Forensic Science International: Genetics 53 (2021) 102535

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

# Geographical and linguistic structure in the people of Kenya demonstrated using 21 autosomal STRs

Jane Mbithe Muinde [a], Devi R. Chandra Bhanu [b], Rita Neumann [b], Richard Okoth Oduor [a], Wangu Kanja [c], Joseph Kagunda Kimani [d], Marion W. Mutugi [e], Lisa Smith [f], Mark A. Jobling [b,*], Jon H. Wetton [b,*]

[a] Department of Biochemistry, Microbiology & Biotechnology, Kenyatta University, Nairobi, Kenya
[b] Department of Genetics & Genome Biology, University of Leicester, Leicester, UK
[c] Wangu Kanja Foundation, Nairobi, Kenya
[d] Forensic Biology Section, Government Chemist Department, Nairobi, Kenya
[e] Amref International University, Nairobi, Kenya
[f] Department of Criminology, University of Leicester, Leicester, UK

ARTICLE INFO

*Keywords:*
Kenya
Autosomal STRs
Population structure
Linguistic structure
GlobalFiler

ABSTRACT

Kenya is a diverse and populous nation that employs DNA evidence in its criminal justice system, and therefore requires reliable information on autosomal STR allele frequency variation across the country and in its many ethnic groups. In order to provide reference data and to assess population structure, we analysed the 21 autosomal STRs in the GlobalFiler multiplex in a sample of 510 indigenous Kenyans representing the country's eight former provinces, 43 of its 47 counties, three main linguistic families and all 29 ethnic groups that each comprise >0.5% of the 2019 census population. The indigenous population originated from successive migrations of Cushitic, Nilotic and Bantu speaking groups who settled in regions that suited their distinctive sustenance lifestyles. Consequently, they now largely reside in a patchwork of communities with strong associations with particular counties and provinces and limited degrees of inter-group marriage, as shown by DNA donors' ancestry details. We found significant genetic differentiation between the three Nilotic language sub-families, with Western Nilotes (the Luo ethnic group) showing greater similarity to the Bantu than the Southern and Eastern Nilotes which themselves showed closer affinity to the Cushitic speakers. This concurs with previous genetic, linguistic and social studies. Comparisons with other African populations also showed that linguistic affiliation is a stronger factor than geography. This study revealed several rare off-ladder alleles whose structure was determined by Sanger sequencing. Among the unusual features that could affect profile interpretation were a deletion of Amelogenin Y but no other forensic marker (autosomal or Y-chromosomal), a triallelic pattern at TPOX and an extremely short SE33 allele falling within the expected size range of D7S820. Compared with the currently implemented Identifiler multiplex, Random Match Probabilities decreased from $6.4 \times 10^{-19}$ to $3.9 \times 10^{-27}$. The appreciation of local population structure provided by the geographically and ethnically representative sample in this study highlights the structured genetic landscape of Kenya.

## 1. Introduction

Kenya has an ancient history of human occupation dating back to the dawn of our species [1], and today has a complex population of 53 million people, including speakers of 58 languages within three divergent language families (Fig. 1), and over 70 recognised ethnic groups [2]. Samples from two of these groups, Maasai in Kinyawa, Kenya [MKK] and Luhya in Webuye, Kenya [LWK] formed part of the International Haplotype Mapping (HapMap) Consortium project [3], and two others, Kikuyu and Kalenjin, were included in the African Genome Variation Project [4]. These have been genotyped using genome-wide single-nucleotide polymorphism (SNP) chips, and the LWK sample has also been whole-genome sequenced as part of the 1000 Genomes Project [5]. Several further Kenyan populations were

* Corresponding authors.
*E-mail addresses:* maj4@le.ac.uk (M.A. Jobling), jw418@le.ac.uk (J.H. Wetton).

https://doi.org/10.1016/j.fsigen.2021.102535
Received 7 March 2021; Received in revised form 11 May 2021; Accepted 16 May 2021
Available online 19 May 2021
1872-4973/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

included in a pan-African STRUCTURE analysis based on 848 autosomal STR markers which demonstrated that the country was the nexus of several distinct genetic clusters that correlated strongly with linguistic families [6]. Although only nine forensic STR markers were included in that screening panel [7] it highlighted that considerable genetic diversity could be expected in the region. Beyond this, much about the genetic diversity of Kenya remains unknown, and yet such knowledge is key to understanding the history and prehistory of East Africa. At a practical level, such information is essential for the unbiased interpretation of forensic DNA evidence in the Kenyan criminal justice system.

Administratively, Kenya was until 2010 divided into eight provinces (Fig. 1a) and today has a finer-scale division comprising 47 counties (Fig. S1). Within this recently established geographical framework, individual ethnic groups are highly structured. Although English and Swahili are official national languages, ethnic groups also have their own languages and these too are strongly geographically differentiated. Bantu speakers comprise approximately 60% of the population and are predominantly associated with Kikuyu, Luhya and Kamba ethnic groups, which occupy the central and southern parts of the country. Bantu languages are otherwise widespread in sub-Saharan Africa, owing their expansion to an agriculturally mediated transition beginning in West Central Africa around 5000 years ago [8]. Nilotic speakers (~31%), representing all three major subfamilies, such as the Southern-Nilotic-speaking Kalenjin ethnic group, the Eastern-Nilotic-speaking Maasai and the Western-Nilotic-speaking Luo are largely concentrated in the west of Kenya. This language family is also found in adjacent regions including South Sudan, Eastern Uganda and parts of Tanzania. Speakers of Cushitic languages, a branch of Afro-Asiatic, are predominantly found in the eastern and northern parts of Kenya and include ethnic groups such as the Somali, and Borana. To the north, the Oromo of Ethiopia are also Cushitic-speaking. Some regions of Kenya, including the Capital province, Nairobi, are cosmopolitan and thus home to diverse people speaking a complex mixture of languages.

Published data on Kenyan autosomal short-tandem repeat (STR) variation are scanty and limited to 15-locus Identifiler profiles collected in three cities (Nairobi, Mombasa, Kisumu) where random sets of 50 individuals each were sampled [9]. Given the complex ethnolinguistic diversity of the country, data on more extensive and better-defined samples would provide a picture of autosomal genetic structure and would be helpful in interpretation of forensic evidence. This study applies the 21 autosomal STRs of the GlobalFiler multiplex to a set of samples from 510 Kenyan males with wide geographical, linguistic and ethnic group representation. We compile allele frequency data and calculate forensic statistics, and ask whether there is significant population structure with regard to geographical and linguistic affiliation.

## 2. Materials and methods

### 2.1. DNA sampling

Recruitment and sampling of DNA donors was done in accordance with the ethical guidelines of *FSI: Genetics* [10]. Ethical review for donor recruitment and DNA analysis was provided independently by the research ethics committees of Kenyatta University (ref. KU/ER-C/EXTEN.APPR.1.VOL.1 [10]) and the University of Leicester (ref. 16000-maj4-ls/gg). Informed written consent was provided by all participants. We sampled males only, to support future Y-STR reference data collection.

Samples were collected between November 2018 and February 2020 from 510 unrelated indigenous Kenyan males (students enrolled at Kenyatta University) who provided information on the birthplaces, languages and ethnic group affiliations of themselves, their parents, and their grandparents. We classified them in terms of ancestry in one of the eight historical provinces (Fig. 1a; Central N = 43; Coast N = 69; Eastern N = 61; Nairobi N = 22; North Eastern N = 4; Nyanza N = 129; Rift Valley N = 120; Western N = 62), and by linguistic affiliation based upon information in Glottolog [11] (Fig. 1c; Bantu N = 293; Cushitic N = 25; Nilotic N = 192 (comprising Western N = 94; Southern N = 73; Eastern N = 25)). In total, 29 of the officially recognised 45 indigenous ethnic groups were sampled, including all those comprising >0.5% of the population (based on the 2019 census, available at www.knbs.or.ke), and 43 of the 47 administrative counties. Proportional coverage by province, ethnic group and linguistic grouping with respect to both our dataset and the census population is shown in Tables S1–S3 and Fig. S1. The close correspondence between actual population density (Fig. 1b) and sampling density can best be seen at the county level (Fig. S1b).

