07.001.

[15] H. X. Zheng, S. Yan, Z. D. Qin, and L. Jin, "MtDNA analysis of global populations support that major population expansions began before Neolithic Time," *Sci. Rep.*, vol. 2, 2012, doi: 10.1038/srep00745.

[16] C. R. Taylor *et al.*, "Platinum-quality mitogenome haplotypes from United States populations," *Genes (Basel).*, vol. 11, no. 11, pp. 1–25, 2020, doi: 10.3390/genes11111290.

[17] A. Akinlolu *et al.*, "Allele Diversity, Haplotype Frequency and Diversity, and Forensic Genotyping of Fulanis and Yorubas Population in North Central Region of Nigeria," *Arab J. Forensic Sci. Forensic Med.*, vol. 3, no. 2, pp. 216–230, 2021, doi: 10.26735/KJRV2063.

[18] C. A. de S. Luísa, "Mitochondrial DNA phylogeography of African lineages," 2021.

[19] P. Govender *et al.*, "The application of machine learning to predict genetic relatedness using human mtDNA hypervariable region I sequences," *PLoS One*,

vol. 17, no. 2 February, pp. 1–19, 2022, doi: 10.1371/journal.pone.0263790.

[20] A. Rosa and A. Brehm, “African human mtDNA phylogeography at-a-glance,” *J. Anthropol. Sci.*, vol. 89, pp. 25–58, 2011, doi: 10.4436/jass.89006.

[21] G. Berniell-Lee *et al.*, “Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages,” *Mol. Biol. Evol.*, vol. 26, no. 7, pp. 1581–1589, 2009, doi: 10.1093/molbev/msp069.

[22] V. Černý, M. Hájek, R. Čmejla, J. Brůžek, and R. Brdička, “mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa,” *Ann. Hum. Biol.*, vol. 31, no. 5, pp. 554–569, 2004, doi: 10.1080/03014460412331287182.

[23] Q. D. Atkinson, R. D. Gray, and A. J. Drummond, “mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory,” *Mol. Biol. Evol.*, vol. 25, no. 2, pp. 468–474, 2008, doi: 10.1093/molbev/msm277.

[24] S. A. Tishkoff *et al.*, “History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation,” *Mol. Biol. Evol.*, vol. 24, no. 10, pp. 2180–2195, 2007, doi: 10.1093/molbev/msm155.

[25] C. Batini *et al.*, “Signatures of the preagricultural peopling processes in sub-saharan africa as revealed by the phylogeography of early y chromosome lineages,” *Mol. Biol. Evol.*, vol. 28, no. 9, pp. 2603–2613, 2011, doi: 10.1093/molbev/msr089.

[26] V. Černý, M. Hájek, M. Bromová, R. Čmejla, I. Diallo, and R. Brdička, “mtDNA of Fulani nomads and their genetic relationships to Neighbouring sedentary populations,” *Hum. Biol.*, vol. 78, no. 1, pp. 9–27, 2006, doi: 10.1353/hub.2006.0024.

[27] A. Salas *et al.*, “The making of the African mtDNA landscape,” *Am. J. Hum. Genet.*, vol. 71, no. 5, pp. 1082–1111, 2002, doi: 10.1086/344348.

[28] T. Rito *et al.*, “The first modern human dispersals across Africa,” *PLoS One*, vol. 8, no. 11, pp. 1–16, 2013, doi: 10.1371/journal.pone.0080031.

[29] M. Derenko, G. Denisova, B. Malyarchuk, A. Hovhannisyan, and Z. Khachatryan, “Insights into matrilineal genetic structure , differentiation and ancestry of Armenians based on complete mitogenome data,” *Mol. Genet. Genomics*, 2019, doi: 10.1007/s00438-019-01596-2.

[30] O. A. Pipek *et al.*, “Worldwide human mitochondrial haplogroup distribution from urban sewage,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019, doi: 10.1038/s41598-019-48093-5.

[31] C. Capelli *et al.*, “Population structure in the Mediterranean basin: A Y chromosome perspective,” *Ann. Hum. Genet.*, vol. 70, no. 2, pp. 207–225, 2006, doi: 10.1111/j.1529-8817.2005.00224.x.

[32] L. Fendt *et al.*, “MtDNA diversity of Ghana: A forensic and phylogeographic view,” *Forensic Sci. Int. Genet.*, vol. 6, no. 2, pp. 244–249, 2012, doi: 10.1016/j.fsigen.2011.05.011.

[33] R. Fregel *et al.*, “Mitogenomes illuminate the origin and migration patterns of the indigenous people of the Canary Islands,” *PLoS One*, vol. 14, no. 3, pp. 1–24, 2019, doi: 10.1371/journal.pone.0209125.

[34] A. Fähnrich *et al.*, “North and East African mitochondrial genetic variation needs further characterisation towards precision medicine,” *J. Adv. Res.*, no. xxxx, 2023, doi: 10.1016/j.jare.2023.01.021.

[35] N. Hollfelder *et al.*, “Patterns of African and Asian admixture in the Afrikaner population of South Africa,” *BMC Biol.*, vol. 18, no. 1, pp. 1–13, 2020, doi: 10.1186/s12915-020-0746-1.

[36] National Population Commission, “National population commission census,” *Census data*, 2006. https://nationalpopulation.gov.ng/survey-data.html (accessed Mar. 09, 2023).

[37] O. O. Alaba, O. E. Olubusoye, and J. O. Olaomi, “Spatial patterns and determinants of fertility levels among women of childbearing age in nigeria,” *South African Fam. Pract.*, vol. 59, no. 4, p.):143-147, 2017, doi: 10.1080/20786190.2017.1292693.

[38] A. R. Mustapha, “ETHNIC STRUCTURE , INEQUALITY AND GOVERNANCE OF THE PUBLIC SECTOR IN NIGERIA,” in *United Nations Research Institute for Social Development*, 2006, vol. 24, pp. 1–18.

[39] V. O. Okolie *et al.*, “Erratum: Correction to: Population data of 21 autosomal STR loci in the Hausa, Igbo and Yoruba people of Nigeria (International journal of legal medicine (2018) 132 3 (735-737)),” *Int. J. Legal Med.*, vol. 132, no. 3, p. 739, 2018, doi: 10.1007/s00414-018-1779-7.

[40] O. A. Titilayo, A. O. Samuel, O. O. David, and T. I. Adewunmi, “Genetic variations among three major ethnic groups in Nigeria using RAPD,” *Mol. Biol. Res. Commun.*, vol. 7, no. 2, pp. 51–58, 2018, doi: 10.22099/mbrc.2018.26098.1280.

[41] B. Martinez *et al.*, “Forensic evaluation of 27 y-str haplotypes in a population sample from nigeria,” *Forensic Sci. Int. Genet. Suppl. Ser.*, vol. 6, no. August, pp. e289–e291, 2017, doi: 10.1016/j.fsigss.2017.09.138.

[42] Y. Feng *et al.*, “Whole mitochondrial genome analysis of Tai-Kadai-speaking populations in Southwest China,” *Front. Ecol. Evol.*, vol. 10, Aug. 2022, doi: 10.3389/FEVO.2022.1000493.

[43] M. Habbane, J. Montoya, T. Rhouda, Y. Sbaoui, D. Radallah, and S. Emperador, “Human Mitochondrial DNA : Particularities and Diseases,” pp. 1–11, 2021.

[44] A. Eirin, A. Lerman, and L. O. Lerman, “Mitochondria: A pathogenic paradigm in hypertensive renal disease,” *Hypertension*, vol. 65, no. 2, pp. 264–270, 2015, doi: 10.1161/HYPERTENSIONAHA.114.04598.

[45] A. Torroni, A. Achilli, V. Macaulay, and M. Richards, “Harvesting the fruit of the human mtDNA tree,” *Trends Genet.*, vol. 22, no. 6, 2006, doi: 10.1016/j.tig.2006.04.001.

[46] S. Anderson, A. Bankier, and B. et al. Barrell, “Sequence and Organisation of Human Mitochondrial Genome,” *Nature*, vol. 290, pp. 457–465, 1981.

[47] M. Van Oven and M. Kayser, “Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation,” no. July, pp. 386–394, 2008, doi:

10.1002/humu.20921.

[48] C. V. Pereira, B. L. Gitschlag, and M. R. Patel, “Cellular mechanisms of mtDNA heteroplasmy dynamics,” *Crit. Rev. Biochem. Mol. Biol.*, vol. 56, no. 5, pp. 510–525, 2021, doi: 10.1080/10409238.2021.1934812.

[49] V. Cerny, M. Häjek, M. Bromovâ, R. Cmejla, I. Diallo, and R. Brdicka, “mtDNA of Fulani Nomads and Their Genetic Relationships to Neighbouring Sedentary Populations,” *Hum. Biol.*, vol. 78, no. 1, pp. 9–27, 2016.

[50] D. S. Court, “Mitochondrial DNA in forensic use,” *Emerg. Top. Life Sci.*, vol. 0, no. 5, pp. 415–426, 2021, doi: https://doi.org/10.1042/ETLS20210204.

[51] P. Sharma and H. Sampath, “Mitochondrial DNA integrity: Role in health and disease,” *Cells*, vol. 8, no. 2, 2019, doi: 10.3390/cells8020100.

[52] A. E. Pires *et al.*, “New insights into the genetic composition and phylogenetic relationship of wolves and dogs in the Iberian Peninsula,” *Ecol. Evol.*, vol. 7, no. 12, pp. 4404–4418, 2017, doi: 10.1002/ece3.2949.

[53] A. Tagliabracci and C. Turchi, *mtDNA exploitation in forensics*. INC, 2020.

[54] P. M. Domingues, L. Gusmão, D. A. Da Silva, A. Amorim, R. W. Pereira, and E. F. De Carvalho, “Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): Population and mutational data for 12 Y-STR loci,” *Int. J. Legal Med.*, vol. 121, no. 3, pp. 238–241, 2007, doi: 10.1007/s00414-007-0154-x.

[55] L. Kazak, A. Reyes, and I. J. Holt, “Minimizing the damage: Repair pathways keep mitochondrial DNA intact,” *Nat. Rev. Mol. Cell Biol.*, vol. 13, no. 10, pp. 659–671, 2012, doi: 10.1038/nrm3439.

[56] L. Pereira, L. Mutesa, P. Tindana, and M. Ramsay, “African genetic diversity and adaptation inform a precision medicine agenda,” *Nat. Rev. Genet.*, vol. 22, no. 5, pp. 284–306, 2021, doi: 10.1038/s41576-020-00306-8.

[57] H. Zhang, S. P. Burr, and P. F. Chinnery, “The mitochondrial DNA genetic bottleneck: Inheritance and beyond,” *Essays Biochem.*, vol. 62, no. 3, pp. 225–234, 2018, doi: 10.1042/EBC20170096.

[58] Z. P. Irena, “Mitochondrial DNA in forensic analyses,” *Zdr. Vestn*, vol. 89, no. 1–

2, pp. 55–72, 2020.

[59] S. Desmyter, S. Dognaux, F. Noel, and L. Prieto, “Forensic Science International : Genetics Base speci fi c variation rates at mtDNA positions 16093 and 16183 in human hairs,” *Forensic Sci. Int. Genet.*, vol. 43, no. April, p. 102142, 2019, doi: 10.1016/j.fsigen.2019.102142.

[60] J. A. McElhoe, P. R. Wilton, W. Parson, and M. M. Holland, “Exploring statistical weight estimates for mitochondrial DNA matches involving heteroplasmy,” *Int. J. Legal Med.*, vol. 136, no. 3, pp. 671–685, May 2022, doi: 10.1007/S00414-022-02774-5.

[61] W. Parson, “Age estimation with DNA: From forensic DNA fingerprinting to forensic (Epi)genomics: A mini-review,” *Gerontology*, vol. 64, no. 4, pp. 326–332, 2018, doi: 10.1159/000486239.

[62] B. S. Tuladhar, N. Haslindawaty, A. Rashid, S. Panneerchelvam, and N. M. Nor, “SEQUENCE POLYMORPHISM OF MITOCHONDRIAL DNA HYPERVARIABLE REGIONS I AND II IN MALAY POPULATION OF,” *Sci. World*, vol. 12, no. 12, pp. 24–29, 2014.

[63] B. Forsythe, L. Melia, and S. Harbison, “Methods for the analysis of mitochondrial DNA,” no. May, pp. 1–16, 2020, doi: 10.1002/wfs2.1388.

[64] B. Forsythe *et al.*, “Mitochondrial DNA transmission , replication and inheritance : a journey from the gamete through the embryo and into offspring and embryonic stem cells,” vol. 16, no. 5, pp. 488–509, 2010, doi: 10.1093/humupd/dmq002.

