Forensic Science International: Synergy 11 (2025) 100633

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

Forensic biobank; towards comprehensive forensic genetic frequency database for the Kenyan population

Eva K. Aluvaala Nambati [a,*], Edward Muge [b], Belinda Azzam [a], Diana Maritim [a], Ngure Kirosh [a], Luna Kamau [a], Solomon Mpoke [a], Winfrida Cheriro [c], Lewis Karani [a], Martin Sang [a], Abdiaziz Gosar [a], Olipher Makwaga [a], Lydia Eyase [a], Herzel Tiffany Wandera [a], Sharon Ariga [a], James Nyabera [a], Milka Mwangi [a], Francis Kimani [a], Wallace Bulimo [a]

[a] Kenya Medical Research Institute, P.O. Box 54840 00200 off Raila Odinga Way, Nairobi, Kenya
[b] University of Nairobi, Kenya. P.O. Box 30197-00100, Nairobi, Kenya
[c] Moi University, Kenya. P.O. Box 3900 - 30100, Eldoret, Kenya

ARTICLE INFO

Keywords:
Forensic biobank
Comprehensive
Forensic genetic frequency databases
Kenya

ABSTRACT

Forensic genetic frequency databases (FGFD) are used in estimating the probability of a DNA match in forensic investigations. They provide reference population data that can be used for statistical estimation for the rarity of a genotype, haplotype or a DNA profile in a population hence giving probative value for forensic evidence.

Currently, three FGFD databases are recommended by the International Society for Forensic Genetics (ISFG) for forensic use; the Y-Chromosome Haplotype Reference Database (YHRD), the EDNAP Mitochondrial DNA Population Database (EMPOP), and the STRs for Identity ENFSI Reference Database (STRidER).

There is need to generate updated and comprehensive genetic frequency data for the Kenyan population in compliance with ethical standards.

This study sought to develop a forensic biobank to facilitate generation of comprehensive genetic frequency data for the Kenyan population. A total of 893 samples were collected from study volunteers in compliance with prescribed ethical standards. The data set comprises 60.8 % Bantu, 24.9 % Nilotic, and 14.3 % Cushitic samples closely mirroring current population structure in Kenya. The samples are currently stored in duplicate as FTA cards and extracted DNA.

132 quality mitogenome reference data has been generated for the coastal region in Kenya. With the broad consent obtained, the resource will be used to generate additional mitogenome reference data for other geographical regions, Y-chromosome haplotype and autosomal STRs for inclusion in recommended forensic databases as per revised guidelines. With the emergence of new technologies in forensic genetics, we anticipate the resource will be valuable in forensic genetics validation studies in future.

1. Introduction

1.1. Forensic genetic frequency databases

Forensic genetic frequency databases (FGFDs) are a collection of allele frequencies from autosomal short tandem repeats (STRs), Y-chromosome STRs (Y-STRs), and mitochondrial DNA haplogroups[1–3]. Unlike other forensic DNA databases such as the offender and crime scene index, FGFDs do not have contributor information and are derived from random individuals in a population. Their sole purpose is to provide probative value to DNA evidence [4]. When there is a match between a suspect's DNA and the crime scene DNA profile, these databases are used to calculate the random match probability (RMP) [5]; RMP is the odds of someone else in the population having the same DNA profile. Meaningful RMP and accurate statistical inferences can only be made when the appropriate population specific forensic genetic frequency databases are used [6].

Currently three high-quality FGFD databases have been endorsed by the International Society for Forensic Genetics (ISFG). They include the Y-chromosome Haplotype Reference Database (YHRD), the EDNAP

* Corresponding author.
E-mail addresses: ealuvaala@kemri.go.ke, ealuvaala@gmail.com (E.K. Aluvaala Nambati).

https://doi.org/10.1016/j.fsisyn.2025.100633
Received 4 June 2025; Received in revised form 17 July 2025; Accepted 22 July 2025
Available online 29 July 2025
2589-871X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Mitochondrial DNA Population Database (EMPOP) and the STRs for Identity ENFSI Reference Database (STRidER) [7].

The amount of data in forensic genetic frequency databases is still limited and there is inequity in data representation from diverse world populations. Representation is skewed towards European and Asian populations. African data is underrepresented accounting for 5.7 % of data in EMPOP, https://empop.online/empop_stats, 8.9 % in STRidER https://strider.online/frequencies and 12.3 % in YHRD, https://yhrd.org/pages/resources/stats.

