Title: Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries

Author: <ce:author id="aut0005" author-id="S1872497316302484-b0ab6753abc11ea279333ea87f62d566"> Giuseppe Iacovacci<ce:author id="aut0010" author-id="S1872497316302484-8cf7ce6741690c7aac1fbb1f7c33134c"> Eugenia D'Atanasio<ce:author id="aut0015" author-id="S1872497316302484-8bdfda08868ec075b0c081883b796122"> Ornella Marini<ce:author id="aut0020" author-id="S1872497316302484-0b019feef4fcf073b2b2d547dd8a26f2"> Alfredo Coppa<ce:author id="aut0025" author-id="S1872497316302484-c0b62bb7ca354d7d63c2472d72283915"> Daniele Sellitto<ce:author id="aut0030" author-id="S1872497316302484-b3232d5747d6218279cd0f63f4f5bb2d"> Beniamino Trombetta<ce:author id="aut0035" author-id="S1872497316302484-706960cfaacb3bb6a4f25b94f9926f87"> Andrea Berti<ce:author id="aut0040" author-id="S1872497316302484-2779ff62cffdfab1a9f85f0d063668b6"> Fulvio Cruciani

PII: S1872-4973(16)30248-4
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2016.12.015
Reference: FSIGEN 1636

To appear in: *Forensic Science International: Genetics*

Received date: 6-9-2016
Revised date: 19-12-2016
Accepted date: 29-12-2016

Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries.

Giuseppe Iacovacci [a^], Eugenia D'Atanasio[b^], Ornella Marini[b^], Alfredo Coppa[c], Daniele Sellitto[d], Beniamino Trombetta[b], Andrea Berti[a], Fulvio Cruciani[b,d*]

[a] Carabinieri, Reparto Investigazioni Scientifiche di Roma, Sezione di Biologia, Rome, Italy

[b] Dipartimento di Biologia e Biotecnologie "Charles Darwin", Sapienza Università di Roma, Rome, Italy

[c] Dipartimento di Biologia Ambientale, Sapienza Università di Roma, Rome, Italy

[d] Istituto di Biologia e Patologia Molecolari, Consiglio Nazionale delle Ricerche, Rome, Italy

* Author for correspondence:

Fulvio Cruciani, Dipartimento di Biologia e Biotecnologie "C. Darwin", Sapienza Università di Roma, Rome, Italy, Tel: +39 06 49912826, Fax: +39 06 4456866; e-mail: fulvio.cruciani@uniroma1.it

^*These authors contributed equally to this work*

Highlights

1. Sequence information regarding microvariants at four Y-STR loci is reported.
2. Diversity at the palindromic DYF387S1 STR is shaped by non-allelic gene conversion.
3. An excess of diallelic patterns at RM Y-STRs from cell lines was observed.
4. Low intra- and high inter-population diversity values are observed.
5. Ten additional Y-STRs in Yfiler Plus *vs.* Yfiler resulted in a 15% DC increase

Abstract

By using the recently introduced 6-dye Yfiler® Plus multiplex, we analyzed 462 males belonging to 20 ethnic groups from four eastern African countries (Eritrea, Ethiopia, Djibouti and Kenya). Through a Y-STR sequence analysis, combined with 62 SNP-based haplogroup information, we were able to classify observed microvariant alleles at four Y-STR loci as either monophyletic (DYF387S1 and DYS458) or recurrent (DYS449 and DYS627). We found evidence of non-allelic gene conversion among paralogous STRs of the two-copy locus DYF387S1. Twenty-two diallelic and triallelic patterns observed at 13 different loci were found to be significantly over-represented ($p < 10^{-6}$) among profiles obtained from cell lines compared to those from blood and saliva. Most of the diallelic/triallelic patterns from cell lines involved recurrent mutations at rapidly mutating loci (RM Y-STRs) included in the multiplex ($p < 10^{-2}$). At haplotype level, intra-population diversity indices were found to be among the lowest so far reported for the Yfiler® Plus, while statistically significant differences among countries and ethnic groups were detected when considering haplotype frequencies alone (F_{ST}) or by using molecular distances among haplotypes (Φ_{ST}). The strong population subdivision observed is probably the consequence of the patrilineal social organization of most eastern African ethnic groups, and suggests caution in the use of country-based haplotype frequency distributions for forensic inferences in this region.