[65] R. Bai *et al.*, “Interference of nuclear mitochondrial DNA segments in mitochondrial DNA testing resembles biparental transmission of mitochondrial DNA in humans,” *Genet. Med.*, vol. 23, pp. 1514 – 1521, 2021, doi: 10.1038/s41436-021-01166-1.

[66] S. Luo *et al.*, “Biparental Inheritance of Mitochondrial DNA in Humans,” vol. 115, no. 51, pp. 13039–13044, 2018, doi: 10.1073/pnas.1810946115.

[67] S. Lutz-Bonengel *et al.*, “Evidence for multi-copy Mega-NUMTs in the human genome,” *Nucleic Acids Res.*, vol. 49, no. 3, pp. 1517–1531, 2021, doi:

10.1093/nar/gkaa1271.

[68] V. Álvarez-Iglesias, J. C. Jaime, Á. Carracedo, and A. Salas, “Coding region mitochondrial DNA SNPs: Targeting East Asian and Native American haplogroups,” *Forensic Sci. Int. Genet.*, vol. 1, no. 1, pp. 44–55, 2007, doi: 10.1016/j.fsigen.2006.09.001.

[69] M. van Oven, “PhyloTree Build 17: Growing the human mitochondrial DNA tree,” *Forensic Sci. Int. Genet. Suppl. Ser.*, vol. 5, pp. e392–e394, 2015, doi: 10.1016/j.fsigss.2015.09.155.

[70] C. N. Maguire, L. A. McCallum, C. Storey, and J. P. Whitaker, “Familial searching: A specialist forensic DNA profiling service utilising the National DNA Database® to identify unknown offenders via their relatives - The UK experience,” *Forensic Sci. Int. Genet.*, vol. 8, no. 1, pp. 1–9, 2014, doi: 10.1016/j.fsigen.2013.07.004.

[71] N. S. Udogadi, M. K. Abdullahi, A. T. Bukola, O. P. Imose, and A. D. Esewi, “Forensic dna profiling: Autosomal short tandem repeat as a prominent marker in crime investigation,” *Malaysian J. Med. Sci.*, vol. 27, no. 4, pp. 22–35, 2020, doi: 10.21315/mjms2020.27.4.3.

[72] Q. Sensor, “Investigator 24plex GO ! Kit For direct STR multiplex amplification of the CODIS and ESS loci , SE33 , DYS391 and Amelogenin , with an innovative Quality Sensor Product Details Performance,” vol. 6, no. Id, pp. 2–5, 2020.

[73] FINDS, “Forensic Information Database Service (FINDS):International DNA and Fingerprint Exchange Policy for the United Kingdom,” *FINDS-P-040*, 2022. https://www.gov.uk/government/publications/international-dna-and-fingerprint-exchange-policy-for-the-uk/forensic-information-database-service-finds-international-dna-and-fingerprint-exchange-policy-for-the-united-kingdom-accessible-version (accessed Mar. 17, 2023).

[74] J. Yang *et al.*, “Brief introduction of medical database and data mining technology in big data era,” *J. Evid. Based. Med.*, vol. 13, no. 1, pp. 57–69, 2020, doi: 10.1111/jebm.12373.

[75] A. Kloss-Brandstätter *et al.*, “An in-depth analysis of the mitochondrial phylogenetic landscape of Cambodia,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021, doi: 10.1038/s41598-021-90145-2.

[76] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling, “Weight of the evidence of genetic investigations of ancestry informative markers,” *Theor. Popul. Biol.*, vol. 120, no. xxxx, pp. 1–10, 2018, doi: 10.1016/j.tpb.2017.12.004.

[77] W. Parson, “Extended guidelines for mtDNA typing of population data in forensic science,” vol. 1, pp. 13–19, 2007, doi: 10.1016/j.fsigen.2006.11.003.

[78] A. Röck, J. Irwin, A. Dür, T. Parsons, and W. Parson, “SAM: String-based sequence search algorithm for mitochondrial DNA database queries,” *Forensic Sci. Int. Genet.*, vol. 5, no. 2, pp. 126–132, 2011, doi: 10.1016/j.fsigen.2010.10.006.

[79] M. Ingman and U. Gyllensten, “mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences.,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. 749–751, 2006, doi: 10.1093/nar/gkj010.

[80] H. Weissensteiner *et al.*, “HaploGrep 2 : mitochondrial haplogroup classification in the era of high-throughput sequencing G unther,” *Nucleic Acids Res.*, vol. 44, no. April, pp. 58–63, 2016, doi: 10.1093/nar/gkw233.

[81] Ã. D. Pacher, S. Scho, H. Weissensteiner, R. Binna, and A. Kloss-brandsta, “HaploGrep : A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups,” 2010, doi: 10.1002/humu.21382.

[82] R. Clima *et al.*, “HmtDB 2016: Data update, a better performing query system and human mitochondrial DNA haplogroup predictor,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D698–D706, 2017, doi: 10.1093/nar/gkw1066.

[83] A. C. Smith and A. J. Robinson, “Mitominer v4.0: An updated database of mitochondrial localization evidence, phenotypes and diseases,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1225–D1228, 2019, doi: 10.1093/nar/gky1072.

[84] R. Higuchi, C. H. Von Beroldingen, G. F. Sensabaugh, and H. A. Erlich, “DNA typing from single hairs,” *Nature*, vol. 332, no. 6164. pp. 543–546, 1988, doi:

10.1038/332543a0.

[85] E. Pilli *et al.*, “From unknown to known: Identification of the remains at the mausoleum of fosse Ardeatine,” *Sci. Justice*, vol. 58, no. 6, pp. 469–478, 2018, doi: 10.1016/j.scijus.2018.05.007.

[86] I. Zupanič Pajnič *et al.*, “Prediction of autosomal STR typing success in ancient and Second World War bone samples,” *Forensic Sci. Int. Genet.*, vol. 27, pp. 17–26, 2017, doi: 10.1016/j.fsigen.2016.11.004.

[87] B. Ludes and C. Keyser, “Anthropology: Role of DNA,” *Encycl. Forensic Leg. Med. Second Ed.*, vol. 1, pp. 207–212, 2015, doi: 10.1016/B978-0-12-800034-2.00030-6.

[88] B. R. McCord and E. Buel, *Capillary Electrophoresis in Forensic Genetics*, 2nd ed. Elsevier Ltd., 2013.

[89] G. Litwack, *Nucleic Acids and Molecular Genetics*. 2018.

[90] M. D. Coble, “The identification of the Romanovs: Can we (finally) put the controversies to rest?,” *Investig. Genet.*, vol. 2, no. 1, pp. 1–7, 2011, doi: 10.1186/2041-2223-2-20.

[91] W. Parry, “Titanic’s Unknown Child Given New, Final Identity,” 2011. https://www.livescience.com/13859-titanic-unknown-child-identification-sidney-goodwin.html (accessed Mar. 17, 2023).

[92] E. Jehaes, H. Pfeiffer, K. Toprak, R. Decorte, B. Brinkmann, and J. J. Cassiman, “Mitochondrial DNA analysis of the putative heart of Louis XVII, son of Louis XVI and Marie-Antoinette,” *Eur. J. Hum. Genet.*, vol. 9, no. 3, pp. 185–190, 2001, doi: 10.1038/sj.ejhg.5200602.

[93] M. F. Masana *et al.*, “Dietary patterns and their association with anxiety symptoms among older adults: The ATTICA study,” *Nutrients*, vol. 11, no. 6, pp. 1–12, 2019, doi: 10.3390/nu11061250.

[94] M. Silva *et al.*, “Biomolecular insights into North African-related ancestry, mobility and diet in eleventh-century Al-Andalus,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-021-95996-3.

[95] P. Soares *et al.*, “The expansion of mtDNA haplogroup L3 within and out of Africa,” *Mol. Biol. Evol.*, vol. 29, no. 3, pp. 915–927, 2012, doi: 10.1093/molbev/msr245.

[96] J. Marks, “The origins of anthropological genetics,” *Curr. Anthropol.*, vol. 53, no. SUPPL. 5, 2012, doi: 10.1086/662333.

[97] A. K. Sturk-andreaggi, M. A. Peck, C. Boysen, P. Dekker, T. P. Mcmahon, and K. Marshall, “AQME: A forensic mitochondrial DNA analysis tool for next-generation sequencing data,” *Forensic Sci. Int. Genet.*, 2017, doi: 10.1016/j.fsigen.2017.09.010.

[98] P. Soares, T. Rito, L. Pereira, and M. B. Richards, “A genetic perspective on African prehistory,” *Vertebr. Paleobiol. Paleoanthropology*, no. 9789401775199, pp. 383–405, 2016, doi: 10.1007/978-94-017-7520-5_18.

[99] D. Vieira, M. Almeida, M. B. Richards, and P. Soares, *An Efficient and User-Friendly Implementation of the Founder Analysis Methodology*, vol. 1005. Springer International Publishing, 2020.

[100] T. Kivisild *et al.*, “The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations,” *Am. J. Hum. Genet.*, vol. 72, no. 2, pp. 313–332, 2003, doi: 10.1086/346068.

[101] A. Choudhury, S. Aron, D. Sengupta, S. Hazelhurst, and M. Ramsay, “African genetic diversity provides novel insights into evolutionary history and local adaptations,” *Hum. Mol. Genet.*, vol. 27, no. R2, pp. R209–R218, 2018, doi: 10.1093/hmg/ddy161.

[102] A. Choudhury *et al.*, “Author Correction: High-depth African genomes inform human migration and health (Nature, (2020), 586, 7831, (741-748), 10.1038/s41586-020-2859-7),” *Nature*, vol. 592, no. 7856, p. E26, 2021, doi: 10.1038/s41586-021-03286-9.

[103] S. Choudhuri, “Phylogenetic Analysis, in Bioinformatics for Beginners,” in *Chapter 9*, Choudhuri., Academic Press, Oxford., 2014, pp. 209–218.

[104] J. Freudenstein, “Parsimony Methods in Phylogenetics,” in *ncyclopedia of*

Evolutionary Biology, Kliman RM., Academic Press, Oxford, 2016, pp. 220–224.

[105] A. Röhl, H. J. Bandelt, and P. Forster, "Median-joining networks for inferring intraspecific phylogenies.," *Mol. Biol. Evol.*, vol. 16, no. 1, pp. 37–48, 1999.

[106] C. Chaisiri, X. Liu, Y. Lin, Y. Fu, F. Zhu, and C. Luo, "Phylogenetic and haplotype network analyses of diaporthe eres species in china based on sequences of multiple loci," *Biology (Basel).*, vol. 10, no. 3, pp. 1–21, 2021, doi: 10.3390/biology10030179.

[107] G. Evanno, S. Regnaut, and J. Goudet, "Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study," *Mol. Ecol.*, vol. 14, no. 8, pp. 2611–2620, 2005, doi: 10.1111/j.1365-294X.2005.02553.x.

[108] D. Wen, Y. Yu, and L. Nakhleh, "Bayesian Inference of Reticulate Phylogenies Under the Multispecies Network Coalescent A Bayesian formulation The likelihood [Supplementary Material]," *PLoS Genet.*, pp. 1–25, 2016, doi: 10.5061/dryad.n2606.

[109] K. Erickson, "The jukes-cantor model of molecular evolution," *Primus*, vol. 20, no. 5, pp. 438–445, 2010, doi: 10.1080/10511970903487705.

[110] J. Yang *et al.*, "The advances in DNA mixture interpretation," *Forensic Sci. Int.*, vol. 301, pp. 101–106, 2019, doi: 10.1016/j.forsciint.2019.05.024.

[111] H. A. Shneewer, N. G. Al-loza, M. A. Kareem, and I. H. Hameed, "Sequence analysis of mitochondrial DNA hypervariable region III of 400 Iraqi volunteers," vol. 14, no. 26, pp. 2149–2156, 2015, doi: 10.5897/AJB2015.14489.

[112] Y. Zhou *et al.*, "Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions," *Heredity (Edinb).*, vol. 118, no. 5, pp. 503–510, 2017, doi: 10.1038/hdy.2017.5.

[113] C. Phillips *et al.*, "Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing," *Electrophoresis*, vol. 34, no. 8, pp. 1151–1162, 2013, doi: 10.1002/elps.201200621.

[114] J. Zhu, Y. Yu, and L. Nakhleh, "In the light of deep coalescence: Revisiting trees within networks," *BMC Bioinformatics*, vol. 17, no. Suppl 14, 2016, doi:

10.1186/s12859-016-1269-1.

[115] D. A. Morrison, “Genealogies: Pedigrees and Phylogenies are Reticulating Networks Not Just Divergent Trees,” *Evol. Biol.*, vol. 43, no. 4, pp. 456–473, 2016, doi: 10.1007/s11692-016-9376-5.

[116] Y. Yu, C. Jermaine, and L. Nakhleh, “Exploring phylogenetic hypotheses via Gibbs sampling on evolutionary networks,” *BMC Genomics*, vol. 17, no. Suppl 10, 2016, doi: 10.1186/s12864-016-3099-y.