Recently the ISFG revised recommendations for FGFDs in line with ethical standards. Researchers and practitioners intending to submit population data must meet the pre-requisite requirements for ethical standards. In yesteryears, genetic frequency databases were operational without strict observance of these international standards. However, with the evolving ethical standards ecosystem, compliance with the set guidelines is now mandatory, rendering some old datasets unsuitable for inclusion in present-day forensic frequency databases [7,8]. This has further compounded the issue of limited data. Furthermore, technologies have changed over time and new markers included, making it necessary to update these databases. A good example would be the EMPOP database. Before validation of next generation sequencing (NGS) in forensic purposes, mitochondrial DNA analysis was focused on the hypervariable regions (HVR) hence the data in EMPOP was primarily for this region. However, recently entire mitogenomes are being used in forensic investigations and hence the need to update the database with whole mitogenome data [9].

1.2. Forensic biobanks

Biobanking for medical genetics and genomic research is well established. However, there are few biobanks that have specifically been developed for forensic genetics and genomic purposes [10]. Whilst medical biobanks are meant for deciphering genetic basis of disease and development of medical interventions, forensic biobanks are focused on genetic markers that are useful in human identification and aiding forensic investigations [10,11]. Depending on the consent obtained, legislation and ethical guidelines, samples from a medical biobank may be used for forensic purposes. Particularly, they can be used to generate forensic genetic frequency data as the resulting database doesn't include personal data and therefore cannot be traced to specific individuals [12]. However, collaborative research between forensic geneticists and medical geneticists seldom occurs and sharing samples may not always be feasible. This necessitates the need to have biobanks that are specific for forensic purposes, particularly for countries that are yet to generate forensic genetic frequency data [13,14]. Forensic genetics involves the use of human subject samples and hence the need to comply with ethical guidelines [15,16]. Recently the ISFG made compliance with these guidelines a mandatory pre-requisite for submitting data to quality-controlled databases, such as STRider, YHRD and EMPOP [17]. Establishment of forensic biobanks in countries yet to develop population frequency databases may be a viable option for addressing the current shortage of data in FGFDs. Furthermore, even for populations that have already generated adequate frequency data, a forensic biobank can be useful for validating new markers and technologies [18].

2. Material and methods

Ethical approval

The study was approved by the KEMRI Scientific and Ethics Review Unit (SERU), under protocol number KEMRI/SERU/CBRD/PROP216/044/4148. A research permit was obtained from the National Commission for Science, Technology & Innovation (NACOSTI), license no NACOSTI/P/22/19201.

2.1. Community engagement

The county government's health department was the entry point and engagements were through the County Executive Committee (CEC) that is responsible for implementing policies and programs within the county government. We were required to provide the county committees with approved research protocols, Institutional Review Board IRB) approvals and research permits from NACOSTI and pitch the research concept to the committees or an appointed official to explain the essence of the research study. Upon satisfaction, the CEC would issue an authorization letter to conduct research in the county. Through the CEC, community health volunteers and chiefs were assigned to us to facilitate navigation of the county. The community health volunteers helped with creating awareness on the project and also assisted in the identification of native and indigenous communities in the county.

2.2. Informed consent

The study participants were presented with informed consent primarily in Kiswahili (Kenya's national language) and English (Kenya's official language). However, where required translation was done in the native language to ensure that the participants understood and were able to give informed consent. During the consenting process, the risks, benefits and rights of the participant were explained. The participation was voluntary, and participants had the right to withdraw participation at any time. Monetary benefits were not given to participants. However, a transport compensation fee was given to participants who had to travel to the sample collection point. Each participant was given a copy of the signed informed consent form. Broad consent for use of samples for future studies was also sought.

2.3. Privacy and confidentiality

To protect the privacy of study participants, all the study samples were anonymized. Personal information was de-linked from the sample. To ensure restricted access to personal data, hard copies of the signed consent form containing personal information were kept filled in cabinets under lock and key.

2.4. Study design and study sites

The study design was a population-based cohort study. Samples were randomly collected from consenting individuals from the 47 recognized ethnolinguistic groups from different geographical locations in Kenya. Participant recruitment was based on self-identified ethnicity. Samples were collected from diverse populations from all the geographical regions including coastal, western, Nyanza, Central, Eastern and Northern region.

The coastal region samples were collected from the five counties: Kwale, Kilifi, Taita Taveta, Tana River and Lamu. Samples were collected from 18 ethnolinguistic groups present at the Kenyan coast which included the Pokomo, Malakote, Boni, Sambaa, Makonde, Pare, Shirazi, Segeju, Pemba, Mijikenda, Tswaka, Vumba, Wakifundi, Zegeu, Taita, Taveta, Swahili and Bajuni. In the western region samples were collected from the Luhya (Bakhayo, Banyala, Isukha, Banyore, Bukusu, Idakho, Kabras, Kisa, Marachi, Maragoli, Marama, Samia, Tiriki, Tsotso, Wanga) and Teso. In the Nyanza region samples were collected from Luo, Suba, Gusii and Kuria. In the Rift Valley region samples were collected from Kalenjin's (Keiyo, Kipsigis, Lembus, Marakwet, Ogiek, Pokot, Sabaot, Terik, Tugen). In the Central region samples were collected from the Kikuyu, Meru and Embu. In the Eastern Region samples were collected from the Kamba. In the Northern Region samples were collected from the Somali (Ajuran, Degodia, Gurreh, Hawiyah, Murile, Ogaden, Leysan) El Molo, Borana, Burji, Gabbra, Sakuye, Boni, Waata/Waliangulu, Dahalo, Galla and Rendille.