### 2.2. DNA extraction and quantification

Buccal cells collected with sterile CytoSoft* Cytology Brushes (Fisher Scientific) were suspended and lysed in 1 mL of NDS solution (0.5 M EDTA, 10 mM Tris-HCl, 1%[w/v] N-lauroylsarcosine [pH 9.5] [12]) in a

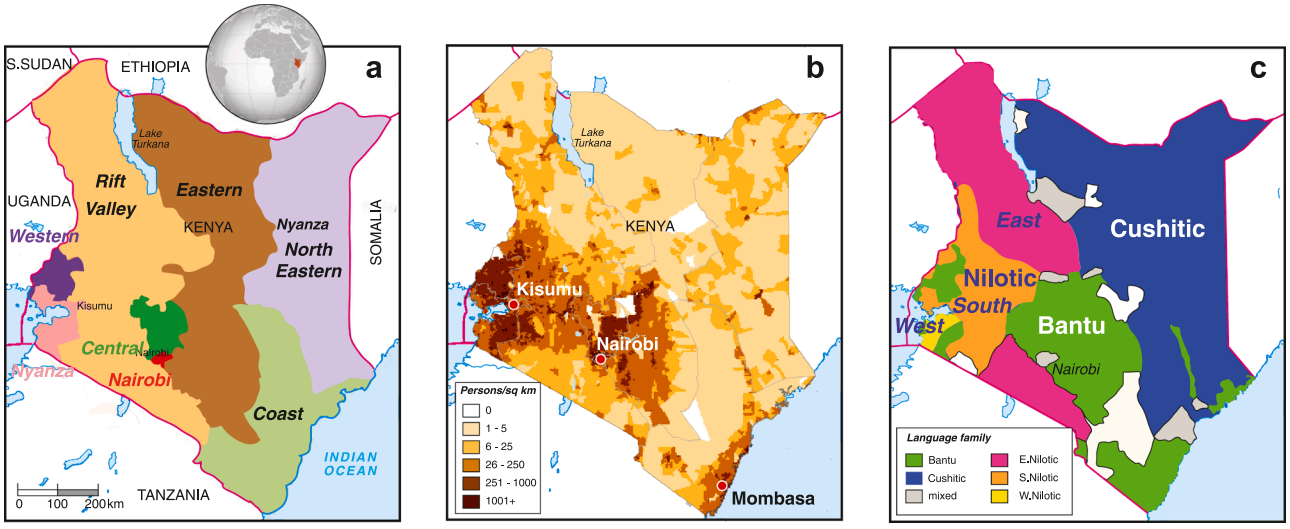


**Fig. 1.** Provinces, population density and language families within Kenya. (a) Map showing boundaries of the eight Kenyan provinces; (b) Population density. (c) Language families. Approximate distributions of divisions of Nilotic are indicated as East, West and South.
Part (b) adapted from Global Rural-Urban Mapping Project (sedac.ciesin.columbia.edu/gpw/). Part (c) drawn from data in Ethnologue (www.ethnologue.com/country/KE and [11]).

2-mL screw-cap tube for ambient storage. DNA was extracted from 200 μL cell lysate using Wizard Genomic DNA Purification Kits (Promega) and a vacuum manifold protocol as recommended by the manufacturer. The quality and quantity of the extracted DNA was estimated using a Nanodrop ND-8000 spectrophotometer (Thermo Fisher Scientific).

### 2.3. DNA amplification and fragment detection

Autosomal profiles based on 21 STRs were generated using the GlobalFiler® PCR amplification kit, analysing the loci D3S1358, vWA, D16S539, CSF1PO, TPOX, D8S1179, D21S11, D18S51, D2S441, D19S433, TH01, FGA, D22S1045, D5S818, D13S317, D7S820, SE33, D10S1248, D1S1656, D12S391 and D2S1338, along with three additional markers used as a sex test (DYS391, Y indel and Amelogenin). Amplification was performed in an MJ Research Tetrad thermal cycler and genotyping done on an ABI3500xL Genetic Analyzer (Thermo Fisher Scientific) following the manufacturer's recommendations. Allele calling and interpretation were carried out using GeneMapper IDX software V1.5.

This study followed the publication guidelines of *FSI: Genetics* for population genetic data [13–15] and allele nomenclature [16]. The dataset was submitted to STRidER [17] for quality control and approved (reference STR000351).

### 2.4. DNA sequencing

Off-ladder and otherwise unusual alleles were re-amplified in singleplex using unlabelled primers (Table S4; [18]), separated via agarose gel electrophoresis, and bi-directionally sequenced using standard Sanger technology. Sequences are available from GenBank under accession numbers MZ090893–MZ090901.

### 2.5. Forensic and statistical analysis

The software package ML-RELATE [19], in conjunction with Yfiler Plus profiling (data not shown) and ancestry information, was used to screen for any undeclared brothers within the dataset. STRAF 1.0.5 [20] was used to calculate allele frequencies, Random Match Probability (PM), Power of Discrimination (PD), Power of Exclusion (PE), Typical Paternity Index (TPI), Genetic Diversity (GD, also equivalent to expected heterozygosity) and observed heterozygosity. Arlequin v 3.5.2.2 [21] was used to test Hardy-Weinberg equilibrium, to calculate expected heterozygosity, and to perform AMOVA on genotypes and $R_{ST}$ genetic distance matrices to investigate genetic diversity within and between the eight provinces and five language groups. Genetic distances were visualised using Principal Coordinates Analysis (PCoA) within the GenAlEx suite [22] for language groups, provinces and counties within Kenya using the full GlobalFiler profiles. Distances were calculated from the Pairwise Population Matrix of Mean Population Codominant Genotypic Genetic Distances and the scaling on the axes corresponds to the resultant Eigenvalues. Pairwise $F_{ST}$ values calculated with POPTREE2 [23] were based upon allele frequencies at the 15 STR loci targeted by the AmpF*l*STR Identifiler kit (CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, THO1, TPOX and vWA) that were shared with other African population studies. The $F_{ST}$ values were also used to generate multidimensional scaling (MDS) plots with the (MASS) package [24] in the R library. The data generated here were compared with data from other populations as follows: African Americans in USA N = 543 [25], amaXhosa and amaZulu in South Africa N = 120 and N = 100 respectively [26]; Angolans in Cabinda N = 152 [27]; Botswana N = 990 [28]; Equatorial Guineans in Madrid N = 134 [29]; Guinea Bissauans in Portugal N = 70 [27]; Kenya N = 150 [9]; Mozambique N = 42 [30]; Namibian Ovambo (Bantu) N = 195 [31]; Rwandan Tutsi (Bantu) N = 124 [32]; Saudi Arabians N = 523 [33]; Somalians resident in Denmark N = 404 [34]; Sudan – N = 498 [35]; Meru (Bantu) from Northern Tanzania N = 172 [36]; Ugandan Karamoja N = 218 [37].

## 3. Results

### 3.1. Ethnolinguistic composition of the sample set

Following recruitment of 510 DNA donors, we analysed supplied personal details to understand their linguistic, geographic and ethnic backgrounds, and to seek evidence for recent admixture. In our dataset linguistic and ethnic group affiliation were perfectly correlated for Western-Nilotic, which was spoken solely by members of the Luo ethnic group, and Southern-Nilotic by the Kalenjin.

All 25 donors belonging to the Cushitic linguistic group had both parents from the same ethnic group as themselves. This was also true of 95.9% (70/73) of Southern-Nilotic, 90.4% (85/94) of Western-Nilotic and 84.0% (21/25) of Eastern-Nilotic speakers. When a parent was from a different ethnic group, this was in all cases the mother, who was from a Bantu group. Among the Bantu speakers themselves, 89.4% of parents were from the same ethnic group, 6.5% were from different Bantu groups, seven (2.4%) had a Western-Nilotic, four (1.4%) a Southern-Nilotic and one (0.3%) a Cushitic-speaking parent; in all but one case the non-Bantu parent was the mother. Overall, 91.4% of donors' parents came from the same ethnic group and 95.1% from the same language group.

Most DNA donors had both parents born in their own province of birth: North-Eastern (100%, N = 4), Nyanza (93.8%, N = 129), Coast (88.4%, N = 69), Eastern (86.9%, N = 61), Central (83.7%, N = 43), Western (80.6%, N = 62), and Rift Valley (76.7%, N = 120). This was not true of donors who were born in Nairobi Province; while both of their parents still tended to be born in the same province as each other (68.2%, N = 22) all but two of these couples (9.1%) had moved to Nairobi Province from another part of Kenya before the birth of their child, reflecting the influx to the capital in recent years. Overall, 87.1% of parents were both born in the same province and 69.2% in the same county. Only five of 1020 parents were born outside of Kenya, in either Ethiopia or Uganda.