[117] T. Agarwal, P. Gambette, and D. Morrison, “Who is Who in Phylogenetic Networks: Articles, Authors and Programs,” pp. 1–10, 2016, [Online]. Available: http://arxiv.org/abs/1610.01674.

[118] M. Richards *et al.*, “Tracing european founder lineages in the near eastern mtDNA pool,” *Am. J. Hum. Genet.*, vol. 67, no. 5, pp. 1251–1276, 2000, doi: 10.1016/S0002-9297(07)62954-1.

[119] L. Sá *et al.*, “Phylogeography of Sub-Saharan Mitochondrial Lineages Outside Africa Highlights the Roles of the Holocene Climate Changes and the Atlantic Slave Trade,” *Int. J. Mol. Sci.*, vol. 23, no. 16, pp. 1–11, 2022, doi: 10.3390/ijms23169219.

[120] V. Macaulay and M. Richards, “Median Networks : Speedy Construction and Greedy Reduction , One Simulation , and Two Case Studies from Human mtDNA,” vol. 16, no. 1, pp. 8–28, 2000, doi: 10.1006/mpev.2000.0792.

[121] H. Bandelt, “Time dependency of molecular rate estimates : tempest in a teacup,” pp. 1–2, 2008, doi: 10.1038/sj.hdy.6801054.

[122] V. M. Cabrera, “Human molecular evolutionary rate, time dependency and transient polymorphism effects viewed through ancient and modern mitochondrial DNA genomes,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–8, 2021, doi: 10.1038/s41598-021-84583-1.

[123] P. Forster, R. Harding, A. Torroni, and H. J. Bandelt, “Origin and evolution of native American mtDNA variation: A reappraisal,” *Am. J. Hum. Genet.*, vol. 59, no. 4, pp. 935–945, 1996.

[124] D. Mishmar *et al.*, “Natural selection shaped regional mtDNA variation in humans,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 1, pp. 171–176, 2003, doi: 10.1073/pnas.0136972100.

[125] P. Soares *et al.*, “Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock,” *Am. J. Hum. Genet.*, vol. 84, no. 6, pp. 740–759, 2009, doi: 10.1016/j.ajhg.2009.05.001.

[126] D. C. Culver, J. E. Kowalko, and T. Pipan, “Natural selection versus neutral mutation in the evolution of subterranean life: A false dichotomy?,” *Front. Ecol. Evol.*, vol. 11, no. January, pp. 1–10, 2023, doi: 10.3389/fevo.2023.1080503.

[127] J. J. Hublin *et al.*, “New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens,” *Nature*, vol. 546, no. 7657, pp. 289–292, 2017, doi: 10.1038/nature22336.

[128] C. Batini and M. A. Jobling, “The jigsaw puzzle of our African ancestry: Unsolved, or unsolvable?,” *Genome Biol.*, vol. 12, no. 6, pp. 1–4, 2011, doi: 10.1186/gb-2011-12-6-118.

[129] C. M. Vidal *et al.*, “Age of the oldest known Homo sapiens from eastern Africa,” vol. 601, no. March 2021, 2022.

[130] M. Cerezo *et al.*, “Comprehensive Analysis of Pan-African Mitochondrial DNA Variation Provides New Insights into Continental Variation and Demography,” *J. Genet. Genomics*, vol. 43, no. 3, pp. 133–143, 2016, doi: 10.1016/j.jgg.2015.09.005.

[131] M. C. Campbell, J. B. Hirbo, J. P. Townsend, and S. A. Tishkoff, “The peopling of the African continent and the diaspora into the new world,” *J. Int. Soc. Burn Inj.*, vol. 43, no. 5, pp. 909–932, 2017, doi: 10.1016/j.gde.2014.09.003.The.

[132] R. Nielsen *et al.*, “Tracing the peopling of the world through genomics,” vol. 541, no. 7637, pp. 302–310, 2017, doi: 10.1038/nature21347.Tracing.

[133] E. K. F. Chan *et al.*, “Human origins in a southern African palaeo-wetland and first migrations,” *Nature*, vol. 575, no. 7781, pp. 185–189, 2019, doi: 10.1038/s41586-019-1714-1.

[134] N. Brucato *et al.*, “Evidence of Austronesian Genetic Lineages in East Africa and South Arabia: Complex Dispersal from Madagascar and Southeast Asia,” *Genome Biol. Evol.*, vol. 11, no. 3, pp. 748–758, 2019, doi: 10.1093/gbe/evz028.

[135] M. Silva *et al.*, “60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2,” *Sci. Rep.*, vol. 5, no. March, pp. 1–13, 2015, doi: 10.1038/srep12526.

[136] P. A. Maier, G. Runfeldt, R. J. Estes, and M. G. Vilar, “African mitochondrial haplogroup L7: a 100,000-year-old maternal human lineage discovered through reassessment and new sequencing,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, 2022, doi: 10.1038/s41598-022-13856-0.

[137] C. Barbieri *et al.*, “Refining the Y chromosome phylogeny with southern African sequences,” *Hum. Genet.*, vol. 135, no. 5, pp. 541–553, 2016, doi: 10.1007/s00439-016-1651-0.

[138] M. H. Beltrame, M. A. Rubel, and S. A. Tishkoff, “Inferences of African evolutionary history from genomic data,” *Curr Opin Genet Dev.*, vol. 41, pp. 159–166, 2016, doi: doi:10.1016/j.gde.2016.10.002.

[139] D. M. Behar *et al.*, “The Dawn of Human Matrilineal Diversity,” *Am. J. Hum. Genet.*, vol. 82, no. 5, pp. 1130–1140, 2008, doi: 10.1016/j.ajhg.2008.04.002.

[140] V. Erný *et al.*, “Genetic structure of pastoral and farmer populations in the African Sahel,” *Mol. Biol. Evol.*, vol. 28, no. 9, pp. 2491–2500, 2011, doi: 10.1093/molbev/msr067.

[141] B. Lorente-Galdos *et al.*, “Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations,” *Genome Biol.*, vol. 20, no. 1, pp. 1–15, 2019, doi: 10.1186/s13059-019-1684-5.

[142] E. C. Ienco *et al.*, “May ‘ Mitochondrial Eve ’ and Mitochondrial Haplogroups Play a Role in Neurodegeneration and Alzheimer ’ s Disease ?,” vol. 2011, 2011, doi: 10.4061/2011/709061.

[143] R. Cann, “All about mitochondrial eve: an interview with Rebecca Cann. Interview

by Jane Gitschier.," *PLoS Genet.*, vol. 6, no. 5, 2010, doi: 10.1371/journal.pgen.1000959.

[144] T. M. K. Göbel *et al.*, "Mitochondrial DNA variation in Sub-Saharan Africa: Forensic data from a mixed West African sample, Côte d'Ivoire (Ivory Coast), and Rwanda," *Forensic Sci. Int. Genet.*, vol. 44, no. November 2019, p. 102202, 2020, doi: 10.1016/j.fsigen.2019.102202.

[145] B. M. Henn *et al.*, "Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 31, pp. 10693–10698, 2008, doi: 10.1073/pnas.0801184105.

[146] E. D'Atanasio *et al.*, "Rapidly mutating Y-STRs in rapidly expanding populations: Discrimination power of the Yfiler Plus multiplex in northern Africa," *Forensic Sci. Int. Genet.*, vol. 38, pp. 185–194, 2019, doi: 10.1016/j.fsigen.2018.11.002.

[147] T. Rito *et al.*, "A dispersal of Homo sapiens from southern to eastern Africa immediately preceded the out-of-Africa migration," *Sci Rep*, no. 9, pp. 28–47, 2019.

[148] N. Ansari Pour, C. A. Plaster, and N. Bradman, "Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people," *Eur. J. Hum. Genet.*, vol. 21, no. 4, pp. 423–429, 2013, doi: 10.1038/ejhg.2012.176.

[149] A. Olivieri *et al.*, "Mitogenome Diversity in Sardinians : A Genetic Window onto an Island ' s Past," vol. 34, no. 5, pp. 1230–1239, 2017, doi: 10.1093/molbev/msx082.

[150] M. Richards, V. Macaulay, and A. Torroni, "In Search of Geographical Patterns in European Mitochondrial DNA," vol. 1, pp. 1168–1174, 2002.

[151] C. Barbieri, M. Whitten, K. Beyer, H. Schreiber, M. Li, and B. Pakendorf, "Contrasting maternal and paternal histories in the linguistic context of Burkina Faso," *Mol. Biol. Evol.*, vol. 29, no. 4, pp. 1213–1223, 2012, doi: 10.1093/molbev/msr291.

[152] I. Lankheet, M. Vicente, C. Barbieri, and C. Schlebusch, "The performance of

common SNP arrays in assigning African mitochondrial haplogroups," pp. 1–8, 2021.

[153] F. Gomez, J. Hirbo, S. A. Tishkoff, K. M. Weiss, and B. W. Lambert, "Genetic Variation and Adaptation in Africa :," *Cold Spring Harb Perspect Biol*, pp. 1–21, 2014.

[154] M. Scientiae and W. Cape, "Gadean Brecht December 2019 Genetic analysis of mitochondrial DNA within Southern African populations . A thesis submitted in fulfilment of the requirements of Magister Scientiae in the Department of Biotechnology , University of the Western Cape . Superv," no. December, 2019.

[155] J. Nováčková *et al.*, "Subsistence strategy was the main factor driving population differentiation in the bidirectional corridor of the African Sahel," *Am. J. Phys. Anthropol.*, vol. 171, no. 3, pp. 496–508, 2020, doi: 10.1002/ajpa.24001.

[156] T. Kivisild *et al.*, "Ethiopian Mitochondrial DNA Heritage : Tracking Gene Flow Across and Around the Gate of Tears," pp. 752–770, 2004.

[157] M. Lipson *et al.*, "Ancient West African foragers in the context of African population history," *Nature*, vol. 577, no. 7792, pp. 665–670, 2020, doi: 10.1038/s41586-020-1929-1.

[158] B. Yunusbayev *et al.*, "The caucasus as an asymmetric semipermeable barrier to ancient human migrations," *Mol. Biol. Evol.*, vol. 29, no. 1, pp. 359–365, 2012, doi: 10.1093/molbev/msr221.

[159] S. Oppenheimer, "Out-of-Africa , the peopling of continents and islands : tracing uniparental gene trees across the map," pp. 770–784, 2012, doi: 10.1098/rstb.2011.0306.

[160] Q. Kong, Y. Yao, C. Sun, and C. Zhu, "Phylogeny of East Asian Mitochondrial DNA Lineages Inferred from Complete Sequences," no. 2002, pp. 671–676, 2003.

[161] D. E. Kelly *et al.*, "The Genetic and Evolutionary Basis of Gene Expression Variation in East Africans," *bioRxiv*, p. 2022.02.16.480765, 2022, doi: 10.1186/s13059-023-02874-4.

[162] F. Oldoni, K. K. Kidd, and D. Podini, "Microhaplotypes in forensic genetics,"

Forensic Sci. Int. Genet., vol. 38, pp. 54–69, 2019, doi: 10.1016/j.fsigen.2018.09.009.

[163] A. Torroni *et al.*, "Asian affinities and continental radiation of the four founding Native American mtDNAs," *Am. J. Hum. Genet.*, vol. 53, no. 3, pp. 563–590, 1993.

[164] C. A. de S. Luísa *et al.*, "Genetic variations among three major ethnic groups in Nigeria using RAPD," *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, 2020, doi: 10.1016/j.mgene.2020.100837.

[165] A. Salas *et al.*, "The African Diaspora: Mitochondrial DNA and the Atlantic Slave Trade," *Am. J. Hum. Genet.*, vol. 74, no. 3, pp. 454–465, 2004, doi: 10.1086/382194.

[166] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson, "African populations and the evolution of human mitochondrial DNA," *Science (80-.).*, vol. 253, no. 5027, pp. 1503–1507, 1991, doi: 10.1126/science.1840702.

[167] M. K. Gonder, H. M. Mortensen, F. A. Reed, A. De Sousa, and S. A. Tishkoff, "Whole-mtDNA Genome Sequence Analysis of Ancient African Lineages," 2006, doi: 10.1093/molbev/msl209.

[168] S. Oliveira *et al.*, "Matriclans shape populations : Insights from the Angolan Namib Desert into the maternal genetic history of southern Africa," no. December, pp. 1–18, 2017, doi: 10.1002/ajpa.23378.

[169] A. M. Oliveira, L. Gusmão, P. M. Schneider, and I. Gomes, "Detecting the Paternal Genetic Diversity in West Africa using Y-STRs and Y-SNPs," *Forensic Sci. Int. Genet. Suppl. Ser.*, vol. 5, pp. e213–e215, 2015, doi: 10.1016/j.fsigss.2015.09.085.