2.5. Sample size determination

A sample size of 20–25 was targeted per ethnolinguistic group. This sample size was calculated based on recommendations on population genetics and genomics studies [19,20]. Based on this, a total sample size of between 940 and 1175 was targeted. For ethnic groups with sub-tribes, 10 individuals per sub-tribe were targeted.

2.6. Sample collection

Samples were collected from consenting individuals who met the inclusion criteria; male or female participant, adults or minors who are maternally unrelated and belong to one of the 47 recognized Kenyan ethnic communities. Sterile Forensic grade COPAN's 4N6FLOQ swabs (Copan Diagnostics, CA, USA) were used for the collection of epithelial cells from the oral cavity of the study volunteers. The cells on the swab were transferred on indicating COPAN^{TM} nucleic card for long term storage and the swab was used for DNA extraction.

2.7. Sample purification

The total DNA Genomic DNA (nuclear and Mitochondrial) was extracted using Promega casework extraction kit and the DNA subsequently purified using the DNA IQ^{TM} casework prokit for Maxwell ®16 following manufacturer's instructions.

Briefly, each swab was placed at the bottom of a spin basket placed in a labelled ClickFit Microtube and extraction mix consisting of 286 µl Casework Extraction Buffer, 10 µl Proteinase K (18 mg/ml), and 4 µl Thioglycerol added to each sample. The samples were vortexed vigorously for 5 s and allowed to incubate at 56 °C for 30 min for cell lysis. The lysed samples were then centrifuged at room temperature for 2 min at 20,800×*g*. The spin basket containing the swab was discarded and an additional 200 µl of lysis buffer added to the lysate. The lysate was loaded on the Maxwell® 16 cartridge and loaded on the Maxwell® 16 automated extractor for purification using a preset program that includes lysis, wash steps and DNA elution. The eluted DNA was stored at -80^0C.

2.8. Sample catalogue and biobanking

The extracted DNA samples were then stored at KEMRI's Sample Management and Reference Facility (SMRF). This a physical biobank that uses an inventory management system for efficient and secure handling of samples. This system includes sample barcoding and electronic tracking, enabling accurate identification, cataloging, and retrieval of each sample stored within the facility. The inventory system is fully integrated with SMRF's freezer units, which allows real-time monitoring of sample locations, temperature conditions, and usage status, ensuring that all samples remain under optimal conditions for future research. In addition, duplicate samples were stored at room temperature on FTA cards.

The sociodemographic, geographic and ancestry data gathered during sample collection was transferred into the Human Genome Biorepository within the KEMRI Knowledge base, Data in this repository is linked to barcoded sample identifiers, within KEMRI's SMRF, where the physical samples are securely stored. The biorepository has an intuitive user interface that supports domain restricted data input and also a change log for every access and modifications made.

The biorepository integrates Kenya's Privacy and Data Protection ACT of 2019 [21], adhering to secure data handling via practices such as end-to-end encryption, role-based access, document version control and a multi-tier authentication system to safeguard data. Access is strictly regulated at different levels; users cannot download data directly but may submit requests for data access.

2.9. Mitogenome reference data

Using the extracted DNA, 132 mitogenomes from 5 coastal counties; Kwale, Kilifi, Taita Taveta, Tana River and Lamu were sequenced using NGS approach on the Illumina Miseq system [22]. Long-range PCR, spanning the entire human mitochondrial genome (16,569 bp), was performed using two primer sets to generate overlapping targets of ~8.5 kb. Library preparation was performed using the Colibri™ ES DNA Library prep kit for Illumina Systems (Thermo Fisher Scientific, USA). Fragments between 350 and 500 bp were targeted.

The libraries were pooled and normalized to a concentration of 4 nM. The normalized libraries were denatured and diluted to 8 pM and 1 % PhiX added as per the Illumina Miseq Denature and Dilute Libraries Guide (MiSeq System Denature and Dilute Libraries Guide (15039740). To generate paired-end sequence data, the denatured libraries were loaded on the v3 Miseq 600cycle cartridge and sequenced on the Illumina Miseq System (Illumina, Inc. San Diego, California).