There is a clear association between language and province of birth which is particularly marked for the Southern and Eastern Nilotes (97.3% and 88.0% from Rift Valley respectively) and Western Nilotes of whom 90.3% were from Nyanza. The sampled Cushitic speakers originated from the Coast (36%), Eastern (28%), North Eastern (16%) and Nairobi Provinces (20%), whilst the Bantu were broadly distributed across all but North Eastern Province though at a lower proportion from the predominantly Nilotic Rift Valley (Table S5, Fig. S1, online resources https://microreact.org/project/qACZKEyuTqviiDDkSTy3A4/c92f2238, https://microreact.org/project/eBpydzMtpSz3RV3SiFNnRN/d011b03f [Province and County map respectively]).

### 3.2. Description of data and forensic statistics

The 21 autosomal STRs included in the GlobalFiler kit were analysed in 510 Kenyan men. In Table 1 we present allele frequency data and forensic statistics for the entire Kenyan dataset; Table S6 provides the same measures for each of the eight Kenyan provinces (Fig. 1a) and five language groups (Fig. 1c).

The least variable loci are THO1 and D16S539, each with eight allelic variants in the Kenyan dataset, and the most variable locus was SE33 with 45 alleles. While SE33 gave the highest Probability of Discrimination (0.9889), the lowest (0.8770) was provided by D3S1358. In comparison to the Identifiler multiplex currently used in Kenyan forensic investigations, GlobalFiler decreased the random match probability from $6.4 \times 10^{-19}$ to $3.9 \times 10^{-27}$.

### 3.3. Rare variants and off-ladder alleles

Forty-two alleles were each observed only once in the entire dataset. Of these, six were also globally rare, and were among the eight off-ladder alleles recorded at these loci: CSF1PO (9.3 [$N_{obs}$ = 1]), D12S391 (24.2

**Table 1**
Kenyan allele frequencies and forensic statistics.

| Allele | CSF1PO | D10S1248 | D12S391 | D13S317 | D16S539 | D18S51 | D19S433 | D1S1656 | D21S11 | D22S1045 | D2S1338 | D2S441 | D3S1358 | D5S818 | D7S820 | D8S1179 | FGA | SE33 | TH01 | TPOX | vWA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **–1** | | | | | | | | | | | | | | | | | | 0.002 | | | |
| **4.2** | | | | | | | | | | | | | | | | | | 0.002 | | | |
| **5.2** | | | | | | | | | | | | | | | | | | 0.002 | | | |
| **6** | | | | | | | | | | | | | | | | | | | 0.180 | 0.051 | |
| **7** | 0.044 | | | | | | | | | | | | | | 0.006 | | | | 0.391 | 0.009 | |
| **7.1** | | | | | | | | | | | | | | | 0.001 | | | | | | |
| **7.2** | | | | | | | | | | | | | | | | | | 0.001 | | | |
| **8** | 0.044 | | | 0.043 | 0.031 | | | | | | | | | 0.068 | 0.204 | | | | 0.225 | 0.291 | |
| **9** | 0.080 | 0.006 | | 0.023 | 0.211 | | 0.003 | | | | | 0.002 | | 0.028 | 0.128 | 0.004 | | | 0.149 | 0.307 | |
| **9.3** | 0.001 | | | | | | | | | | | | | | | | | | 0.044 | | |
| **10** | 0.262 | 0.002 | | 0.030 | 0.099 | 0.003 | 0.013 | 0.003 | | 0.027 | | 0.046 | | 0.060 | 0.363 | 0.016 | | 0.002 | 0.008 | 0.072 | |
| **10.1** | | | | | | | | | | | | | | | 0.001 | | | | | | |
| **10.2** | | | | | | 0.002 | 0.001 | | | | | | | | | | | 0.002 | | | |
| **11** | 0.210 | 0.064 | | 0.292 | 0.289 | 0.002 | 0.068 | 0.047 | | 0.174 | | 0.314 | 0.001 | 0.174 | 0.185 | 0.055 | | 0.003 | 0.001 | 0.227 | 0.004 |
| **11.1** | | | | | | | | | | | | 0.002 | | | | | | | | | |
| **11.2** | | | | | | | 0.001 | | | | | | | | | | | 0.010 | | | |
| **11.3** | | | | 0.001 | | | | | | | | 0.044 | | | | | | | | | |
| **12** | 0.294 | 0.133 | | 0.393 | 0.232 | 0.055 | 0.112 | 0.062 | | 0.043 | | 0.150 | 0.002 | 0.391 | 0.097 | 0.145 | | 0.005 | 0.001 | 0.039 | 0.001 |
| **12.1** | | | | | | | | | | | | | | 0.001 | | | | | | | |
| **12.2** | | | | | | | 0.025 | | | | | | | | | | | 0.008 | | | |
| **12.3** | | | | | | | | | | | | 0.004 | | | | | | | | | |
| **13** | 0.052 | 0.258 | | 0.153 | 0.128 | 0.043 | 0.220 | 0.115 | | 0.009 | | 0.044 | 0.003 | 0.256 | 0.015 | 0.159 | | 0.015 | | 0.003 | 0.010 |
| **13.2** | | | | | | 0.004 | 0.066 | | | | | | | | | | | 0.011 | | | |
| **14** | 0.011 | 0.280 | 0.006 | 0.064 | 0.008 | 0.055 | 0.233 | 0.216 | | 0.060 | | 0.325 | 0.066 | 0.020 | | 0.312 | | 0.028 | | | 0.090 |
| **14.2** | | | | | | 0.001 | 0.069 | | | | | | | | | | | 0.009 | | | |
| **14.3** | | | | | | | | 0.019 | | | | 0.001 | | | | | | | | | |
| **15** | 0.002 | 0.157 | 0.094 | 0.001 | 0.001 | 0.126 | 0.095 | 0.188 | | 0.251 | 0.002 | 0.061 | 0.325 | 0.003 | | 0.224 | | 0.048 | | 0.001 | 0.197 |
| **15.2** | | | | | | 0.005 | 0.054 | | | | | | | | | | | 0.004 | | | |
| **15.3** | | | | | | | | 0.023 | | | | | | | | | | | | | |
| **16** | | 0.080 | 0.097 | | | 0.175 | 0.018 | 0.148 | | 0.274 | 0.048 | 0.007 | 0.312 | | | 0.071 | | 0.075 | | | 0.249 |
| **16.1** | | | | | | | | | | | | | | | | | 0.001 | | | | |
| **16.2** | | | | | | 0.006 | 0.022 | | | | | | 0.001 | | | | | | | | |
| **16.3** | | | | | | | | 0.096 | | | | | | | | | | | | | |
| **17** | | 0.019 | 0.175 | | | 0.172 | 0.002 | 0.026 | | 0.154 | 0.129 | | 0.239 | | | 0.012 | 0.002 | 0.102 | | | 0.183 |
| **17.1** | | | 0.002 | | | | | | | | | | | | | 0.001 | | | | | |
| **17.2** | | | | | | 0.005 | | | | | | | 0.002 | | | | 0.001 | 0.002 | | | |
| **17.3** | | | 0.001 | | | | | 0.036 | | | | | | | | | | | | | |
| **18** | | 0.001 | 0.251 | | | 0.133 | | 0.004 | | 0.009 | 0.061 | | 0.047 | | | 0.003 | 0.021 | 0.143 | | | 0.143 |
| **18.1** | | | 0.001 | | | | | | | | | | | | | | | | | | |
| **18.2** | | | | | | 0.001 | | | | | | | | | | | 0.002 | 0.001 | | | |
| **18.3** | | | | | | | | 0.012 | | | | | | | | | | | | | |
| **19** | | | 0.137 | | | 0.128 | | 0.001 | | | 0.155 | | 0.002 | | | | 0.036 | 0.125 | | | 0.085 |
| **19.1** | | | 0.005 | | | | | | | | | | | | | | | | | | |
| **19.2** | | | | | | | | | | | | | | | | | 0.006 | | | | |
| **19.3** | | | | | | | | 0.005 | | | | | | | | | | | | | |
| **20** | | | 0.070 | | | 0.056 | | | | | 0.108 | | | | | | 0.049 | 0.063 | | | 0.028 |
| **20.2** | | | | | | | | | | | | | | | | | | 0.004 | | | |
| **21** | | | 0.055 | | | 0.020 | | | | | 0.120 | | | | | | 0.095 | 0.051 | | | 0.008 |
| **21.2** | | | | | | | | | | | | | | | | | | 0.009 | | | |
| **22** | | | 0.031 | | | 0.009 | | | | | 0.149 | | | | | | 0.201 | 0.011 | | | 0.001 |
| **22.2** | | | | | | | | | | | | | | | | | | 0.004 | | | |
| **23** | | | 0.044 | | | | | | | | 0.080 | | | | | | 0.165 | 0.003 | | | |
| **23.2** | | | | | | | | | | | | | | | | | | 0.006 | | | |
| **23.3** | | | | | | | | | 0.002 | | | | | | | | 0.002 | | | | |
| **24** | | | 0.015 | | | | | | | | 0.074 | | | | | | 0.146 | 0.001 | | | |