[170] L. Quintana-Murci *et al.*, "Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 5, pp. 1596–1601, 2008, doi: 10.1073/pnas.0711467105.

[171] L. Quintana-Murci *et al.*, "Strong Maternal Khoisan Contribution to the South

African Coloured Population: A Case of Gender-Biased Admixture," *Am. J. Hum. Genet.*, vol. 86, no. 4, pp. 611–620, 2010, doi: 10.1016/j.ajhg.2010.02.014.

[172] P. Brotherton *et al.*, "Supplementary information - Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans," *Nat. Commun.*, pp. 1–10, 2015.

[173] B. Ely, J. L. Wilson, F. Jackson, and B. A. Jackson, "African-American mitochondrial DNAs often match mtDNAs found in multiple African ethnic groups," *BMC Biol.*, vol. 4, pp. 1–14, 2006, doi: 10.1186/1741-7007-4-34.

[174] A. B. Lane *et al.*, "Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies," *Am. J. Phys. Anthropol.*, vol. 119, no. 2, pp. 175–185, 2002, doi: 10.1002/ajpa.10097.

[175] D. C. Johnson *et al.*, "Mitochondrial DNA diversity in the African American population," *Mitochondrial DNA*, vol. 26, no. 3, pp. 445–451, 2015, doi: 10.3109/19401736.2013.840591.

[176] M. M. Osman *et al.*, "Mitochondrial HVRI and whole mitogenome sequence variations portray similar scenarios on the genetic structure and ancestry of northeast Africans," *Meta Gene*, vol. 27, no. August 2020, p. 100837, 2021, doi: 10.1016/j.mgene.2020.100837.

[177] V. G. Olivares, A. M. Barrera, J. M. L. Salazar, A. I. Campos, R. G. Montelongo, and C. Flores, "OPEN A benchmarking of human mitochondrial DNA haplogroup classifiers from whole - genome and whole - exome sequence data," *Sci. Rep.*, pp. 1–11, 2021, doi: 10.1038/s41598-021-99895-5.

[178] Y. S. Chen, A. Torroni, L. Excoffier, A. S. Santachiara-Benerecetti, and D. C. Wallace, "Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups," *Am. J. Hum. Genet.*, vol. 57, no. 1, pp. 133–149, 1995.

[179] C. Fortes-Lima *et al.*, "Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt," *Mol. Biol. Evol.*, vol. 39, no. 10, pp. 1–19, 2022, doi: 10.1093/molbev/msac209.

[180] V. Montano *et al.*, “The Bantu expansion revisited: A new analysis of y chromosome variation in Central Western Africa,” *Mol. Ecol.*, vol. 20, no. 13, pp. 2693–2708, 2011, doi: 10.1111/j.1365-294X.2011.05130.x.

[181] S. Oliveira *et al.*, “The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola,” *Eur. J. Hum. Genet.*, vol. 27, no. 3, pp. 475–483, 2019, doi: 10.1038/s41431-018-0304-2.

[182] V. Fernandes *et al.*, “Genetic stratigraphy of key demographic events in Arabia,” *PLoS One*, vol. 10, no. 3, pp. 1–27, 2015, doi: 10.1371/journal.pone.0118625.

[183] N. M. Silva, T. Kivisild, A. Torroni, and R. Villems, “ARTICLE A ‘“ Copernican ”’ Reassessment of the Human Mitochondrial DNA Tree from its Root,” pp. 675–684, 2012, doi: 10.1016/j.ajhg.2012.03.002.

[184] D. M. Behar *et al.*, “The genome-wide structure of the Jewish people,” *Nature*, vol. 466, no. 7303, pp. 238–242, 2010, doi: 10.1038/nature09103.

[185] S. Davidovic, B. Malyarchuk, T. Grzybowski, and M. Stevanovic, “Complete mitogenome data for the Serbian population : the contribution to high-quality forensic databases,” 2020.

[186] T. Tau *et al.*, “Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017, doi: 10.1038/s41598-017-06365-y.

[187] M. Lucas-Sánchez, K. Fadhlaoui-Zid, and D. Comas, “The genomic analysis of current-day North African populations reveals the existence of trans-Saharan migrations with different origins and dates,” *Hum. Genet.*, vol. 142, no. 2, pp. 305–320, 2022, doi: 10.1007/s00439-022-02503-3.

[188] L. Pereira, L. Gusmão, C. Alves, A. Amorim, and M. J. Prata, “Bantu and European Y-lineages in Sub-Saharan Africa,” *Ann. Hum. Genet.*, vol. 66, no. 5–6, pp. 369–378, 2002, doi: 10.1046/j.1469-1809.2002.00130.x.

[189] A. Salas, H. Bandelt, V. Macaulay, and M. B. Richards, “Phylogeographic investigations : The role of trees in forensic genetics,” vol. 168, pp. 1–13, 2007, doi: 10.1016/j.forsciint.2006.05.037.

[190] G. Sirugo, S. M. Williams, and S. A. Tishkoff, “The Missing Diversity in Human Genetic Studies,” *Cell*, vol. 177, no. 1, pp. 26–31, 2019, doi: 10.1016/j.cell.2019.02.048.

[191] M. B. Richards, P. Soares, and A. Torroni, “Palaeogenomics: Mitogenomes and migrations in Europe’s past,” *Curr. Biol.*, vol. 26, no. 6, pp. R243–R246, 2016, doi: 10.1016/j.cub.2016.01.044.

[192] E. Watson, K. Bauer, R. Aman, G. Weiss, A. Von Haeseler, and S. Paabo, “mtDNA Sequence Diversity in Africa,” *Am. J. Hum. Genet.*, vol. 59, pp. 437–444, 1996.

[193] C. A. Lambert and S. A. Tishkoff, “Genetic structure in African populations: Implications for human demographic history,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 74, pp. 395–402, 2009, doi: 10.1101/sqb.2009.74.053.

[194] L. Quintana-murci, O. Semino, H. Bandelt, G. Passarino, K. Mcelreavey, and A. S. Santachiara-benerecetti, “Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa,” vol. 23, no. december, pp. 437–441, 1999.

[195] C. L. Hernández *et al.*, “Early holocenic and historic mtDNA African signatures in the Iberian Peninsula: The Andalusian region as a paradigm,” *PLoS One*, vol. 10, no. 10, pp. 1–24, 2015, doi: 10.1371/journal.pone.0139784.

[196] C. Z. Wang, X. E. Yu, M. Sen Shi, H. Li, and S. H. Ma, “Whole mitochondrial genome analysis of the Daur ethnic minority from Hulunbuir in the Inner Mongolia Autonomous Region of China,” *BMC Ecol. Evol.*, pp. 1–10, 2022, doi: 10.1186/s12862-022-02019-4.

[197] F. Gandini *et al.*, “Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes,” *Sci. Rep.*, vol. 6, no. January, pp. 1–13, 2016, doi: 10.1038/srep25472.

[198] M. Lucas-Sánchez, J. M. Serradell, and D. Comas, “Population history of North Africa based on modern and ancient genomes,” *Hum. Mol. Genet.*, vol. 30, no. R1, pp. R17–R23, 2021, doi: 10.1093/hmg/ddaa261.

[199] T. Falola and M. M. Heaton, *A history of Nigeria*. 2008.

[200] C. Gomes *et al.*, "Genetic insight into Nigerian population groups using an X-chromosome decaplex system," *Forensic Sci. Int. Genet. Suppl. Ser.*, vol. 7, no. 1, pp. 501–503, 2019, doi: 10.1016/j.fsigss.2019.10.067.

[201] I. T. Sabiu, F. A. Zainol, and M. S. Abdullahi, "HAUSA PEOPLE OF NORTHERN NIGERIA AND THEIR DEVELOPMENT," *Asian People J.*, vol. 1, no. 1, pp. 179–189, 2018.

[202] G. C. Uzomba, C. A. Obijindu, and U. K. Ezemagu, "Considering the lip print patterns of Ibo and Hausa Ethnic groups of Nigeria : checking the wave of ethnically driven terrorism," *Crime Sci.*, pp. 1–7, 2023, doi: 10.1186/s40163-023-00183-6.

[203] B. Martínez *et al.*, "Mitochondrial genetic profile of the Yoruba population from Nigeria," *Forensic Sci. Int. Genet. Suppl. Ser.*, no. September, pp. 0–1, 2019, doi: 10.1016/j.fsigss.2019.10.185.

[204] E. Paradis and K. Schliep, "Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R," *Bioinformatics*, vol. 35, no. 3, pp. 526–528, 2019, doi: 10.1093/bioinformatics/bty633.

[205] T. Hall, "BioEdit_a_user_friendly_biological_seque," *Nucleic Acids Symp Serie*, vol. 41, no. 8. pp. 95–98, 1999.

[206] K. Tamura, G. Stecher, and S. Kumar, "MEGA11: Molecular Evolutionary Genetics Analysis Version 11," *Mol. Biol. Evol.*, vol. 38, no. 7, pp. 3022–3027, 2021, doi: 10.1093/molbev/msab120.

[207] S. Xu, S. Gupta, and L. Jin, "PEAS V1.0: A package for elementary analysis of SNP data," *Mol. Ecol. Resour.*, vol. 10, no. 6, pp. 1085–1088, 2010, doi: 10.1111/j.1755-0998.2010.02862.x.

[208] J. W. Leigh and D. Bryant, "POPART: Full-feature software for haplotype network construction," *Methods Ecol. Evol.*, vol. 6, no. 9, pp. 1110–1116, 2015, doi: 10.1111/2041-210X.12410.

[209] B. M. Muhire, A. Varsani, and D. P. Martin, "SDT : A Virus Classification Tool

Based on Pairwise Sequence Alignment and Identity Calculation," *PLoS One*, vol. 9, no. 9, 2014, doi: 10.1371/journal.pone.0108277.

[210] M. Clement, D. Posada, and K. A. Crandall, "TCS: a computer program to estimate gene genealogies," *Mol. Ecol.*, no. 9, pp. 1657–1659, 2000.

[211] M. Dos Santos, A. Cabezas, M. P. Tavares, A. I. Xavier, and M. Branco, "TcsBU: A tool to extend TCS network layout and visualization," *Bioinformatics*, vol. 32, no. 4, pp. 627–628, 2016, doi: 10.1093/bioinformatics/btv636.

[212] A. Amorim, T. Fernandes, and N. Taveira, "Mitochondrial DNA in human identification : a review," vol. 2, 2019, doi: 10.7717/peerj.7314.

[213] B. Budowle, M. R. Wilson, J. A. Dizinno, C. Stauffer, M. A. Fasano, and M. M. Holland, "Mitochondrial DNA regions HVI and HVII population data," vol. 103, pp. 23–35, 1999.

[214] M. Y. Diallo *et al.*, "Circum-Saharan Prehistory through the Lens of mtDNA Diversity," *Genes (Basel).*, vol. 13, no. 3, 2022, doi: 10.3390/genes13030533.

[215] A. A. Adeyemo, G. Chen, Y. Chen, and C. Rotimi, "Genetic structure in four West African population groups," *BMC Genet.*, vol. 38, no. 6, pp. 1–9, 2005, doi: 10.1186/1471-2156-6-38.

[216] A. Sani and M. U. Mustapha, "Global Growing Impact of Hausa and the Need for its Documentation," *Contemp. J. Lang. Lit.*, vol. 1, no. September, 2018.

[217] A. A. Liman, "MEMORIALIZING A LEGENDARY FIGURE : BAYAJIDDA THE PRINCE OF BAGDAD IN HAUSA LAND," *Afrika Focus*, vol. 32, no. 1, pp. 125–136, 2019.

[218] K. Kim, D. Kim, and K. Kim, "Mitochondrial Haplogroup Classification of Ancient DNA Samples Using Haplotracker," vol. 2022, 2022.

[219] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, no. 22, pp. 4673–4680, 1994, doi: 10.1093/nar/22.22.4673.

[220] H. T. Jukes and R. C. Cantor, *Evolution of protein molecules*, vol. (I-XII + 4.

ACADEMIC PRESS, INC., 1978.

[221] K. R. Veeramah *et al.*, “Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria,” pp. 1–17, 2010.

[222] R. Gomaa, “Analysis of Mitochondrial DNA Variation in the Egyptian Population and Its Implications for Forensic DNA Analysis,” *PQDT - Glob.*, no. July, p. 242, 2010, [Online]. Available: https://manchester.idm.oclc.org/login?url=https://search.proquest.com/docview/2377668580?accountid=12253%0Ahttp://man-fe.hosted.exlibrisgroup.com/openurl/44MAN/44MAN_services_page?genre=dissertations+%26+theses&atitle=&author=Gomaa%2C+Raina&volume=&issue=.