Data analysis and haplogroup assignment were performed using the CLC genomics workbench v.22 the AFDIL-QIAGEN mtDNA Expert (AQME) plugin.

3. Results

3.1. Ethnolinguistic and geographical distribution

The participants were broadly classified as Bantu, Cushitic, and Nilotic. A total of 893 samples are currently biobanked with the following distribution. The total number of Bantus were 543 (60.8 %), Nilotes 222 (24.9 %) and Cushites 128 (14.3 %) (Fig. 1). The Bantus were further classified as Coastal, Western, Central and Eastern based on geographical regions. The Coastal Bantus had the highest representation at 286 (32 %) followed by western Bantus at 171 (9.63 %) due to presence of sub-ethnic groups. Central and Eastern Bantus had a lower representation at 86 (19.1 %), due to absence of sub-ethnicities. The highland and plain Nilotes predominantly found in the Rift valley region were represented at 84 (9.41 %), 79 (8.85 %) respectively, whilst the river lake Nilotes present in Nyanza and some parts of western Kenya were represented at 59 (6.61 %.) For the Cushitic group, largely found in the North Eastern region, the Somali had the highest representation at 60 (6.8 %) owing to the sub-ethnic groups, while the rest of the Cushitic ethnic groups cumulatively accounted for 68 (7.7 %).

3.2. Socio demographic characteristics

The sociodemographic data captured in the study included age, educational level, employment status, marital status, and gender. The mean age of participants was 41.9 years (Table 1).

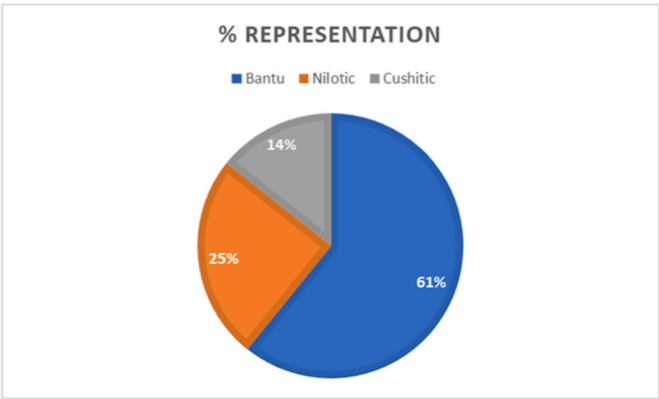

Fig. 1. Pie chart showing the percentage representation of the major ethnic groups in Kenya.

Table 1
Sociodemographic characteristics (Age, Education Level and Employment Status) captured in the study.

No	Sociodemographic Characteristic		Value						
1	**Age Distribution**								
		Mean Age	41.9						
		Median Age	40						
		SD Age	15.4165						
		Min Age	17						
		Max Age	93						
2	**Education Level**		**College**	**Other**	**Primary**	**Secondary**	**University**	**N/A**	**Total**
		Count	163	2	304	236	78	110	893
		Percent %	18.25	0.22	34.04	26.43	8.73	12.32	100
3	**Employment Status**		**Formal**	**Informal**	**Self-employed**	**Student**	**Unemployed**	**NA**	**Total**
		Count	191	182	279	40	184	17	893
		Percent %	21.4	20.4	31.2	4.5	20.6	1.9	100

In terms of education, 34 % attended primary education, followed by 26.4 % with secondary education, 18 % with college education, and 8.73 % with university education. Additionally, 12 % of participants were uneducated, and 0.2 % attended informal education (Table 1).

For employment status, (Table 1, most of the participants were self-employed at (31.2 %), with formally employed accounting for (21.4 %), informal employment was at (20.4 %), and unemployment, 20.6 %. Students accounted for 4.5 % and 1.9 % did not declare their employment status.

The gender distribution of participants was 43 % female and 57 % male (Fig. 2).

3.3. Mitogenome haplogroup diversity at the Kenyan coast

Diverse mitochondrial haplogroups were observed at the coastal region, this was predominantly of macro-haplogroup 'L'.

Diverse haplogroups were observed including L3, L0, L2, L1, L4, M, L5, and N. 66 individuals clustered as L3 haplogroup making it the most prevalent at 50 %. Twenty-eight individuals clustered as L0 represent 21.2 % of haplogroups in this geographical region. Twenty-two individuals clustered as L2 account for 16.6 %. Haplogroup L1 and L4 each had 6 individuals accounting for 4.5 %. Haplogroup M was observed in 2 individuals accounting for 1.5 % of all haplogroups Fig. 3. Fifty-seven unique haplogroups were observed. L3 had 23 unique sub haplogroups with L3e3a having the highest frequency of 13.5 %. L0a2a2a was the most frequent L0 sub haplogroup with a frequency at 5.3 %.