(*continued on next page*)

**Table 1** (*continued*)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **24.2** | | | 0.001 | | | | | | | | | | | | | | 0.022 | | | | |
| **24.3** | | | | | | | | | 0.013 | | | | | | | | | | | | |
| **25** | | | 0.010 | | | | | | | | 0.041 | | | | | 0.110 | | | | | |
| **25.2** | | | 0.001 | | | | | | | | | | | | | | 0.033 | | | | |
| **25.3** | | | | | | | | | 0.002 | | | | | | | | | | | | |
| **26** | | | 0.004 | | | | | | 0.003 | | 0.026 | | | | | 0.063 | | | | | |
| **26.1** | | | | | | | | | | | | | | | | 0.001 | | | | | |
| **26.2** | | | | | | | | | | | | | | | | | 0.056 | | | | |
| **27** | | | 0.001 | | | | | | 0.054 | | 0.007 | | | | | 0.039 | 0.001 | | | | |
| **27.2** | | | | | | | | | | | | | | | | | 0.049 | | | | |
| **28** | | | | | | | | | 0.232 | | | | | | | 0.026 | 0.003 | | | | |
| **28.2** | | | | | | | | | | | | | | | | | 0.033 | | | | |
| **29** | | | | | | | | | 0.213 | | | | | | | 0.015 | | | | | |
| **29.2** | | | | | | | | | | | | | | | | | 0.030 | | | | |
| **30** | | | | | | | | | 0.190 | | | | | | | 0.007 | | | | | |
| **30.2** | | | | | | | | | 0.005 | | | | | | | 0.002 | 0.006 | | | | |
| **31** | | | | | | | | | 0.072 | | | | | | | 0.003 | | | | | |
| **31.2** | | | | | | | | | 0.036 | | | | | | | 0.001 | 0.006 | | | | |
| **32** | | | | | | | | | 0.014 | | | | | | | | | | | | |
| **32.1** | | | | | | | | | 0.001 | | | | | | | | | | | | |
| **32.2** | | | | | | | | | 0.063 | | | | | | | | 0.006 | | | | |
| **33** | | | | | | | | | 0.004 | | | | | | | 0.001 | 0.001 | | | | |
| **33.1** | | | | | | | | | 0.002 | | | | | | | | | | | | |
| **33.2** | | | | | | | | | 0.034 | | | | | | | | 0.003 | | | | |
| **34** | | | | | | | | | 0.015 | | | | | | | | 0.002 | | | | |
| **34.2** | | | | | | | | | 0.003 | | | | | | | | | | | | |
| **35** | | | | | | | | | 0.024 | | | | | | | | | | | | |
| **36** | | | | | | | | | 0.014 | | | | | | | | | | | | |
| **37** | | | | | | | | | 0.004 | | | | | | | | | | | | |
| **38** | | | | | | | | | 0.002 | | | | | | | | | | | | |
| **41.2** | | | | | | | | | | | | | | | | 0.001 | | | | | |
| **42.2** | | | | | | | | | | | | | | | | 0.002 | | | | | |
| **43.2** | | | | | | | | | | | | | | | | 0.002 | | | | | |
| **44.2** | | | | | | | | | | | | | | | | 0.001 | | | | | |
| | **CSF1PO** | **D10S1248** | **D12S391** | **D13S317** | **D16S539** | **D18S51** | **D19S433** | **D1S1656** | **D21S11** | **D22S1045** | **D2S1338** | **D2S441** | **D3S1358** | **D5S818** | **D7S820** | **D8S1179** | **FGA** | **SE33** | **THO1** | **TPOX** | **vWA** |
| **N** | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 | 1020 |
| **Nall** | 10 | 10 | 20 | 9 | 8 | 20 | 16 | 16 | 23 | 9 | 13 | 12 | 11 | 9 | 9 | 11 | 28 | 45 | 8 | 9 | 12 |
| **PIC** | 0.757 | 0.774 | 0.844 | 0.688 | 0.759 | 0.866 | 0.843 | 0.851 | 0.832 | 0.774 | 0.881 | 0.729 | 0.686 | 0.704 | 0.732 | 0.770 | 0.868 | 0.924 | 0.699 | 0.721 | 0.806 |
| **PM** | 0.073 | 0.068 | 0.037 | 0.120 | 0.076 | 0.028 | 0.034 | 0.033 | 0.040 | 0.066 | 0.024 | 0.099 | 0.123 | 0.109 | 0.086 | 0.067 | 0.026 | 0.011 | 0.108 | 0.097 | 0.050 |
| **PD** | 0.927 | 0.932 | 0.963 | 0.880 | 0.924 | 0.972 | 0.966 | 0.967 | 0.960 | 0.934 | 0.976 | 0.901 | 0.877 | 0.891 | 0.914 | 0.933 | 0.974 | 0.989 | 0.892 | 0.903 | 0.950 |
| **PE** | 0.498 | 0.563 | 0.705 | 0.495 | 0.560 | 0.716 | 0.625 | 0.724 | 0.685 | 0.570 | 0.787 | 0.542 | 0.491 | 0.546 | 0.444 | 0.581 | 0.756 | 0.820 | 0.459 | 0.525 | 0.592 |
| **TPI** | 1.947 | 2.277 | 3.446 | 1.932 | 2.257 | 3.592 | 2.684 | 3.696 | 3.228 | 2.318 | 4.811 | 2.161 | 1.917 | 2.179 | 1.723 | 2.383 | 4.180 | 5.667 | 1.783 | 2.073 | 2.452 |
| **Hobs** | *0.743* | *0.780* | *0.855* | *0.741* | *0.778* | *0.861* | *0.814* | *0.865* | *0.845* | *0.784* | *0.896* | *0.769* | *0.739* | *0.771* | *0.710* | *0.790* | *0.880* | *0.912* | *0.720* | *0.759* | *0.796* |
| **GD** | *0.789* | *0.802* | *0.859* | *0.730* | *0.791* | *0.879* | *0.859* | *0.866* | *0.850* | *0.803* | *0.892* | *0.764* | *0.734* | *0.743* | *0.767* | *0.799* | *0.880* | *0.929* | *0.740* | *0.761* | *0.829* |
| **pHW** | *0.881* | *0.075* | *0.072* | *0.663* | *0.535* | *0.021* | *0.888* | *0.784* | *0.889* | *0.503* | *0.491* | *0.084* | *0.062* | *0.294* | *0.104* | *0.816* | *0.646* | *0.404* | *0.639* | *0.916* | *0.456* |

Notes: N: total number of alleles; Nall: number of alleles observed per locus; PIC: polymorphism information content; PM: random match probability; PD: power of discrimination; PE: power of exclusion; TPI: typical paternity index; Hobs: observed heterozygosity; GD: gene diversity (equivalent to Hexp: expected heterozygosity); pHW: Hardy-Weinberg equilibrium p-value.

[1], 25.2 [1]), D21S11 (23.3 [2], D8S1179 (17.1 [1]), FGA (33 [1]) and SE33 (5.2 [2], 7.2 [1]). Off-ladder alleles of similar length have been logged previously in STRBase (http://strbase.nist.gov/index.htm) [38]. Both of the rare off-ladder alleles that were recorded twice (D21S11 23.3 and SE33 5.2) were among the Kalenjin of the Rift Valley, although each allele was traced to different subclans and counties, suggesting the variants may not be uncommon in this ethnic group.

Sanger sequences of off-ladder alleles are presented in Table S7 following both the original NIST STRBase nomenclature and current ISFG guidance [15] to report the forward strand of the human genome reference sequence in outputs from massively parallel sequencing (MPS) studies, as implemented in the NCBI STRSeq database (NCBI Accession: PRJNA380127). These recommendations have led to nomenclature changes for several loci including CSF1PO, which was originally described as an $[AGAT]_n$ repeat but should now be re-designated as the reverse complement $[ATCT]_n$. The 9.3 intermediate CSF1PO allele results from deletion of the base that delimits the boundary between the repeat region and the flanking sequence separating a highly variable $[ATCT]_n$ array from a typically invariant $A[ATCT]_3$ motif. In this case, deletion of the intervening A increases the number of contiguous ATCT tetramer repeats by three despite reducing the overall length by 1 bp. This mutation has been reported previously in NCBI dbSNP as rs1392493641; unfortunately, there is no information on the geographic origin of this rare database variant.