[223] M. Merheb *et al.*, “Mitochondrial DNA, a powerful tool to decipher ancient human civilisation from domestication to music, and to uncover historical murder cases,” *Cells*, vol. 8, no. 5, 2019, doi: 10.3390/cells8050433.

[224] C. L. Holt, K. M. Stephens, P. Walichiewicz, K. D. Fleming, and E. Forouzmand, “Human Mitochondrial Control Region and mtGenome : Design and Forensic Validation of NGS Multiplexes , Sequencing and Analytical Software,” 2021.

[225] E. Tasker, B. LaRue, C. Beherec, D. Gangitano, and S. Hughes-Stamm, “Analysis of DNA from post-blast pipe bomb fragments for identification and determination of ancestry,” *Forensic Sci. Int. Genet.*, vol. 28, pp. 195–202, 2017, doi: 10.1016/j.fsigen.2017.02.016.

[226] M. D. Brandhagen, R. S. Just, and J. A. Irwin, “Validation of NGS for mitochondrial DNA casework at the FBI Laboratory,” *Forensic Sci. Int. Genet.*, vol. 44, p. 102151, 2020, doi: 10.1016/j.fsigen.2019.102151.

[227] N. Saitou and M. Nei, “The Neighbour-joining method: a new method for reconstructing phylogenetic trees.,” *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, 1987, doi: 10.1093/oxfordjournals.molbev.a040454.

[228] J. R. Luis, H. Lacau, K. Fadhlaoui-Zid, M. A. Alfonso-Sanchez, R. Garcia-Bertrand, and R. J. Herrera, “Afghanistan: conduits of human migrations identified

using AmpFlSTR markers," *Int. J. Legal Med.*, vol. 133, no. 6, pp. 1659–1666, 2019, doi: 10.1007/s00414-019-02018-z.

[229] M. E. Prendergast and E. A. Sawchuk, *Genetics and the African Past What is Ancient DNA and How is It Studied ?*, no. October. 2022.

[230] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.," *Nucl. Acids. Symp. Ser.*, vol. 41, pp. 95–98, 1994.

[231] F. Tajima, "Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism," *Genetics*, vol. 123, pp. 585–595, 1989.

[232] A. R. Templeton, K. A. Crandall, and S. C. F., "A Cladistic Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping and DNA Sequence Data. III. Cladogram Estimation," *Genetics*, vol. 132, no. 2, pp. 619–633., 1992.

PAPER PUBLICATION

Indian Journal of Forensic Medicine and Pathology
Volume 14 Number 2, April - June 2021 (Special Isuue)
DOI: http://dx.doi.org/10.21088/ijfmp.0974.3383.14221.53

REVIEW RATICLE

African Y-STR Haplotyping and Y-Chromosome Profiling: A Review

Ibrahim El-ladan Shehu[1], Priyanka Chhabra[2]

ABSTRACT

Forensic genetics is an indispensable tool in forensic analysis that uses genetic evidence in crime investigation and/or identification of missing individuals or victims of mass disaster. It is highly reliable and can be used to substantiate evidence to prove the guilt or innocence of a suspect in question. There is paucity of data on African forensic DNA profiling. This is partly due to lack of funding and expertise. Moreover, there are very limited forensic genetic commercial kits that incorporate markers that are specific for African populations, markers that will provide highly specific information on the African Y-STR markers. Therefore, the purpose of this study is to consolidate the published Y-STR data of African population for forensic and population genetic reference. The review presents the Y-haplotype and genetic diversity of African male population. The review dissects the data into different regions of the African continent, viz., the Northern, Southern, Eastern, Western and Central African regions.

KEYWORDS | forensic genetics, dna profiling, y-str, y-haplotype, africa

Authors' Affiliations:
[1]*Researech Scholar,*
[2]*Assistant Professor, School of Basic and Applied Sciences, Galgotias University, Greater Noida, 201310, Uttar Pradesh, India.*

Corresponding Author:
Priyanka Chhabra, Assistant Professor, School of Basic and Applied Sciences, Galgotias University, Greater Noida 201310, Uttar Pradesh, India.

Email:
pchhabra188@gmail.com

INTRODUCTION

The use of DNA profiling in criminal cases was first used a little above 30 years ago. The pioneer of this work was Professor Sir Alec Jeffreys, whose groundbreaking forensic work was able to link the assault and murder of two young girls to Colin Pitchfork in 1983 and 1986. The case served as the landmark criminal case that gave birth to the use of DNA fingerprinting in the criminal justice system.[1]

The Y-chromosome is male-specific in humans and follows a strict mode of paternal inheritance. It comprises of a major non-recombining region (NRY) that makes it suitable to providing one of the highest resolution tools for studying human population genetics.[2] This high resolution provides it with high discrimination power between individuals suitable for forensic investigations involving male victims or suspects.

The number of multiple alleles that are remarkably differentiated between individuals by the number of repeat units on Y-Chromosome is referred to as Y-Chromosomal Short Tandem Repeats (Y-STRs).[3] The forensic use of Y-STR genotyping has become instrumental in the identification of males involved in sexual assault, paternity and ancestral determination, missing and disaster victims investigations.[4] High mutation rates in Y-STR markers referred to as rapidly mutating Y-STRs or RM Y-STRs have been reported recently.[3] Some commercially available forensic analysis kits, for example, Y-Filer plus amplification kit of Thermo Fisher Scientific, USA, have introduced the RM Y-STRs with anticipation that they will assist in discriminating close male

How to cite this article
Ibrahim El-Ladan Shehu. African Y-STR Haplotyping and Y Chromosome Profiling: A Review. Indian J Forensic Med Pathol. 2021;14(3 Special):379-385.

The African continent is inhabited by people with enormous linguistic, cultural and genetic diversity of more than 2,000 different languages and ethnic populations. The demographic timeline of the continent has recorded oscillations in population size, admixture, long and short-term migrations leading to rich and diverse variations and in modern populations. It can be said that the most genetically diverse region of the world is Africa.[5] Africa by size and population is the second-largest continent, comprising numerous countries and diverse populations. However, the information available on the Y-STR haplotype and allele frequencies in these populations is very little.[6]

There are several databases that the scientific community use to compare their data with already published data, forensic laboratories and other security agencies also use the databases, the most common databases include: YHRD and with Relia-Gene database and PowerPlex Y Haplotype Database.[7] The data provided in this review will help in consolidating the valuable information on the haplotype and allele frequencies of African Y-STR profile for population genetics and forensic reference.

Y-haplotype and Genetic Diversity in Central Africans

Studies on Central African population conducted by Arroyo-Pardo *et al.,* (2005) studied 16 Y-STR loci in 101 male samples of Equatorial Guinea origin who live in Madrid, Spain. Of the 101 studied individuals, 94 different haplotypes were obtained in the study. Another study conducted on 873 samples from Gabon and Cameroon from Central Africa by typing 18 Y-STRs found a total of 728 different haplotypes indicating high discrimination between the populations. They also observe high frequency of modal Bantu haplotype and its one-step neighbors as described by other literature in all the 24 Bantu populations in the study.[8] They also observed modal Bantu haplotype in two pygmy samples from Gabon and one of its one-step neighbors in one pygmy sample from Cameroon. Multi-Dimensional Scale Plot (MDS) showed all the Bantu population clustering together obviously separated from the Pygmies, indicating population homogeneity among the Bantus and some population admixture between the two populations.

Another study on Y STR among 165 Bantu population living on Bangui, Central African Republic, was conducted by Lecerf et al., (2007) who reported 88 different haplotypes, of which 83 were unique. The authors observed DYS385 and DYS392 to have the highest (0.9305) and lowest (0.1685) gene diversity (GD), respectively.

Data of Y-haplotype in Fang and Bubi populations from Bioko (Equatorial Guinea) were studied by Barrot et al., (2007), of the 133 samples studied in Bubi population, 102 and 87 different and unique haplotypes were reported respectively. Gene diversity in the study revealed DYS385 as the most polymorphic system.

The data generated by the researchers on the Y-haplotype diversity among Central Africans revealed that the populations are ideal for forensic casework due to their unique haplotype profile. It has also been established that DYS385 is the most polymorphic marker in the Central African population. It has also been revealed that the Central African population has intermediate allele 13.2 at DYS385 locus.

Y-Haplotype and Genetic Diversity in East Africans

Eighteen Y-STR profile of the population living in Maputo from Mozambique was characterized by Alves et al., 2003 who reported 101 defined haplotype out of a total of 112 studied samples, two individuals with seven shared haplotypes, while the most frequent was shared by five individuals. The most diverse and less polymorphic loci were DYS385 and DYS392 respectively.[11]

Twelve Y-STR of 40 Karimojong males from Karamoja, Uganda revealed 32 different haplotypes with high discrimination power. Comparison with the Y-STR data of Uganda, Mozambique, Cabinda and Equatorial Guinea revealed a large genetic distance between the populations.[12] Another study of 17 Y-STR haplotypes involving 118 males from the Nilotes population of Karamoja region in Uganda by Gomes et al., (2010) reported 94 different haplotypes, a total of 19 shared, 14 and 5 in two and three individuals respectively.[13]

Another study employing 27 Y-STR haplotypes among the Tigray populations of Northern Ethiopia by Haddish et al., (2019) revealed that the recent expansion of Yfiler to

study 27 loci produced a haplotype diversity and high discrimination capacity of 100%. Seventeen Y-chromosomal STR haplotypes in 69 Rwanda-Hutu unrelated male individuals from East Central Africa revealed a total of 62 unique haplotypes out of the 69 individuals.[6] The authors reported the lowest and highest gene diversity at locus DYS392 and DYS385 respectively.

Based on the data generated by the researchers, it is revealed that the highly discriminating or rather highly polymorphic allele in the East African population is DYS385 making it highly relevant in discriminating East African populations using the allele in forensic casework. It has also been established that DYS392 is at least a polymorphic marker in East African populations. The data presented by the researchers have demonstrated the uniqueness of the East African population with very few shared haplotypes among the populations which are of high forensic relevance.

Y-haplotype and Genetic Diversity in North Africans

A study on 185 individual Y-STR haplotype among four different populations: Southern Moroccan Berbers, Mozabites, Moroccan Arabs and Saharawis from North West Africa. The most informative markers in the populations were DYS390, DYS389II, DYS389I and DYS391 in the respective order of magnitude, the highest and lowest haplotype diversity were observed in Moroccan Berbers and Mozabites respectively.[14]

The first study on Y-STR of the Tunisian population was conducted by El Khil et al., (2001) who studied six Y-STR loci among 135 males from different ethnic groups: Berbers, Blacks of Jerba Island and Arabs. There was no significant difference between the Arabs and the Berbers except on locus DYS390. Contrastingly, a significant difference between Blacks and the other two islander groups was observed except for the locus DYS39I.[15]

In another study of 135 Jerban males, 67 different haplotypes were reported: 33 haplotypes out of 42 Jerban of Sub-Saharan Africa, 27 haplotypes out of 46 Arabs and 18 haplotypes out of 47 Berbers.[16] The study population had a haplotype diversity ranging from 0.987 to 0.827 where the Jerbans of Sub-saharan origin had the uppermost value, whereas the Berbers had the lowest value.

Another study on 13 Y-STR of 105 Southern Tunisian population by Ayadi et al., (2006) identified 81 different haplotypes, out of which 67 were unique, the most frequent haplotype was shared by five individuals in the study population.[17] The loci with the highest and lowest polymorphism are DYS385a/b and DYS436 respectively. In another study by Onofri et al., (2008) in Northern African populations: 52 Tunisian and 51 Moroccan samples, a lower haplotype diversity of 29 unique haplotypes out of 39 different haplotypes were observed in the Tunisian population, while higher haplotype diversity of 44 unique out of 47 different haplotypes in Moroccan population were reported.[16]

A study on a total of 267 Moroccan ethnic populations (Sahrawi, n=68, Berber-speaking, n=69 and Arab-speaking, n=130) revealed a total of 257 (96.25%) different haplotypes, out of which 10 alleles were found in two individuals each, 237 unique haplotypes were observed.[19] Highest Gene Diversity was recorded on alleles DYS385 (0.887) and DYS458 (0.820), the discrimination capacity (DC) and Haplotype Diversity (HD) were 0.963 and 0.9991 respectively.[19] Another study by Palet et al., (2010) on the Moroccan population from Figuig Oasis revealed 52 different haplotypes, and 36 unique in an overall total population of 96. The loci with the highest and lowest respective diversity were DSY458 and DYS392.