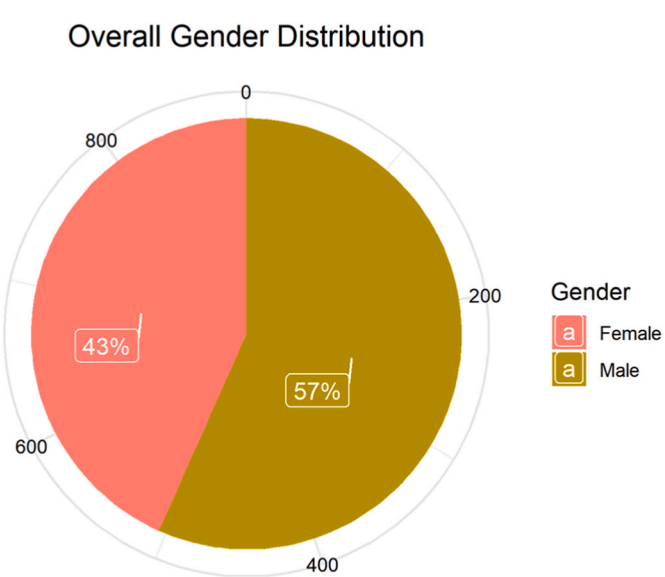

Fig. 2. Pie chart showing the percent gender distribution for data set present in the biobank.

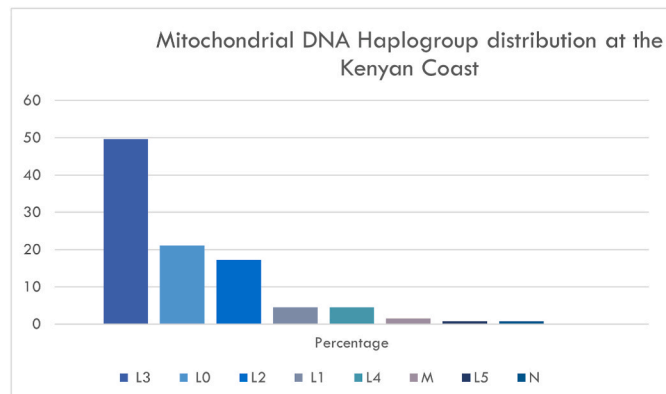

Fig. 3. Percentage Haplogroup distribution as observed at the Kenya coast.

L2a1b1a had the highest frequency within L2 haplogroup at 5.3 %. The remaining sub-haplogroups within the L1, L4, M,L5 and N had low frequencies of between 0.75 and 1.5 % (Table 2).

The distribution of haplogroups and sub-haplogroups was heterogenous with clustering observed in some ethnolinguistic groups.

4. Discussion

The majority representation of Bantus at (60.8 %), followed by Nilotes (24.9 %) and Cushites (14.3 %), reflects on the general population distribution in Kenya hence equitable representation of the ethnolinguistic groups was achieved. With 87.7 % of the population having gone through formal education, the literacy levels of the participants is indicative of their ability to give informed consent. A mean age of 41.9 years reveals a largely middle-aged cohort that did not require a guardian for consenting purposes. 73 % of the participants were involved in an economic activity, this is significant especially for vulnerable groups who may be coerced to participate in a study for economic gain. In this study, the participants were not offered any monetary compensation in exchange for samples. In our experience community engagement and informed consent enhanced participant willingness to volunteer in the study without financial incentives. A near-equal distribution of male (57 %) and female (43 %) participants demonstrates that there was gender inclusivity.

High quality mitogenome sequence data was achieved from this study [22]. Analysis of 132 samples from the Kenya coast, revealed haplogroups predominantly belonging to the macro-haplogroup L, which is characteristic of individuals of African ancestry [23]. Haplogroup L3 was the most prevalent, consistent with previous studies highlighting its prominence in African populations. In a previous study, haplogroup L3 was also found to be the most prevalent in the Nairobi database [24,25] In addition, the L0 haplogroup considered sub-Saharan

Table 2
Unique mitogenome haplogroups at the Kenyan coast and respective frequencies n = 132.