The two D12S391 off-ladder alleles (24.2 and 25.2) both possess a partial repeat (AT) within the initial $[AGAT]_n$ block of the canonical $[AGAT]_n[AGAC]_n[AGAT]$ structure, suggesting they have a common origin, and share this feature with a 22.2 allele in the STRseq database (accession MH167176.1). The two short SE33 alleles (5.2 and 7.2) also display a shared 14-bp deletion removing a dinucleotide and three tetranucleotide repeats ($CT[CTTT]_3$), that is usually considered as part of the flanking region following immediately after the last variable $[CTTT]_n$ block. The same deletion was noted in an 11.2 allele, a length that was observed ten times in this dataset, and in a 10.2 allele that is included in the STRSeq database (accession MH232698.1). The end of the repeat array is also the location of the variant giving rise to the D8S1179 17.1 allele, but in this case an additional T is inserted within the terminal [TCTA] repeat to form a [TCTTA]. A 13-bp deletion starting 12 bp into the 3′ flanking DNA is the cause of the 23.3 allele at D21S11 which has 27 repeats within the region where length variation is normally seen. A 24.3 allele with the same deletion and one extra tetramer in the last repeat block is recorded in STRSeq (MT298832.1); this variant has been observed four times in a UK-based study [39] in three North East Africans and one West African. Unlike the other off-ladder alleles which were of intermediate repeat size, the FGA 33 allele has an unusually long major repeat tract but otherwise matched a previously observed structure of mixed tetramer repeat motifs (similar to FGA allele 30, MH232639.1).

An apparent case of a type-2 [40] tri-allelic pattern with three evenly balanced peaks (8,10,11.3) was noted at D7S820 in an individual who also appeared to be homozygous at SE33 for the uncommon allele 21 (Fig. S2a). However, Sanger sequencing of the apparent D7S820 allele 11.3 showed this instead to be an extremely truncated SE33 allele with a 60-bp 5′ flanking deletion and an apparent length of −1 repeat, similar to an allele previously reported in a Somali individual [41] and an identical structure to a "2-repeat" allele in a Saudi man which had three more tetramers in the longest uninterrupted stretch of CTTT [42]. The one other observation of an apparent D7S820 11.3 allele in the Kenyan dataset was associated with a D7S820 allele 10 peak of twice the height and an apparently homozygous SE33 allele 15; this is also likely to be a SE33 heterozygote for the same truncated allele. A true type-2 tri-allelic pattern (6,10,11) was reproducibly detected at TPOX in one individual (Fig. S2b); we assume that this individual carries an extra copy of the locus, which in 98% of tri-allelic Africans is associated with presence of an allele 10 [43].

One male yielded no peak at Amelogenin Y, despite a standard result at the Y indel and DYS391 loci (Fig. S2c) and a complete Yfiler Plus profile (data not shown). This individual would be wrongly identified as a female if the Amelogenin Y locus alone were to be tested. Most Amelogenin Y-negative cases result from large-scale deletions which encompass some neighbouring Y-STR loci [44], but in this instance the cause appears to be more genomically localised. Attempts to amplify the locus with primers outside of those used in the GlobalFiler kit failed to yield a product from the Y, and Sanger sequencing showed only the X-linked copy, suggesting that a Y-chromosomal deletion had occurred but on a smaller scale than that usually associated with Amelogenin Y deficiency.

### 3.4. Analysis of genetic structure by geography and language family

None of the loci deviated from Hardy-Weinberg equilibrium in the combined Kenyan dataset following Bonferroni correction; however, D10S1248 displayed a significant deficiency of heterozygotes in the Cushitic subset ($p \leq 0.0001$) whilst other loci were clearly in Hardy-Weinberg equilibrium suggesting a possible null allele may be segregating in this population.

As migration into and within Kenya has led to an inter-mixed and patchy distribution of ethnic groupings and languages, we explored population differentiation at the linguistic, geographic and ethnic group levels. AMOVA with respect to language groups showed evidence of inbreeding ($F_{IS} = 0.01331$, $p < 0.01$), which was also significant within the ethnically structured Bantu group ($F_{IS} = 0.01376$, $p < 0.05$). Overall $F_{ST}$ was not significant but pairwise comparisons highlighted several cases of differentiation: Bantu/Southern- and Eastern-Nilotic, Western-Nilotic/Southern- and Eastern-Nilotic, Cushitic/Western-Nilotic (all $p < 0.0001$) and Cushitic/Southern-Nilotic ($p < 0.05$) while Bantu/Cushitic was significant only prior to Bonferroni correction (Table 2).

A similar analysis of the provinces also detected evidence of inbreeding ($F_{IS} = 0.01381$, $p < 0.01$) but this was not the case for any of the provinces individually. Once again overall $F_{ST}$ was not significant but pairwise comparisons showed that Rift Valley was differentiated from both Coast and Nyanza ($p < 0.001$) (Table 3). As both analyses showed that overall $F_{ST}$ was approaching significance, PCoA was used to examine the whole dataset. The relative placement of the sub-populations in the PCoA plot closely mirrored the geographic arrangement of the provinces (Fig. 2). Allele frequency data from published sources was used to extend this to a comparison with neighbouring countries through an MDS plot (Fig. 3) based on pairwise $F_{ST}$ for the 15 Identifiler loci (CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, THO1, TPOX and vWA). In the first instance the Kenyan dataset was treated as a whole and compared with an existing dataset (referred to here as 'Kenya-KMN') comprising 50 individuals randomly sampled from blood transfusion centres in each of the three largest Kenyan cities: Kisumu (sampling Nyanza and the broader Western Province), Mombasa (sampling coastal Kenya) and Nairobi (sampling the Nairobi region and part of the wider Central Province) [9]. The two Kenyan datasets fell close together (Fig. 3a) but ours was displaced towards the Ugandan and Sudanese datasets (both largely Southern- and Eastern-Nilotic) and Somalia (predominantly Cushitic). This probably represents under-sampling of these two groups in the Kenya-KMN dataset, derived largely from Western-Nilotic (Nyanza)

**Table 2**
Pairwise $F_{ST}$ between Kenyan linguistic groups.

| Language family | Bantu | Cushitic | W. Nilotic | S. Nilotic | E. Nilotic |
|---|---|---|---|---|---|
| Bantu | | 0.00684 | 0.30566 | <0.00001 | <0.00001 |
| Cushitic | 0.00442 | | <0.00001 | 0.00391 | 0.23926 |
| W. Nilotic | 0.00031 | 0.00779 | | <0.00001 | <0.00001 |
| S. Nilotic | 0.00377 | 0.00539 | 0.00442 | | 0.09375 |
| E. Nilotic | 0.00848 | 0.00237 | 0.00849 | 0.00299 | |

Note: p-values are given above the diagonal.

**Table 3**
Pairwise $F_{ST}$ between Kenyan provinces.

| Province | Central | Coast | Eastern | Nairobi Province | Rift Valley | N Eastern | Nyanza | Western |
|---|---|---|---|---|---|---|---|---|
| Central | | 0.00977 | 0.82031 | 0.49805 | 0.05371 | 0.34082 | 0.00586 | 0.08496 |
| Coast | 0.00362 | | 0.17969 | 0.77344 | <0.00001 | 0.08008 | 0.06836 | 0.0166 |
| Eastern | 0.00001 | 0.00104 | | 0.9834 | 0.02539 | 0.34766 | 0.03711 | 0.43555 |
| Nairobi | 0.00024 | 0.00001 | 0.00001 | | 0.49902 | 0.40039 | 0.74023 | 0.62012 |
| Rift Valley | 0.00212 | 0.00618 | 0.00202 | 0.00023 | | 0.27246 | <0.00001 | 0.0127 |
| N Eastern | 0.00481 | 0.01363 | 0.00502 | 0.00315 | 0.00767 | | 0.15918 | 0.21289 |
| Nyanza | 0.00296 | 0.00129 | 0.00147 | 0.00001 | 0.00298 | 0.00955 | | 0.97363 |
| Western | 0.00207 | 0.00253 | 0.00016 | 0.00001 | 0.00199 | 0.00737 | 0.00001 | |

Note: p-values are given above the diagonal.