Another study on Berber and Arab-speaking populations in Morocco by[20] reported 74 different haplotypes out of the total 85 individuals. A non-significant difference in gene diversity between the Arab-speaking samples having higher (0.566) than the Berbers (0.472) was reported by the authors. The high polymorphic alleles in Arab-speaking and Berbers were DYS385 and DYS458 respectively, while the lowest polymorphic marker was DYS392 in both the populations. Two new alleles of DYS458 locus were observed in one Berber and one Arab.[20]

A study on 17 Y-STR of 208 individuals from South(Upper) Egypt reported 204 different haplotypes, of which 200 were unique and 4 alleles were found twice each.[21] The most polymorphic allele was DYS385a/b followed by DYS458. Another study on Y-STR of 238 Benghazi

population, East Libya revealed a total of 238 different haplotypes, out of which 214 were unique, and 24 shared haplotypes.[22] The most polymorphic loci were DYS385a/b and DYS458 with haplotype diversity of 0.82 and 0.73 respectively.[22] Another study on Libyan population by Triki-Fendri *et al.,* (2013) revealed 142 different haplotypes out which 124 were unique in a total population of 176 individuals.[23]

D'Atanasio *et al.,* (2019) studied the discrimination power of the Y-Filer Plus multiplex kit in 11 North African populations from Egypt, Libya, Algeria and Morocco. The authors observed null alleles at three different loci (FYF387S1, DYS448 and DYS389II). They determined the genetic diversity (GD) with the exclusion of the null alleles and observed DYS385 and DYS481 with the highest values of 0.86 and 0.85 respectively. DYF387S1 showed the fourth highest value comparable to the GD values of RM Y-STRs which the authors attributed to an observed low GD value in the Algerian population under study.[24]

The most polymorphic loci in North African populations are DYS385 and DYS458, while the least polymorphic locus in all the North African populations was DYS392. Some Tunisian populations revealed DYS710 as the most polymorphic marker. Intermediate alleles were observed in some of the North African populations. The North African populations also exhibited unique haplotype diversity even after search and comparison on Y-STR databases which makes them suitable for forensic casework.

Y-haplotype and Genetic Diversity in West Africans

A study on the population of Guinea Bissau was conducted by.[25] The authors studied Y-STR population data of 215 unrelated healthy males whose ancestors were known to have lived in Guinea Bissau for three generations. The authors observed that the range of the studied loci and the allele frequencies are similar to the ones observed in other Sub-saharan Africa. The authors noted high prevalence of alleles 11 for DYS392 (88%), 14 for DYS437 (72%), 11 for DYS438 (65%), 21 for DYS390 (67%), and 15 for DYS19 (42%). The highest genetic diversity was observed in DYS19 and DYS389II (0.7182 and 0.7239) respectively, while the highest haplotype diversity was observed in DYS385 (0.9031). One hundred and fifty-four distinct haplotypes were observed in 161 fully typed individuals.

Benin and Ivory Coast ethnic populations were studied by Fortes-Lima et al., (2015). The Authors studied 288 individuals. The data from the research showed 30 minimum haplotypes and a total of 45 Y-filer in Yoruba as well as 34 minimum haplotypes and 44 Y-filer in Bariba population. The Yoruba and the Bariba exhibited high genetic diversity values of 0.9937 and 0.9929 respectively.[26]

A study was conducted at the Institute of Legal Medicine, Cologne, Germany. Individuals from different countries of West Africa (Nigeria, Gambia, Niger, Senegal, Benin, Togo, Sierra Leone, Ghana, Ivory Coast and Liberia) were selected for the study.[5] High values of haplotype diversity (1.0000 ± 0.0018) were observed in 86 samples under study, the values obtained were similar to what was reported by other studies.[5]

Y-STR profile of 142 individuals from the three largest ethnic groups in Nigeria (Yoruba, Hausa and Igbo) was studied by Martinez et al., (2017). 140 different haplotypes were observed, comprising of two individuals with two shared haplotypes. The authors reported an increase in the number of shared haplotypes when Y-filer kit was used: four and one haplotypes shared by two and three individuals respectively.[27]

The data presented on the genetic profile of West African populations revealed low genetic diversity with very few shared haplotypes among the individuals which translates to population homogeneity. Upon comparison with other populations on YHRD, very few matches were obtained. DYS385 was found to be the most polymorphic allele in some West African populations.

Y-haplotype and Genetic Diversity in Southern Africans

A study was conducted by Sánchez-Diz P *et al.,*[28] on African population groups from Mozambique. The authors studied a sample of 308 unrelated healthy individuals from the following groups: Nguni, Rongas, Senas, Changanes, Nhungwes, Tswas, Macondes, Chopes, Yao, Bitongas, Chuabos, Shonas, Lomwe, Ndaus, Makuas and

Nyanjas. Lower gene diversity was observed on DYS391 and DYS392 in all the populations under study. Of the total 308 samples studied, only 126 different haplotypes were observed, with the most frequent haplotype present in 22 samples. The observed average haplotype diversity was 97%.

Study on individuals living in KwaZulu-Natal and Western Cape provinces in South Africa was conducted by Lea N *et al*.[29] Three subpopulations (88 Xhosa, 101 English Speaking Caucasian and 77 Asian Indian males) were recruited for the study. Of the total population, 77, 101 and 73 different haplotypes were observed in Xhosa, Caucasian and Asian Indian populations. The number of alleles ranges from three (DYS391 and DYS392) to 21 in DYS710, while the average gene diversity ranged from 0.32 in DYS391 to 0.89 in DYS711 loci. The authors reported DYS710, DYS711, DYS712, DYS713, DYS7114 as novel markers and among the most variable markers.[29]

In another study conducted by D'Amato et al., (2008) on 99 indigenous Xhosa, 100 Caucasian English, 86 Asian Indian, 114 mixed "colored" and 107 Caucasian Afrikaan populations. Of the 506 individuals, 394 different haplotypes were observed, shared haplotypes were observed in 33 individuals. The allele frequency and haplotype diversity in 54 Ovambo male population in Namibia were carried out.[30] The study was conducted on 28 Y-STRs, where a total of 51 different haplotypes and 48 unique haplotypes were observed. Three shared haplotypes were also observed in two individuals. DYS385 and DYS392 had the highest (0.9000) and lowest (0.036) respective diversity values.

Analysis of 17 Y-STR loci in 105 healthy, unrelated Muslim populations of Cape Town, South Africa was conducted.[31] Eighty-three, 102 and 89 Asian-Indian, European-English and native Xhosa respectively were used for the comparison. The most polymorphic (0.958) and least polymorphic (0.449) loci based on GD values reported were DYS385 and DYS391 respectively. Ninety-one unique haplotypes and DC values of 0.866 were observed when considering the nine minimal haplotype Y-STRs, while in the case of the remaining eight loci.

The Y-STR haplotypes in three ethnic groups of Angola were studied by Melo *et al.* (2011). The authors studied 11 Y-STR haplotypes from a total of 166 individuals of three main ethnolinguistic groups of Angola: 53 Ovimbundo, 57 Bakongo and 56 Kimbundo populations. The Ovimbundo ethno-linguistic group showed 39 and 46 different and unique haplotypes respectively with two shared haplotypes that appear twice and one shared haplotype that appeared three times. Fifty-three and 49 different and unique haplotypes were observed respectively in the Bakongo group, four shared haplotypes were observed twice in the Bakongo group. In the Kimbundo group, 53 and 50 different and unique haplotypes were reported, while three shared haplotypes were observed twice. The most polymorphic locus was DYS385. Of the total of 166 individuals, 138 and 120 different and unique haplotypes were observed respectively.

The first study on Y-STR in Botswana population was conducted by Tau *et al.,* (2015), the authors studied 17 Y-STR profiles of 252 individuals among Botswana population: The authors clustered the samples into two regions: Northern [North and North-Western (1 San, 1 Sobea, 1 Herero, 6 Yeyi, 3 Mbukushu) n=12] and Southern Botswana [South and South East(1 Pedi, 2 Ndebele, 5 Tswapong, 11 Birwa, 8 Kgalagadi, 24 Kalanga, and 189 Tswana) n= 240]. The authors observed Haplotype Diversity and Discrimination capacity of 0.9990 and 0.9444 respectively and 238 unique haplotypes. The most common haplotype was observed five times in the populations except in the Tswana and Mbukushu that had four and one most frequent haplotypes.

D'Amato & Kasu, (2017) designed a highly discriminating Y-STR kit to preferentially target and amplify African samples, this genetic tool has been developed to a commercial prototype called UniQTyper Y-10. It was made up of 10 Y-STR loci markers including four RM Y-STRs. The authors studied 957 individuals from native and immigrant South African ethnic populations: English, Afrikaan, Indian, Admixed and native Bantu groups such as Venda, Pedi, Xhosa and Zulu. Of the total 957 studied samples, 870 unique haplotypes were observed with an overall Discrimination Capacity of 0.909. Another study conducted by Lesaoana et al., (2019) also used UniQTyper Y-10 to type the Y-STR profile of 938

individuals in five Bantu ethnic groups living in 10 Lesotho districts composed of South, North and Central regions. The authors reported 698 and 588 different and unique haplotypes respectively. A total of 350 individuals shared 110 haplotypes, the most frequent haplotype was shared by 28 individuals. The same haplotype was also reported to have been observed in 17 unrelated samples in Northern South Africa.

Another study on 27 Y-STR profiles of 200 unrelated individuals of Shona ethnic group of Zimbabwe in Harare province was conducted by Shonhai et al., (2020a) using 5-dye SureID 27Y kit. The authors reported only 159 complete 27 loci profiles, in response to that, the authors downgraded the loci to 12 Y-STR of PowerPlex. A total of 154 unique haplotypes out of the 159 were observed, two and one haplotype appeared twice and four times respectively. With a high genetic diversity depicted by the haplotype diversity of 0.9994, the overall DC of the population was 0.9686 while haplotype match probability was computed as 0.0069. The authors performed a single locus analysis of the whole Y-STR profile where they reported several observations which included but were not limited to triallelic pattern for locus DYS387S1, microvariant allele markers at DYS387S1 and DYS385. The lowest GD values were observed at loci DYS392 (0.03748) and DYS437 (0.096702). Meanwhile, DYS449, DYS481, and DYS518 had the highest GD values of 0.867239, 0.85042, and 0.825179, respectively.

In another study by Shonhai et al., (2020b) on Zimbabwean Shona brother pairs using the same kit used by Shonhai et al., (2020a). Only four brother pairs out of the 18 pairs were distinguishable based on the variation of allele numbers on only one allele marker among the 27 loci studied in addition to RM Y-STR DYS518. The authors observed loci DYS481 and DYS518 to have the highest GD with values of 0.8252 and 0.8502 respectively. Although the authors described the loci DYS393 and DYS458 as mini Y-STRs, they could not clearly explain the reason why variation between the brothers was observed in the markers. It is clear that the kit is not suitable for discriminating between related individuals. However, it can be noted that the kit was discriminatory between unrelated male Shona population of South Africa.[36]

The studies conducted on the Southern African populations have revealed that the most polymorphic markers reported were DYS385 and DYS710 in some populations. DYS391 and DYS392 were reported with the lowest GD values in the South African populations. Analysis of molecular variance in the study population revealed variations within the study groups. Searches on YHRD and Applied Biosystem databases presented very few numbers of shared haplotypes with other African populations. Additionally, very few shared haplotypes were observed within the study samples. The high frequency of unique haplotypes in the populations has highlighted the suitability of Southern African populations for forensic casework.

CONCLUSION

It can be concluded based on the review that the African populations are unique populations with high discrimination haplotypes, thus making them unique for forensic reference. DYS385 is the most polymorphic allele, intermediate allele 13.2 at the same locus was observed in Central African populations. The most informative marker in East African is also DYS385, while DYS392 is the least polymorphic locus. The North African populations bear DYS385 as the most polymorphic marker in addition to DYS458, while DYS392 is also the least polymorphic marker. However, the Tunisian population has exhibited DYS710 as the most informative marker. DYS385 was also the most informative marker in West African populations. DYS385 is also the most informative marker in South African populations in addition to DYS710 in small populations as observed in Tunisian populations of Central Africa. DYS391 and DYS392 are as well the least polymorphic markers in the South African populations. However, based on the number of African populations and the number of countries where the studies were conducted, it can be said that the African genetic data for forensic reference is under-represented. There is also no dedicated African database for forensic reference.

It is recommended that African nations should embrace the use of DNA forensics to minimize the rate of crimes in their countries. The need for

African nations to embrace the use of Combined DNA Information System (CODIS) cannot be over-emphasized, as this will help in solving many mysterious criminal cases involving mass disaster victims, victims of rape, murder, and other criminal cases. Based on the data generated from the review, it is visible that there is a need to develop and validate Y-STR kits that will primarily target and amplify the unique African haplotypes.

RECOMMENDATIONS

It is recommended that Africans should embrace the DNA forensics to minimize the rate of crimes in Africa. The need for Africa to embrace the use of Combined DNA Information System (CODIS) cannot be over-emphasized, as this will help in solving many criminal cases involving mass disaster victims, victims of rape, murder, and other criminal cases. Based on the data generated from the review, it is visible that there is a need to develop and validate Y-STR kits that will primarily target the unique African haplotypes. IJFMP

Acknowledgement:
The authors have made no acknowledgment in this article.
Conflict of Interest:
The authors declare that there is no commercial or financial links that could be construed as conflict of interests.
Source of Funding:
The author declares that there is no funding for this project.