No	Hap L3	L3 Freq	Hap L0	L0 Freq	Hap L2	L2 Freq	Hap L1	L1 Freq	Hap 4	L4 Freq	Hap M	M Freq	Hap 5	L5 Freq	Hap N	N Freq
1	L3e2b	0.0150	L0f1	0.0225	L2a1+143+16189+16192	0.0150	L1c1a1	0.0075	L4b2a1	0.0150	M1a1f	0.0075	L5b2	0.0075	N1a1a	0.0075
2	L3d1a1a	0.0526	L0d2a'b'd	0.0075	L2a1b1a	0.0526	L1b1a	0.0150	L4b2a2	0.0150	M1a1d	0.0075				
3	L3e4a	0.0225	L0f2a	0.0375	L2a1	0.0075	L1c2b1a'b	0.0075	L4b2a	0.0075						
4	L3a2	0.0150	L0a2a2a	0.0526	L2a1g	0.0075	L1c2a3	0.0075	L4b2a2b	0.0075						
5	L3e3a	0.1353	L0a1b1a1	0.0150	L2a1h	0.0451	L1b2a	0.0075								
6	L3e1b2	0.0150	L0a1+16293	0.0075	L2b2a	0.0075										
7	L3b1a1a	0.0225	L0a1e	0.0150	L2b2	0.0075										
8	L3f1b4a1	0.0375	L0a1d	0.0075	L2a1a1	0.0225										
9	L3d1a1a1	0.0451	L0a1b1a1a	0.0075												
10	L3e1d	0.0225	L0a1b2a	0.0075												
11	L3x1a	0.0075	L0a2d	0.0150												
12	L3h1a2a1	0.0150	L0a2a1a	0.0075												
13	L3h1b2	0.0075	L0f	0.0075												
14	L3a1b	0.0075														
15	L3e5a	0.0075														
16	L3f	0.0075														
17	L3e1a3a	0.0075														
18	L3e1a2	0.0075														
19	L3x1a2	0.0075														
20	L3i2*1	0.0150														
21	L3e1b2	0.0075														
22	L3b1a11	0.0075														
23	L3d1d	0.0075														

*Highlighted in black are the sub-haplogroups with the highest frequencies at the Kenyan Coast

African specific was found in relatively high frequency at the coastal region in Kenya [25]. The L3e sub-haplogroup was found to have the highest frequency, suggesting a significant maternal lineage contribution to the genetic makeup of individuals residing along the Kenyan coast.

The distribution of haplogroups exhibited regional variations; whereas L0 prevalence was highest in Tana River and Lamu county. L3 was prominent in Kwale, Taita Taveta and Kilifi County.

The study also elucidated the relationship between mitochondrial haplogroups and ethnolinguistic affiliations. For instance, individuals identifying as Pokomo exhibited a higher prevalence of haplogroup L3 in Tana River County, while the Swahili population in Lamu County displayed a distinct distribution with haplogroup L0 being predominant. The Taita, unlike the other ethnic groups in Taita Taveta, displayed high prevalence of haplogroup L0. Interestingly while L3e was the dominant sub-haplogroup when the entire coastal region is considered, Tavetas are the only ones who typed as L3h and all the Bajunis clustered as L3d. While L0a was the most prevalent L0 sub-haplogroup, and present in all the counties, L0f was primarily observed in Tana River and Taita Taveta county with only 1 observation made in Kilifi. L0f haplogroup has previously been reported as being present in Tanzania but rare or absent in other regions in Africa [26]. Taita Taveta borders Tanzania, and this may explain the presence of this rare haplogroup in this county. L0f apart from being observed among the Taita was also observed in individuals who self-identified as Pare who are indigenous to Tanzania. In addition, haplogroups N1 and M1 are rare in African populations but have been found present in Tanzania [26]. In the Taita Taveta county dataset, this haplogroups were found in individuals that self-identified as Sambaa and Chagaa who are also indigenous to Tanzania.

5. Conclusion

The established forensic biobank has equitable representation of diverse ethnolinguistic groups from different geographical regions in Kenya. This is expected to provide comprehensive datasets that capture genetic diversity in the Kenyan population that can be used in generating forensic genetic frequency data. There has been previous effort to generate frequency data for autosomal STR, YSTR and mitochondrial DNA for the Kenyan population however the studies did not capture the entire ethnic diversity in Kenya [27]. Furthermore, some of the data was generated using fewer markers hence the need to update for the new markers [28]. In addition, there have been technological advancements including NGS that has eased whole mitogenome sequencing Prior to this, mitochondrial DNA data in EMPOP was focused on the HVR data, there is need to update the Kenyan data [24].The biobank is positioned to provide research samples that can be used in generating inclusive population specific forensic genetic frequency reference database for the Kenyan population. Whereas we have focused on the coastal region haplogroup diversity, it's important to note that the forensic biobank contains samples from other Kenyan regions that will be used to generate a comprehensive mtDNA database as well as Y and autosomal STRs reference databases. The biobank can in addition be used for validation studies when new technologies and forensic kits become available. The current size of the biobank can be increased to have larger data sets. Larger data sets can in future be used for fine scale genetic studies that can potentially contribute to the development of predictive models for prediction of biogeographical ancestry in Kenya.