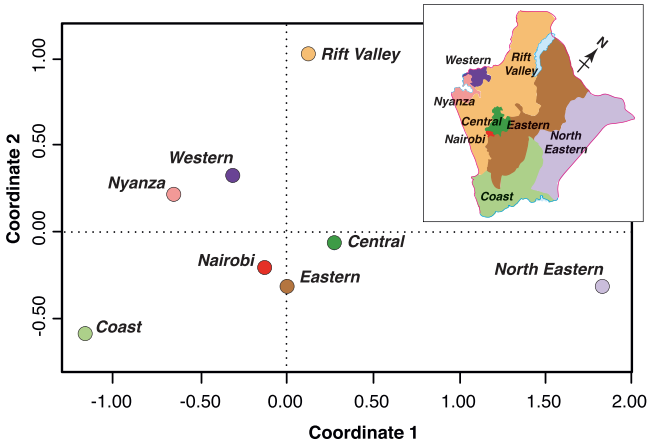


**Fig. 2.** $F_{ST}$ genetic distances between Kenyan provinces. Genetic distances between provinces, visualised through Principal Coordinates Analysis (PCoA); the inset rotated map illustrates the correspondence between genetics and geography.

and Bantu majority provinces (Fig. S3). To explore this further, the current dataset was subdivided into the five linguistic groupings which resulted in the Cushitic subset falling between Somalia and a Saudi Arabian population, the Southern and Eastern Nilotes clustering together with Uganda and Sudan, the Western Nilotes with neighbouring Tanzania and the Bantu closest to the Kenya-KMN dataset and near the other predominantly Bantu datasets (Fig. 3b).

## 4. Discussion

Kenya is a country of great diversity, linguistically, culturally and geographically. Whilst the proportion of Kenyan residents with ancestry outside of Africa is very low (~0.3%, 2019 Census), the indigenous population, broadly sampled in this study, originated largely through ancient migrations from the North (Nilotic-speaking pastoralists), North-East (predominantly Cushitic nomadic groups and coastal traders) and West (Bantu farmers and Luo fishing communities), each group having distinct lifestyles which initially limited mixing [45]. These incoming groups effectively displaced or absorbed the low-density hunter-gatherer populations that previously occupied the region.

Geography played a major role in channelling migration through environments suitable for pursuing these lifestyles. As a result, a patchwork of diverse groups occupying discontinuous but ecologically similar regions was formed. Among neighbouring populations of different origins there was some exchange of agricultural practices, which in some instances progressed to alliances and even mergers between ethnic groups resulting in increased intermarriage, and abandonment of former lifestyles and languages [45]. In this way whilst geography, culture and language contribute to differentiation, these divisions are porous, and groups are open to exchange, merger and complete replacement.

In this study both geography and linguistic affiliation were shown to be significantly associated with genetic differentiation. The Nilotic language family within Kenya is divided into three branches [46]: Eastern, Southern and Western, broadly associated with three ethnic groupings. These are the Plain Nilotes speaking Eastern-Nilotic Maa languages (including the Maasai, Turkana and Samburu ethnic groups of the Rift Valley zone and the isolated Teso ethnic group in Western Province), the Highland Nilotes speaking Southern-Nilotic Kalenjin (the Keiyo, Kipsigis, Marakwet, Nandi, Ogiek, Pokot, Sabaot, Terik and Tugen peoples, which together comprise the Kalenjin ethnic group) and the River Lake Nilotes speaking Western-Nilotic Luo. Eastern and Southern Nilotic were previously grouped as Paranilotic languages [47] largely as a result of their greater proportion of Cushitic loan-words leading to the alternative name of Nilo-Hamitic. This usage has been abandoned as it does not stem from a closer fundamental relationship between these two language groups but may indicate subsequent greater social contact between the Eastern and Southern Nilotic-speaking ethnic groups and Cushitic-speaking peoples [48]. Indeed, we found no evidence from the autosomal STR data that the Plain and Highland ('Paranilotic') groups are differentiated from the Cushitic speakers whilst the River Nilotes (Luo), who largely reside in the west furthest from the predominantly Cushitic regions of Kenya, show significant genetic differentiation from them ($F_{ST}$, $p < 0.0001$). Similarly, the Western-Nilotic Luo are undifferentiated from the neighbouring Bantu population and we found the highest frequency of intermarriage between these groups in the ancestry data provided by donors. The Bantu-speaking Abasuba have taken a further step and become assimilated within the Luo within the past 200 years as a result of intermarriage, shared fishing-based lifestyles, proximity and trade, and have now largely adopted the Western-Nilotic language [49]; at least six Luo donors in this study have Abasuba ancestry. These findings closely mirror those achieved with a much larger panel of non-forensic STRs [6].

Further support for a long-term history of admixture between Nilotic groups and the Bantu comes from anthropometric studies [50,51]. The Sabaot sub-group of the Kalenjin show greater physical similarity with the neighbouring Bantu than with many of their fellow Southern-Nilotic Kalenjin speakers, and the Eastern-Nilotic Teso are not only physically but also culturally closer to the Bantu than other Nilotic-speaking ethnic groups having converted from pastoralism to farming. Amongst the eight Teso donors in our dataset half have at least one Luhya (Bantu) parent or grandparent, and this is also true of one of the three Sabaot donors. Even among Bantu-speaking donors, one Luhya had Sabaot ancestors in both his maternal and paternal lineages but by cultural convention the individual's ethnic identity is that of his father. This reflects the historical pattern of women's flexibility in marriage including marrying into their husband's ethnic group. Female-mediated gene flow will tend to homogenise allele frequencies at autosomal loci but the lower frequency of male-mediated gene flow implied by the supplied ancestry data in this study would suggest that more marked differentiation will be seen with Y-STRs, and this will be explored in a forthcoming paper on Yfiler Plus profiling of the same dataset.

In our analysis we used the eight former provinces to subdivide the population rather than the current 47 administrative counties, in order