REFERENCES

1. **Linacre AMT.** *Forensic sciences | DNA profiling. Encycl Anal Sci. 2019;4(March 2018):17–22.*
2. **Singh M, Sarkar A, Nandineni MR.** *A comprehensive portrait of Y-STR diversity of Indian populations and comparison with 129 worldwide populations. Sci Rep. 2018;8(1):2–8.*
3. **Watahiki H, Fujii K, Fukagawa T, Mita Y, et al.** *Polymorphisms and microvariant sequences in the Japanese population for 25 Y-STR markers and their relationships to Y-chromosome haplogroups. Forensic Sci Int Genet [Internet]. 2019;41(February):e1–7. Available from: https://doi.org/10.1016/j.fsigen.2019.03.004*
4. **Kasu M, Fredericks J, Fraser M, Labuschagne C, Lesaoana M, D'Amato ME.** *Novel Y-chromosome short tandem repeat sequence variation for loci DYS710, DYS518, DYS385, DYS644, DYS612, DYS626, DYS504, DYS481, DYS447 and DYS449. Int J Legal Med. 2019;133(6):1681–9.*
5. **Oliveira AM, Gusmão L, Schneider PM, Gomes I.** *Detecting the Paternal Genetic Diversity in West Africa using Y-STRs and Y-SNPs. Forensic Sci Int Genet Suppl Ser [Internet]. 2015;5:e213–5. Available from: http://dx.doi.org/10.1016/j.fsigss.2015.09.085*
6. **Balamurugan K, Duncan G.** *Y chromosome STR allelic and haplotype diversity in a Rwanda population from East Central Africa. Leg Med [Internet]. 2012;14(2):105–9. Available from: http://dx.doi.org/10.1016/j.legalmed.2011.12.002*
7. **Arroyo-Pardo E, Gusmão L, López-Parra AM, Baeza C, Mesa MS, Amorim A.** *Genetic variability of 16 Y-chromosome STRs in a sample from Equatorial Guinea (Central Africa). Forensic Sci Int. 2005;149(1):109–13.*
8. **Berniell-Lee G, Bosch E, Bertranpetit J, Comas D.** *Y-chromosome diversity in Bantu and Pygmy populations from Central Africa. Int Congr Ser. 2006;1288:234–6.*
9. **Lecerf M, Filali M, Grésenguet G, Ndjoyi-Mbiguino A, Le Goff J, de Mazancourt P, et al.** *Allele frequencies and haplotypes of eight Y-short tandem repeats in Bantu population living in Central Africa. Forensic Sci Int. 2007;171(2–3):212–5.*
10. **Barrot C, Sánchez C, Xifró A, Ortega M, Mas J, Huguet E, et al.** *Data for Y-chromosome haplotypes in Fang and Bubi populations from Bioko (Equatorial Guinea). Forensic Sci Int. 2007;168(1):10–2.*
11. **Alves C, Gusmão L, Barbosa J, et al.** *Evaluating the informative power of Y-STRs: A comparative study using European and new African haplotype data. Forensic Sci Int. 2003;134(2–3):126–33.*
12. **Gusmão L, Sánchez-Diz P, Gomes I, Alves C, Carracedo Á, João Prata M, et al.** *Genetic analysis of autosomal and Y-specific STRs in the Karimojong population from Uganda. Int Congr Ser. 2006;1288:213–5.*
13. **Gomes V, Alves C, Amorim A, Carracedo Á, et al.** *Haplotype data defined by 17 Y-chromosome STRs. Forensic Sci Int Genet. 2010;4(4).*
14. **Bosch E, Calafell F, Pérez-Lezaun A, Comas D, Izaabel H, Akhayat O, et al.** *Y chromosome STR haplotypes in four populations from northwest Africa. Int J Legal Med. 2000;114(1–2):36–40.*
15. **El Khil HK, Marrakchi RT, Loueslati BY, et al.** *Y chromosome microsatellite variation in three populations of Jerba Island (Tunisia). Ann Hum Genet. 2001;65(3):263–70.*
16. **Khodjet El Khil H, Marrakchi RT, et al.** *Distribution of Y chromosome lineages in Jerba island population. Forensic Sci Int. 2005;148(2–3):211–8.*
17. **Ayadi I, Ammar-Keskes L, Rebai A.** *Haplotypes for 13 Y-chromosomal STR loci in South Tunisian population (Sfax region). Forensic Sci Int. 2006;164(2–3):249–53.*
18. **Onofri V, Alessandrini F, Turchi C, et al.** *Y-chromosome markers distribution in Northern Africa: High-resolution SNP and STR analysis in Tunisia and Morocco populations. Forensic Sci Int Genet Suppl Ser. 2008;1(1):235–6.*
19. **Aboukhalid R, Bouabdellah M, et al.** *Haplotype frequencies for 17 Y-STR loci (AmpFISTR®Y-filerTM) in a Moroccan population sample. Forensic Sci Int Genet. 2010;4(3):2009–10.*
20. **Gaibar M, Esteban E, Harich N, et al.** *Genetic differences among North African Berber and Arab-speaking populations revealed by Y-STR diversity. Ann Hum Biol. 2011;38(2):228–36.*

CURRICULUM VITAE

IBRAHIM SHEHU EL-LADAN

Nationality:	Nigerian
Indian Address:	Saffron Living, Knowledge Park 3, Greater Noida, India
Permanent home address:	No. 60, Yusra Estate, Federal Mortgage Housing, Katsina. Katsina State, Nigeria.
Work Address:	Umaru Musa Yar'adua University, Katsina, Katsina State, Nigeria.
Mobile no:	+91 9311248859, +2348035120742
Email:	Ibrahim.el_phd20@galgotiasuniversity.edu.in: ibrahim.shehu@umyu.edu.ng

A. Education and Qualifications

Post-secondary education

i. Galgotias University, Uttar Pradesh, India. PhD Forensic Science (in-view) 2020 to date

Courses taken and passed:

1. Forensic Biology
2. Instrumentation Analysis
3. Research and Publication Ethics
4. Research Methodology
5. Statistical Methods in Research
6. Comprehensive Exam

ii. Ahmadu Bello University, Zaria, Nigeria. MSc Human Anatomy. **(CGPA of 4.09, 23rd May, 2016)**.

Courses taken and marks obtained:

1. Advance Biological Anthropology (**Conflated Mark: 72%**)
   The course reviewed different aspects of biological anthropology such as forensic anthropology, nutritional adaptation, ethnic groups of man, primate classification and primate behaviour, dating techniques, geological time scale, skin colour adaptation, heat and cold adaptation, global warming etcetera. The assessment was based on two written examinations.

2. Biometry (**Conflated Mark: 76%)**
   The course covered aspects mathematical and statistical methods employed analysing scientific data. It involved theoretical statistics and its application in analysing data using SPSS. An unseen written exam and written tests were used in the assessment.

3. Clinical and Applied Anatomy (**Conflated Mark: 72%)**
   The course reviewed the gross anatomy of the human body in relation to clinical and applied anatomy. PowerPoint Oral presentations were conducted on various topics. I presented on **the clinical anatomy of the brachial plexus.** 2 unseen written examination were conducted in the assessment.

4. Molecular Genetics (**Conflated Mark: 62%**)
   The course reviewed the overview of molecular genetics and theoretically introduced the advance molecular techniques such as Gel Electrophoresis, Karyotyping, PCR, DNA sequencing techniques and Blotting techniques. Two unseen written examination were conducted.

5. Cell and Tissue Culture (**Conflated Mark: 60%)**
   The course introduced different types of culture and culture media. The protocols and safety measures in maintaining the cell and tissue cultures. The methods used in determining and counting of viable cells, examination of cell and tissue cultures via visual, microscopic, histochemical and immunohistochemical techniques. The course also introduced the application of different cell lines in biomedical research. The assessment was based on two written examinations.

6. Molecular Cell Biology (**Conflated Mark: 64%)**
   The course theoretically reviewed cellular activities such as cellular transport, growth and differentiation, protein synthesis and cell signalling pathways. The course also covered the aspects of oncology and the role of various proteins/genes in the manifestation of cancer and other cellular disorders. The assessment was based on oral PowerPoint presentations where I presented on the **extrinsic pathways in apoptosis.** The course involved the review of numerous articles and unseen written examination.

7. Teaching and Practical (**Conflated Mark: 60%)**
   The assessment was based on rotational laboratory instruction to undergraduate students.

8. Embalming and Museum Techniques (**Conflated Mark: 66%)**
   The course reviewed the ancient to modern embalming and museum techniques. The assessment was based on two written examination and oral PowerPoint presentations where I presented on the **anatomical plastic models in scientific museum.**

9. Research Methodology (**Conflated Mark: 68%)**
   The course covered research methods in biomedical science. The techniques of collecting, processing, analysing and presenting scientific data. The assessment was based on written exam and study of some articles.

10. Advance Embryology and Teratology (**Conflated Mark: 67%)**
    The course reviewed the embryological development of the human body. It also introduced the current trends in teratology and developmental research protocols and methodologies. The assessment was based on written exam and study of some articles.

11. Advance Histology (**Conflated Mark: 61%)**
    The course reviewed the basic and systemic histology in addition to histological techniques, it also reintroduced advance microscopy such as electron microscopy, confocal microscopy and so on. It involved laboratory practical, PowerPoint oral presentation where I presented on the **histology of the urinary system.** The assessment was based on written examination.

12. Histochemistry and Techniques (**Conflated Mark: 53%)**
    The course covered the histochemical techniques employed in the histochemistry, enzyme histochemsitry, immunocytochemistry and in situ hybridisation techniques. The assessment was based on written examination.

13. Neuroanatomy and Neurotracing Technique (**Conflated Mark: 50%)**
    The course reviewed the anatomy of the nervous system and functional correlations. It also introduced theoretically the techniques employed in neuronal tracing such as the use of Pressure injection, Iontophoretic injection, Insertion of dye crystals and Viral transneuronal injection. PowerPoint oral presentation was conducted where I presented on the **neuroanatomy of the diencephalon.** Written examination was also conducted.

14. MSc Human Anatomy research project – (**Orally Defended May, 2016)**
    Assessment based on laboratory work, approximately 25, 000 words project write-up, 2 project power point oral defence (Internal and External Defence)

ii. Ahmadu Bello University, Zaria, Nigeria. BSc. (Hons.) Human Anatomy. **Grade: 2:2 division**. Year of study: 2004-2008.

All the assessments were based on coursework and unseen examinations.

1. In the Four year programme, I offered several courses from the departments of Human Anatomy, Human physiology, Biochemistry, Pharmacology and Community Medicine. This involved taught lectures, laboratory rotations and seminar presentations.
2. The coursework in the fourth year also included a four (4) month laboratory-based research project, in which I got a distinction.

RELEVANT WORKING EXPERIENCE WITH DATES

- ✓ Teaching, Research and Training at Umaru Musa Yar'adua University Katsina. 2013 –to - date
- ✓ Organizing and Conducting Practical for undergraduate Students. 2016 – date
- ✓ Biology Classroom Teacher, Gobarau Academy, Katsina. 2012
- ✓ National Youth Service Corps, Kano State-Nigeria (as Integrated Science Classroom Teacher) 2009–2010

TEACHING EXPERIENCES WITH DATES

(a) List of Undergraduate courses Taught according to the sessions

2015/2016

- BIO 1201 (General Biology II)
- Bio 2211 (Genetics I)
- BIO 1113 (Experimental Biology I)
- BIO 1126 (Experimental Biology II)
- BIO 3328 (Embryology)
- BIO 3302 (General Cytology)
- ZOO 3202 (Animal Anatomy)

2016/2017

- BIO 2201 (Introductory Ecology)
- BIO 1113 (Experimental Biology I)
- BIO 1201 (General Biology II)
- BIO 3306 (Molecular Biology)
- BIO 3328 (Embryology)
- BIO 3302 (General Cytology)

2017/2018

- BIO 1113 (Experimental Biology I)
- BIO 1126 (Experimental Biology II)
- BIO 2201 (Introductory Ecology)
- BIO 2211 (Genetics I)
- BIO 3328 (Embryology)
- ZOO 3202 (Animal Anatomy)
- BIO 3302 (General Cytology)

(b) List of Postgraduate courses Taught according to the sessions

Nil

(c) Undergraduate Supervision

2015/2016 ACADEMIC SESSION

1- Prevalence of Intestinal helminths among primary school pupils in Zango Local Government Area, Katsina State. Nuradden Sabi'u- U2/13/BIO/0072
2- Prevalence of Hepatitis B virus among blood donors attending General Hospital Katsina. Faruq Abubalar- U2/13/BIO/674
3- Prevalence of Hepatitis B virus among pregnant women attending Federal Medical Center, Katsina. Junaidu Aisha Galadanchi- U2/13/BIO/485

2016/2017 ACADEMIC SESSION

1- Identification and sexual characterisation of Drosophila *melanogaster spp* in Katsina Metropolis. Muhammad Musa Ahmad- U1/13/BIO/0584
2- Identification and taxonomic classification of different Drosophila strains in Katsina Metropolis. Buhari Bara'u- U1/13/BIO/1420
3- Study on the life cycle of Drosophila *melanogaster spp* in Katsina Metropolis. Aisha Ibrahim U1/13/BIO/1652

2017/2018 ACADEMIC SESSION

1- Effects of Deforestation on Fauna and Flora and how it affects biodiversity and food web in Kaita, Katsina State.AMIRA UMAR ISYAKU- U1/14/BIO/0426
2- Effects of Deforestation on Fauna and Flora and how it affects biodiversity and food web in Batsari, Katsina State. U1/14/BIO/2352- Usman Muhammad Z
3- Assessment of the Impact of Malnutrition on Children at General Hospital Katsina. AMINA ISYAKU- U1/14/BIO/2331
4- Prevalence of Gastrointestinal Parasite in Sheep within Katsina Metropolis. AHMED NASIR U1/14/BIO/0444
5- Prophylactic And Antibacterial Properties of *Moringa Oliefera* Leaves (Radish/Zogali). MUHAMMAD DAYYABU. U2/15/BIO/1356

Secondary Education

1. Junior School Certificate from Government Science and Technical School, Batagarawa, Katsina State. Year of Study: 1997-1999
2. Senior School Certificate from Government College Katsina, Katsina State, Nigeria. Year of study: 2000-2003.