CRediT authorship contribution statement

Eva K. Aluvaala Nambati: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Edward Muge:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Belinda Azzam:** Writing – original draft, Methodology. **Diana Maritim:** Writing – review & editing,

Writing – original draft, Methodology. **Ngure Kirosh:** Writing – review & editing, Writing – original draft, Methodology. **Luna Kamau:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Solomon Mpoke:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Winfrida Cheriro:** Writing – review & editing, Methodology. **Lewis Karani:** Writing – original draft, Visualization. **Martin Sang:** Methodology. **Abdiaziz Gosar:** Methodology. **Olipher Makwaga:** Methodology. **Lydia Eyase:** Methodology. **Herzel Tiffany Wandera:** Writing – original draft, Methodology. **Sharon Ariga:** Writing – original draft, Methodology. **James Nyabera:** Methodology. **Milka Mwangi:** Writing – original draft, Methodology. **Francis Kimani:** Writing – original draft, Methodology. **Wallace Bulimo:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Ethics statement

All the study participants gave their informed consent for inclusion before they participated in the study. The protocol was approved by the KEMRI Scientific and Ethics Review Unit (SERU) KEMRI/SERU/CBRD/PROP216/044/4148 and a research license was issued by the National Commission for Science and Technology and Innovation Kenya. All applicable guidelines governing the ethical conduct of research involving human subjects were adhered to. Personal data and metadata of participants is confidential and is not included in analysis and publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Kenya Medical Research Institute grant number: KEMRI/IRG/EC002. The authors would like to thank the county governments of Garissa, Coast region and Nairobi for their support in the sensitization and mobilization of research participants. We would like to thank the participants for agreeing to participate in this study. The entire staff at the Centre for Biotechnology Research and Development (CBRD), Kenya Medical Research Institute (KEMRI) for support in terms of resources required as well as moral support. Special mention to Kiritu Nyaga ILRI; Leonard Ndwiga; KEMRI Welcome and Kimita Gathii; WRAIR for their assistance in instrument access and sequencing troubleshooting.