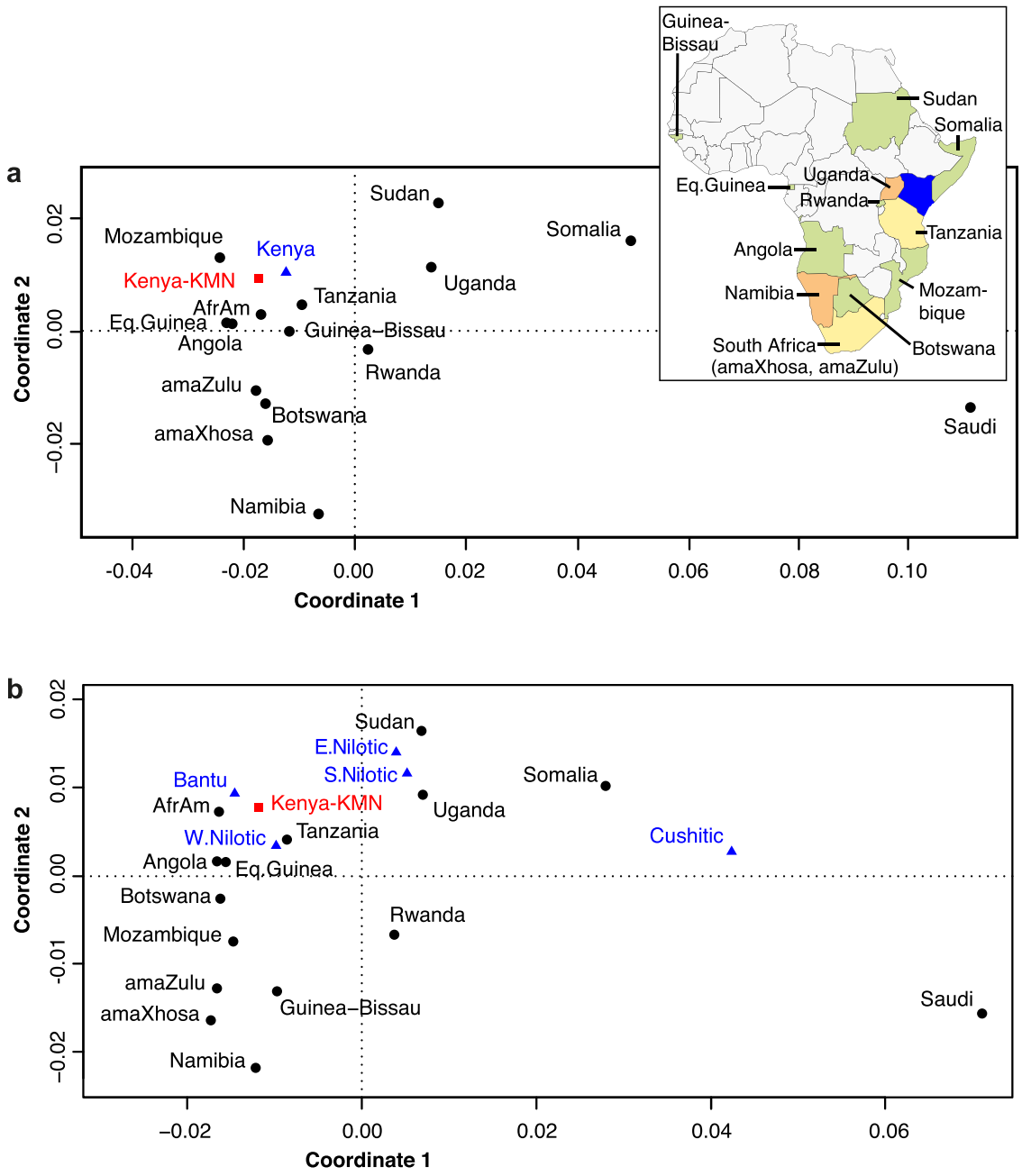


**Fig. 3.** Genetic relationships of Kenya and its linguistic groups with other populations in Africa and Arabia. Multidimensional scaling (MDS) plot based on pairwise $F_{ST}$ values derived from autosomal STR data at the Identifiler loci. Data from the present study are compared with previously published datasets including Kenya-KMN (red, N = 150 [9]) and other African populations, as well as a Saudi Arabian sample, as (a) the undivided Kenyan sample (blue triangle), and (b) the five Kenyan linguistic sub-divisions (Bantu, Cushitic, Western-, Southern- and Eastern-Nilotic – blue). The inset map shows locations of populations, with Kenya highlighted in blue and other countries highlighted in arbitrary colours. Comparative data sources are given in Section 2. AfrAm – African Americans; Eq. Guinea – Equatorial Guinea (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

to provide adequate sample sizes for comparison. The provincial boundaries tend to encompass groups of counties occupied by different ethnic groups within the same linguistic family and so the linguistic and geographic subsets here are closely correlated (see Fig. S1). Due to settlement patterns being determined by the patchy availability of suitable environments, smooth clinal variation in autosomal allele frequencies resulting from the migration of different groups into the Kenyan region would not be expected even though genetically diverse populations colonised the country by migration from different cardinal points [45]. However, a principal coordinates analysis (PCoA) plot (Fig. 2) reflects the relative geographic placement of the eight provinces, suggesting that genetic similarity does decay with distance apart.

Closely related ethnic groups often occur in neighbouring countries separated by political borders drawn during the colonial era with little regard to the distribution of the indigenous people. Consequently, data from population datasets collected outside of Kenya may be relevant to the assessment of allele frequencies within particular ethnic groups. This is particularly true where the majority of the ethnic group live outside of Kenya; the Somali are a good example of a minority group within Kenya who are most abundant in neighbouring Somalia but have only been extensively sampled through expatriate communities living in Europe. PCoA analysis (Fig. 2) revealed a predictable clustering of Kenyan linguistic groups with countries in which related ethnic groups occur. However, the Kenyan Cushitic sample is displaced beyond the Danish

dataset of expatriate Somalians towards a Saudi Arabian population. There are a number of possible reasons for this: it may suggest that the Cushitic sample in this study is not composed exclusively of Somalis, that there are Bantu Somalians among the Danish sample, and/or that Somalis themselves are characterised by strong clan structuring and genetic differentiation [52]. Granularity of ethnic distribution continues on a micro-geographic scale within Nairobi Province where the district of Eastleigh, just two kilometres from the centre of Nairobi, is overwhelmingly inhabited by expatriate Somalis [53], emphasising the need to consider relevant allele frequency databases and the choice of appropriate theta corrections in an unevenly distributed society.

Despite the diversity of the individuals sampled here, there were no truly novel allele lengths detected although many of the rarer variants had previously only been observed in an East African context. Indeed, when developing new multiplexes, the usual approach is to use African-American datasets for validation, and yet their ancestry is predominantly Niger-Kordofanian (~71%), the language family that includes Bantu, with European ancestry (~13%) exceeding that from other African groups (~8%) [6]. Consequently, surveys of East African populations can reveal novel findings. Three observations have significance for the interpretation of autosomal profiles: the presence of an Amelogenin Y null allele without drop-out of other Y markers, the presence of an additional TPOX allele which may be related to a translocation onto another chromosome (potentially the X [43]), and an extremely short SE33 allele that falls outside of the expected size range and could be mistaken for an off-ladder allele at another locus detected with the same dye. Beyond this, the application of GlobalFiler has demonstrated a significant improvement in discrimination power compared with Identifiler (the current multiplex used by Kenyan forensic laboratories). A comparison with the existing Identifiler population frequency database derived from small samples from three of the largest cities in Kenya shows close similarity in allele frequencies with slight deviations that might be linked to undersampling of both Cushitic and Southern- and Eastern-Nilotic speaking ethnic groups. It is hoped that the data from this study will support and improve the reporting of match probabilities within the Kenyan legal system and thus support the robust application of forensics in Kenya.

## Online resources

Interactive versions of the province and county maps (Fig. S1) are available at https://microreact.org/project/qACZKEyuTqviiDDkSTy3A4/c92f2238 and https://microreact.org/project/eBpydzMtpSz3RV3SiFNnRN/d011b03f respectively. Clicking on the pie charts, label and colour buttons reveals additional data and ways of visualising the dataset.

## Conflicts of interest

None.

## Acknowledgements

We thank all DNA donors, Becky Steffen (NIST) for primer sequences, and two anonymous reviewers for helpful comments. This work was supported by a grant to L.S. and M.A.J. from ELRHA (Enhancing Learning and Research for Humanitarian Assistance), a philanthropic donation from Ring for Peace, a University of Leicester International Research Development Fund award, and a GCRF Global Impact Accelerator Account award to the University of Leicester (EPSRC Grant Ref: EP/S515929/1).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102535.