Primary Education

Primary School Certificate from Modoji Primary school, Katsina State, Nigeria. Year of study: 1992-1997.

<u>ADDITIONAL CERTIFICATES</u>

- ✓ *The Article Publishing Process: An Elsevier Author Workshop: (120 minutes)*
- ✓ *The Book Publishing Process: An Elsevier Author Workshop: (120 minutes)*
- ✓ *Programming for Everybody (Getting Started with Python) an online non-credit course authorized by* ***University of Michigan*** *and offered through Coursera*
- ✓ *The Data Scientist's Toolbox an online non-credit course authorized by* ***Johns Hopkins University*** *and offered through Coursera*
- ✓ *Basic Gene Expression Techniques at Center for Biotechnology Bayero University, Kano, Nigeria*
- ✓ *Basic Molecular Techniques Workshop at Inqaba Biotec West Africa*
- ✓ *Basic Bioinformatics Workshop at Inqaba Biotec West Africa*

B. Research experience

1. Recent research activity

PhD Research topic:

"Matrilineal Genetic Diversity and Forensic Characterization of Hausa Population in Nigeria."

Undergraduate Students Supervised in their chosen research topics

i. I supervised three undergraduate students in their chosen research area in the 2015-2016 academic session.
ii. supervision of three undergraduate students, we want to pioneer drosophila research in my institution, we are starting with ecology, life cycle and biology of drosophila population in the area (Katsina, Nigeria). 2017

MSc. Project

I undertook my MSc. research under the supervision and mentorship of Professor Samuel Sunday Adebisi and Professor Wilson Oliver Hamman, Department of Human Anatomy, Ahmadu Bello University, Zaria, Nigeria. The work was aimed at evaluating the intrauterine effects of ethanol exposure on the histology, histochemistry, neuro-behaviour and neurochemistry of the cerebral and cerebellar cortices of neonatal Wistar rats at different phases of development. The neurochemistry analysis was done using Atomic Absorption Spectrophotometry to quantify the amounts of Iron, Copper, Zinc and Manganese in the neonatal rat's Cerebrum and cerebellum. The result showed effects on the histology, morphology, neuro-behaviour and neurochemistry of the affected neonates.

2. Previous laboratory experience

My undergraduate research work studied the effects of intrauterine ethanol exposure on histology of foetal Wistar rats. [Paper published (1)]

INTERNATIONAL CONFERENCE ATTENDED WITH DATES

1. Ibrahim El-ladan Shehu, Arvind Jain Kumar and Gaurav Kumar. PhD thesis presentation: Matrilineal Genetic Diversity and Forensic Characterisation of Hausa Population in Nigeria. 2nd International Conference on Neo Era of Forensic Science and Law Interface. 22nd – 23rd April, 2023. Forensis Agora 2023. Galgotias University.
2. Ibrahim El-ladan Shehu and Priyanka Chhabra. Review on African Y-STR haplotyping and Y Chromosome Profiling. International e-Conference on Forensic Science and Criminology: Bridging the Gap in Criminal Justice System Conference Series; Forensis Agora. 15th -16th May, 2021. Galgotias University

PUBLICATION(S) IN PEER REVIEW JOURNALS

1. **Ibrahim El-ladan Shehu** and Priyanka Chhabra. African Y-STR haplotyping and Y chromosome profiling: A Review. Indian Journal of Forensic Medicine and Pathology. 2021;14(3 special issue):379-385. Scopus-indexed
2. **El-ladan Ibrahim Shehu** and Affan Usman (2021). The effects of intrauterine ethanol exposure on the levels of Iron and Copper in Cerebrum and Cerebellum of neonatal Wistar rats. *Dutse Journal of Pure and Applied Sciences*.
3. **El-ladan I.S.** and Usman A. (2021). Effects of intrauterine ethanol ingestion on zinc metabolism in cerebrum and cerebellum of neonatal wistar rats. *Katsina Journal of Pure and Applied Sciences.*
4. Sadeeq AA, Bauchi ZM, Tanko M, Timbual JA, Mukhtar AI, Fatika AR, Abdullahi A, Muhammad Z, **El-ladan IS**, and Rilwan BH. (2021). Mercury exposure and post exposure recovery of the cerebellar cortex and hippocampus of adult wistar rats. *Journal of Anatomical Sciences*.
5. Sadeeq AA, Amos A, Bauchi ZM, Tanko M, Mukhtar AI, Muhammad Z, **El-ladan IS**, and Rilwan BH. (2021). Solanum esculentum ameliorates paraquat-induced cerebellar and cervical spinal lesions in wistar rats. *Nigerian Biomedical Science Journal*.

6. Sadeeq AA, Bauchi ZM, Tanko M, Hussaini IA, Gilimah MH, Muhammad Z, **El-ladan IS**, and Rilwan BH. (2021). Ameliorative effects of ethanolic extract of turmeric on memory and motor activities in mercury-induced neurotoxicity in rats. *Dutse Journal of Pure and Applied Sciences.*
7. **El-ladan I.S** and AA Sadeeq (2020). Histopathological Effects of Repeated In-Utero Ethanol Exposures on Neonatal Cerebral Cortices in Wistar Rats. Paper ID: BJBS/20/0119
8. AA Sadeeq, A.O Ibegbu, SP Akpulu , Hadiza A.R, L.H Adamu, **El-ladan I.S**, M.O Kwanashie (2017). Evaluation on the effects of Mercury exposure on exploratory motor activity of Adult Wistar Rats. *Nigerian Biomedical Science Journal*, Vol. 13, No. 1,. Pp 18-22
9. AA Sadeeq, A.O Ibegbu, JA Timbuak, M Tanko SP Akpulu , HR Bello, SA Musa, L.H Adamu, **El-ladan I.S**, H.O Kwanashie (2016). Role of Mercury toxicity on memory and Calcium (Ca+) level determination using NAA-1 in the brain tissues of adult Wistar rats. *Nigerian Biomedical Science Journal*, Vol., No. ,
10. Sunday A Musa, **Shehu Ibrahim,** Uduak E Umana, Samuel S Adebisi and Wilson O Hamman, (2012). Pathological lesions in the lungs of neonatal Wistar rats from dams administered ethanol during gestation. *Asian Journal of Medical Sciences* 4(1): 4-7,

PAPERS PREPARED AND COMMUNICATED FROM MY THESIS

1. Matrilineal Genetic Diversity and Forensic Data of Hausa Ethnic Population of Daura Emirate, Nigeria. Journal of Indian Academy of Forensic Medicine: Scopus-indexed….Communicated
2. Population Structure and Stratification among the Hausa Ethnic Group based on HVR1 mtDNA Analysis. Journal of Genetics: Springer; Scopus-indexed….Communicated
3. Mitochondrial DNA Analysis Reveals Complex Genetic Diversity and Population Structure among the Hausa People of Nigeria: Forensic Sciences Research: Taylor and Francis; Scopus-indexed…Communicated

PAPERS PRESENTED AT CONFERENCES AND SEMINAR

1. Purkinje cell loss in neonatal wistar rats exposed to ethanol in utero affects the development of sensory and motor reflexes. The 15th Annual Scientific Conference And General Meeting of The Anatomical Society of Nigeria (ASN). Sept.2018 held at Usmanu Danfodio University Sokoto, Nigeria. - **Oral Presentation**
2. Prevalence of intestinal parasite among pupils in some selected public primary schools in Zango Local Government Area, Katsina State, Nigeria. 41st Annual General Meeting and Scientific Conference of the Nigerian Society of Microbiology Nigeria (NSM), Sept. 2018 held at Umaru Musa Yar'adua University, Katsina, Nigeria. - **Oral Presentation**
3. Neurodevelopmental Consequences of Prenatal Alcohol Exposure on the Cerebellar Cortices of Neonatal Wistar Rats. The 16th Annual Scientific Conference And General Meeting of The Neuroscience Society of Nigeria (NSN). Aug.2018 held at Federal University Dutse, Jigawa, Nigeria. - **Oral Presentation**
4. Altered Iron and Copper cerebellar metabolism in neonatal wistar rats exposed to ethanol in-utero affects the Development of Sensory and Motor Reflexes. The 16th Annual Scientific Conference And General Meeting of The Neuroscience Society of Nigeria (NSN). Aug.2018 held at Federal University Dutse, Jigawa, Nigeria. - **Oral Presentation**
5. Zinc Deficiency in the Neonatal Wistar Rats' Cerebral Cortex Contributes to Intrauterine Ethanol Foeto-Toxicity. The 16th Annual Scientific Conference And General Meeting of The Neuroscience Society of Nigeria (NSN). Aug.2018 held at Federal University Dutse, Jigawa, Nigeria. - **Poster Presentation**
6. Effects of Intrauterine Ethanol Exposure on the Development of Sensory and Motor Reflexes in Neonatal Wistar Rats. 14th Annual Conference and General Meeting of the Neuroscience Society of Nigeria (NSN) 2016. Zaria, Nigeria- **Oral Presentation**
7. Differential Pattern of Intrauterine Ethanol Exposure Effects on Cerebral Cortex of Neonatal Wistar Rats. 14th Annual Conference and General Meeting of the Neuroscience Society of Nigeria (NSN) 2016. Zaria, Nigeria- **Oral Presentation**

C. Grants

1. Nigerian Tertiary Institution Trust Fund (TETFUND): MSc Research Grant.

D. Memberships

1. Member of Ahmadu Bello University Zaria, Human Anatomy Students Association 2004-2008.
2. Member Anatomical Society of Nigeria
3. Member Neuroscience Society of Nigeria
4. Member Backbone Organisation

E. Community-directed Activities

- Community outreach Services under Backbone Organisation 2013 to date
- Member, Local Organising Committee, Communique subcommittee. Nigerian Society for Microbiologist, Scientific Conference and Annual General Meeting. UMYU, Katsina, Nigeria.
- Member, Faculty of Natural and Applied science Scientific Research Group
- Member, University ICT Committee
- Member, Faculty of Medicine Research evaluation committee
- Member, Faculty of Medicine Promotion Assessment Committee

F. Skills

IT Skills: SPSS 20; conversant with basic software and programs.: Learning R and Python Programming, STRUCTURE, BioEdit, PopArt, MEGA11, Mesquite, TCS, SDT, Arlequin

Technical Skills: Rat Experiments; determination of Oestrous Phases in Wistar Rats, Histological and Histochemical Techniques, Basic Gene Expression Techniques, Molecular Genetic Techiques and Bioinformatics

G. Hobbies

Research enthusiast, Travelling, Reading and Discussing and Analyzing Public Policies

H. Referees

1. Dr. Gaurav Kumar, Division of Clinical Research, Department of Biosciences, School of Basic and Applied Sciences, Galgotias University, India
2. Professor Arvind Kumar Jain, Division of Forensic Sciences, Department of Chemistry, School of Basic and Applied Sciences, Galgotias University, India
3. Dr Bashir Abdulkadir, Department of Microbiology, Umar Musa Yar'adua University, Katsina, Nigeria Mobile: +2348065137374, email. bashir.abdulkadir@umyu.edu.ng