References

[1] W. Parson, A. Dür, EMPOP-A forensic mtDNA database, Forensic Sci Int Genet 1 (2007) 88–92, https://doi.org/10.1016/j.fsigen.2007.01.018.
[2] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA commission of the international society for forensic genetics (ISFG) on quality control of autosomal short tandem repeat allele frequency databasing (STRidER), Forensic Sci Int Genet 24 (2016) 97–102, https://doi.org/10.1016/j.fsigen.2016.06.008.
[3] S. Willuweit, L. Roewer, The new y chromosome haplotype reference database, Forensic Sci Int Genet 15 (2015) 43–48, https://doi.org/10.1016/j.fsigen.2014.11.024.
[4] M. Kruijver, J.A. Bright, H. Kelly, J. Buckleton, Exploring the probative value of mixed DNA profiles, Forensic Sci Int Genet 41 (2019) 1–10, https://doi.org/10.1016/j.fsigen.2019.03.010.
[5] W.C. Thompson, R.H. Grady, E. Lai, H.S. Stern, Perceived strength of forensic scientists' reporting statements about source conclusions, Law Probab. Risk 17 (2018) 133–155, https://doi.org/10.1093/lpr/mgy012.
[6] C.D. Steele, D.J. Balding, Choice of population database for forensic DNA profile analysis, Sci. Justice 54 (2014) 487–493, https://doi.org/10.1016/j.scijus.2014.10.004.
[7] M.E. D'Amato, Y. Joly, V. Lynch, H. Machado, N. Scudder, M. Zieger, Ethical considerations for forensic genetic frequency databases: First report conception and development, Forensic Sci Int Genet 71 (2024), https://doi.org/10.1016/j.fsigen.2024.103053.
[8] M. Bodner, W. Parson, The strider report on two years of quality control of autosomal str population datasets, Genes 11 (2020) 1–15, https://doi.org/10.3390/genes11080901.
[9] C.R. Taylor, K.M. Kiesler, K. Sturk-Andreaggi, J.D. Ring, W. Parson, M. Schanfield, P.M. Vallone, C. Marshall, Platinum-quality mitogenome haplotypes from United States populations, Genes 11 (2020) 1–25, https://doi.org/10.3390/genes11111290.
[10] M. Zieger, S. Utz, A "forensic biobank" to establish comprehensive genetic frequency data for Switzerland, Forensic Sci Int Genet 40 (2019) 46–51, https://doi.org/10.1016/j.fsigen.2019.02.004.
[11] L. Caenazzo, P. Tozzo, Microbiome forensic biobanking: a step toward microbial profiling for forensic human identification, Health Care 9 (2021), https://doi.org/10.3390/healthcare9101371.
[12] J.H. Smith, J.S. Horne, Towards ethical forensic practice: undertaking research on biomaterials and applications for forensic science, Forensic Sci. Int. 8 (2024), https://doi.org/10.1016/j.fsisyn.2024.100459.
[13] V. Dranseika, J. Piasecki, M. Waligora, Forensic uses of research biobanks: should donors be informed? Med Health Care Philos 19 (2016) 141–146, https://doi.org/10.1007/s11019-015-9667-0.
[14] G. Sguazzi, G. Fasani, F. Renò, S. Gino, Biobanks: archives or resources? Their secondary use for forensic purposes—A systematic review, Forensic Sci. 4 (2024) 42–61, https://doi.org/10.3390/forensicsci4010004.
[15] B. Budowle, A. Sajantila, Revisiting informed consent in forensic genomics in light of current technologies and the times, Int. J. Leg. Med. 137 (2023) 551–565, https://doi.org/10.1007/s00414-023-02947-w.
[16] M. Bodner, W. Parson, Informed consent for forensic genetic population studies: status quo and a call for harmonization, Forensic Sci Int Genet 78 (2025), https://doi.org/10.1016/j.fsigen.2025.103298.
[17] D'Amato M.E, Joly Y, Lynch V, Ethical Considerations for Forensic Genetic Frequency Databases, (n.d.).
[18] S.M.C. Gibbons, J. Kaye, GOVERNING GENETIC DATABASES: COLLECTION, STORAGE AND USE Europe PMC Funders Group, 2007.
[19] M.L. Hale, T.M. Burg, T.E. Steeves, Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies, PLoS One 7 (2012), https://doi.org/10.1371/journal.pone.0045170.
[20] E.P. Flesch, J.J. Rotella, J.M. Thomson, T.A. Graves, R.A. Garrott, Evaluating sample size to estimate genetic management metrics in the genomics era, Mol Ecol Resour 18 (2018) 1077–1091, https://doi.org/10.1111/1755-0998.12898.
[21] Kenya Gazette, THE KENYA DATA PROTECTION ACT, 2021.
[22] E.K.A. Nambati, E. Muge, B. Azzam, C. Marshall, K.S. Andreaggi, D. Maritim, N. K. Mwiwawi, L. Kamau, S. Mpoke, W. Bulimo, Quality human mitogenome sequence data amenable for use as population reference data obtained from manual processing using the Collibri™ ES DNA library prep kit, Forensic Genomics (2025), https://doi.org/10.1089/forensic.2024.0010.
[23] M. Cerezo, A. Achilli, A. Olivieri, U.A. Perego, A. Gómez-Carballa, F. Brisighelli, H. Lancioni, S.R. Woodward, M. López-Soto, Á. Carracedo, C. Capelli, A. Torroni, A. Salas, Reconstructing ancient mitochondrial DNA links between Africa and Europe, Genome Res. 22 (2012) 821–826, https://doi.org/10.1101/gr.134452.111.
[24] A. Brandstätter, C.T. Peterson, J.A. Irwin, S. Mpoke, D.K. Koech, W. Parson, T. J. Parsons, Mitochondrial DNA control region sequences from Nairobi (kenya): inferring phylogenetic parameters for the establishment of a forensic database, Int. J. Leg. Med. 118 (2004) 294–306, https://doi.org/10.1007/s00414-004-0466-z.
[25] A. Salas, M. Richards, M.-V. Lareu, R. Scozzari, A. Coppa, A. Torroni, V. Macaulay, A. Ngel Carracedo, The African Diaspora: Mitochondrial DNA and the Atlantic Slave Trade, 2004.
[26] M.K. Gonder, H.M. Mortensen, F.A. Reed, A. De Sousa, S.A. Tishkoff, Whole-mtDNA genome sequence analysis of ancient African lineages, Mol. Biol. Evol. 24 (2007) 757–768, https://doi.org/10.1093/molbev/msl209.
[27] J.M. Muinde, D.R. Chandra Bhanu, R. Neumann, R.O. Oduor, W. Kanja, J. K. Kimani, M.W. Mutugi, L. Smith, M.A. Jobling, J.H. Wetton, Geographical and linguistic structure in the people of Kenya demonstrated using 21 autosomal STRs, Forensic Sci Int Genet 53 (2021), https://doi.org/10.1016/j.fsigen.2021.102535.
[28] Kagunda Joseph Kimani, l(Sixteen Loci a Thesis Submitted IN Partial Fulfillment Of The Requirements For The Award Of TIQ: pEGREE Of Mastel} Of Science (Biotechnology), Of Kenyatta University, 2009. https://ir-library.ku.ac.ke/items/5867803b-29ab-4d2f-a5a0-55928ba16ccc (accessed June 2, 2025).