## References

[1] C. Stringer, The origin and evolution of Homo sapiens, Philos. Trans. R. Soc. Lond. B Biol. Sci. 371 (2016), 20150237.
[2] Kenya National Bureau of Statistics, The 2009 Kenya Population and Housing Census; Volume II - Population and Household Distribution by Socio-Economic Characteristics, Nairobi, 2010.
[3] International HapMap Consortium, Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.
[4] D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M.O. Pollard, A. Choudhury, G.R. Ritchie, Y. Xue, J. Asimit, R.N. Nsubuga, E.H. Young, C. Pomilla, K. Kivinen, K. Rockett, A. Kamali, A.P. Doumatey, G. Asiki, J. Seeley, F. Sisay-Joof, M. Jallow, S. Tollman, E. Mekonnen, R. Ekong, T. Oljira, N. Bradman, K. Bojang, M. Ramsay, A. Adeyemo, E. Bekele, A. Motala, S.A. Norris, F. Pirie, P. Kaleebu, D. Kwiatkowski, C. Tyler-Smith, C. Rotimi, E. Zeggini, M.S. Sandhu, The African Genome Variation Project shapes medical genetics in Africa, Nature 517 (2015) 327–332.
[5] 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65.
[6] S.A. Tishkoff, F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A.A. Awomoyi, J.M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M. J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, S.M. Williams, The genetic structure and history of Africans and African Americans, Science 324 (2009) 1035–1044.
[7] S.H. Katsanis, J.K. Wagner, Characterization of the standard and recommended CODIS markers, J. Forensic Sci. 58 (2013) S169–S172.
[8] K. Bostoen, The Bantu expansion, Oxford research encyclopedia of African history, Oxford University Press, Oxford, 2018. http://africanhistoryoxfordrecom/view/101093/acrefore/9780190277734.001.0001/acrefore-9780190277734-e-191.
[9] K.J. Kimani, MSc Thesis: Kenya Population Forensic Data on a Sixteen Loci Microsatellite DNA System, Department of Biotechnology, Kenyatta University, 2009.
[10] M.E. D'Amato, M. Bodner, J.M. Butler, L. Gusmao, A. Linacre, W. Parson, P. M. Schneider, P. Vallone, A. Carracedo, Ethical publication of research on genetics and genomics of biological material: guidelines and recommendations, Forensic Sci. Int Genet 48 (2020), 102299.
[11] H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, Glottolog 4.2.1. Jena: Max Planck Institute for the Science of Human History, 2020. http://glottolog.org (Accessed 24 September 2020).
[12] T.E. King, S.J. Ballereau, K. Schürer, M.A. Jobling, Genetic signatures of coancestry within surnames, Curr. Biol. 16 (2006) 384–388.
[13] A. Carracedo, J.M. Butler, L. Gusmao, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int Genet 4 (2010) 145–147.
[14] A. Carracedo, J.M. Butler, L. Gusmao, A. Linacre, W. Parson, L. Roewer, P. M. Schneider, New guidelines for the publication of genetic population data, Forensic Sci. Int Genet 7 (2013) 217–220.
[15] L. Gusmao, J.M. Butler, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Sci. Int Genet 30 (2017) 160–163.
[16] P.M. Schneider, Scientific standards for studies in forensic genetics, Forensic Sci. Int 165 (2007) 238–243.
[17] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, 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.
[18] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Sci. Int Genet 5 (2011) 329–332.
[19] S.T. Kalinowski, A.P. Wagner, M.L. Taper, ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship, Mol. Ecol. Notes 6 (2006) 576–579.
[20] A. Gouy, M. Zieger, STRAF—a convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int Genet 30 (2017) 148–151.
[21] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564–567.
[22] R. Peakall, P.E. Smouse, GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and researc, Mol. Ecol. Notes 6 (2006) 288–295.
[23] N. Takezaki, M. Nei, K. Tamura, POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface, Mol. Biol. Evol. 27 (2010) 747–752.
[24] W.N. Venables, B.D. Ripley. Modern Applied Statistics with S, fourth ed., Springer, New York, 2002.
[25] T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, J.B. Smerick, B. Budowle, Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples, J. Forensic Sci. 46 (2001) 647–660.
[26] P.G. Ristow, K.W. Cloete, M.E. D'Amato, GlobalFiler((R)) express DNA amplification kit in South Africa: extracting the past from the present, Forensic Sci. Int. Genet. 24 (2016) 194–201.
[27] S. Guerreiro, T. Ribeiro, M.J. Porto, M.J. Carneiro de Sousa, P. Dario, Characterization of GlobalFiler loci in Angolan and Guinean populations inhabiting Southern Portugal, Int J. Leg. Med. 131 (2017) 365–368.

[28] T. Tau, A. Wally, T.P. Fanie, G.L. Ngono, S.W. Mpoloka, S. Davison, M.E. D'Amato, Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci, Sci. Rep. 7 (2017) 6768.
[29] C. Alves, L. Gusmao, A.M. Lopez-Parra, M. Soledad Mesa, A. Amorim, E. Arroyo-Pardo, STR allelic frequencies for an African population sample (Equatorial Guinea) using AmpFlSTR Identifiler and Powerplex 16 kits, Forensic Sci. Int 148 (2005) 239–242.
[30] C. Alves, L. Gusmao, A. Damasceno, B. Soares, A. Amorim, Contribution for an African autosomic STR database (AmpF/STR Identifiler and Powerplex 16 System) and a report on genotypic variations, Forensic Sci. Int 139 (2004) 201–205.
[31] T. Muro, J. Fujihara, S. Imamura, H. Nakamura, T. Yasuda, H. Takeshita, Allele frequencies for 15 STR loci in Ovambo population using AmpFlSTR Identifiler Kit, Leg. Med. 10 (2008) 157–159.
[32] M. Regueiro, J.C. Carril, M.L. Pontes, M.F. Pinheiro, J.R. Luis, B. Caeiro, Allele distribution of 15 PCR-based loci in the Rwanda Tutsi population by multiplex amplification and capillary electrophoresis, Forensic Sci. Int 143 (2004) 61–63.
[33] Y.M. Khubrani, J.H. Wetton, M.A. Jobling, Analysis of 21 autosomal STRs in Saudi Arabia reveals population structure and the influence of consanguinity, Forensic Sci. Int. Genet. 39 (2019) 97–102.
[34] C. Tomas, H.S. Mogensen, S.L. Friis, C. Hallenberg, M.C. Stene, N. Morling, Concordance study and population frequencies for 16 autosomal STRs analyzed with PowerPlex(R) ESI 17 and AmpFlSTR(R) NGM SElect in Somalis, Danes and Greenlanders, Forensic Sci. Int. Genet. 11 (2014) e18–e21.
[35] H.M. Babiker, C.M. Schlebusch, H.Y. Hassan, M. Jakobsson, Genetic variation and population structure of Sudanese populations as indicated by 15 Identifiler sequence-tagged repeat (STR) loci, Investig. Genet. 2 (2011) 12.
[36] W. Charoenchote, AmpFℓSTR® Identifiler™ STR Allele Frequencies and PowerPlex® Y-STR Haplotype Frequencies of the Meru Population of Northern Tanzania, California State University, Sacramento, 2008.
[37] V. Gomes, P. Sanchez-Diz, C. Alves, I. Gomes, A. Amorim, A. Carracedo, L. Gusmao, Population data defined by 15 autosomal STR loci in Karamoja population (Uganda) using AmpF/STR Identifiler kit, Forensic Sci. Int Genet 3 (2009) e55–e58.
[38] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (2001) 320–322.
[39] L. Devesse, L. Davenport, L. Borsuk, K. Gettings, G. Mason-Buck, P.M. Vallone, D. Syndercombe Court, D. Ballard, Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power, Forensic Sci. Int. Genet. 48 (2020), 102356.
[40] T.M. Clayton, S.M. Hill, L.A. Denton, S.K. Watson, A.J. Urquhart, Primer binding site mutations affecting the typing of STR loci contained within the AMPFlSTR SGM Plus kit, Forensic Sci. Int. 139 (2004) 255–259.
[41] S. Hering, R. Nixdorf, J. Edelmann, C. Thiede, J. Dreßler, Further sequence data of allelic variants at the STR locus ACTBP2 (SE33): Detection of a very short off ladder allele, Intern. Congr. Ser. 1288 (2006) 810–812.
[42] H.M. Alsafiah, A. Iyengar, S. Hadi, W.M. Alshlash, W. Goodwin, Sequence data of six unusual alleles at SE33 and D1S1656 STR Loci, Electrophoresis 39 (2018) 2471–2476.
[43] A.B. Lane, The nature of tri-allelic TPOX genotypes in African populations, Forensic Sci. Int Genet 2 (2008) 134–137.
[44] M.A. Jobling, I.C. Lo, D.J. Turner, G.R. Bowden, A.C. Lee, Y. Xue, D. Carvalho-Silva, M.E. Hurles, S.M. Adams, Y.M. Chang, T. Kraaijenbrink, J. Henke, G. Guanti, B. McKeown, R.A. van Oorschot, R.J. Mitchell, P. de Knijff, C. Tyler-Smith, E. J. Parkin, Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing *Amelogenin Y*, Hum. Mol. Genet. 16 (2007) 307–316.
[45] J.L. Newman, The Peopling of Africa: A Geographic Interpretation, Yale University Press, New Haven, CT, 1995.
[46] J.H. Greenberg, The languages of Africa, Int J. Am. Linguist. 29 (Pt 2) (1963).
[47] A.N. Tucker, M.A. Bryan, The Non-Bantu Languages of North-Eastern Africa. Handbook of African Languages Part 3, Oxford University Press, London, 1956.
[48] A. Barnard, J. Spencer, Encyclopedia of Social and Cultural Anthropology, Taylor & Francis, Abingdon, 1996.
[49] H.O. Ayot, A History of the Luo-Abasuba of Western Kenya, from A.D. 1760–1940 (Doctoral Dissertation), University of Nairobi, 1973.
[50] R.R. Sokal, E.M. Winkler, Spatial variation among Kenyan tribes and subtribes, Hum. Biol. 59 (1987) 147–164.
[51] E.M. Winkler, R.R. Sokal, A phenetic classification of Kenyan tribes and subtribes, Hum. Biol. 59 (1987) 121–145.
[52] G. Iacovacci, E. D'Atanasio, O. Marini, A. Coppa, D. Sellitto, B. Trombetta, A. Berti, F. Cruciani, Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries, Forensic Sci. Int. Genet. 27 (2017) 123–131.
[53] A. Lindley, Protracted displacement and remittances: the view from Eastleigh, Nairobi. New Issues in Refugee Research Working Paper No.143, UNHCR, Geneva, 2007.