Keywords: Y-STR, Rapidly mutating Y-STR, Yfiler® Plus, Eastern African populations.

Introduction

Lack of crossing over, haploidy and male-specificity are the main factors that have shaped the peculiar genetic diversity of the male specific region of the human Y chromosome (MSY). Lack of crossing over implies that, barring recurrent mutations, alleles at different MSY polymorphic loci are usually co-inherited. In turn, male-specificity and haploidy determine a small effective population size, which leads to lower intra-population and higher inter-population genetic diversity compared to the autosomes. Both marker co-inheritance and low intra-population diversity make Y chromosome polymorphisms unattractive as an investigative tool for most forensic genetic purposes.

Nevertheless, multiallelic Y chromosome short tandem repeats (Y-STRs) are being used more and more in specific forensic investigations that require male lineage differentiation and could represent the only source of genetic information useful to obtain male specific profiles from unbalanced male-female mixtures [1].

Ballantyne et al. [2,3] reported the identification of thirteen rapidly mutating Y-STRs (RM Y-STRs), characterized by mutation rates higher than 1×10^{-2}. Because of high mutation rates, RM Y-STR haplotypes have been shown to distinguish approximately ¼ of father-son pairs and ½ of brother pairs [4,5]. Using RM Y-STRs, high levels of within population haplotype diversity and low-to-zero haplotype sharing between subjects belonging to different populations have also been reported [4].

Recently, seven RM Y-STRs (DYS449, DYS518, DYS570, DYS576, DYS627 and the two-copy system DYF387S1), along with other three standard Y-STRs, have been added to a commercially available 17 Y-STR multiplex (Yfiler®). The new 27 Y-STR multiplex, termed Yfiler® Plus [6], has, to date, been analyzed in several populations from Eurasia [7-12], but data for Africa are limited to a sample of Somalian immigrants in Denmark [8].

In the present study, we provide a report on the forensic analysis of MSY haplotypes obtained using the 27 Y-STR multiplex Yfiler® Plus in 20 populations from four eastern African countries. Using a stable phylogenetic framework based on 62 Y-SNPs, we classified the observed microvariant alleles and biallelic/triallelic patterns as monophyletic or recurrent and characterized the molecular structure of microvariant alleles through Sanger sequencing. We also observed low levels of intra-population genetic diversity and significant Y-STR haplotype frequency differences between populations and countries, which is possibly a consequence of the widespread patrilineal and patrilocal social organization of most ethnic groups in the region.

Materials and Methods

DNA samples

A total of 462 males belonging to 20 populations from 4 eastern African countries (Eritrea, Ethiopia, Djibouti and Kenya) were investigated for 27 Y-STRs and 62 single nucleotide polymorphisms (SNPs, Fig. 1). Sample sizes, geographic origin and linguistic affiliation for each population are reported in Supplementary Table 1. All the populations have been previously analyzed [13-17] for some of the SNPs used in this study. Some of the samples came from the Coriell Institute Repository (44 Maasai and 44 Luhya from Kenya) and from the National Laboratory for the Genetics of Israeli Populations (17 Ethiopian Jews). Differences in sample sizes for some populations between the present work and previous studies reflect subsequent unavailability of DNA. Appropriate informed consent was obtained from all participants. The research project was approved by the General Command of the Arma dei Carabinieri under the Ministry of Defense and by the “Sapienza Università di Roma” Ethical committee (document number 2755/15).

DNA extraction and quantification

Genomic DNA was extracted from blood (N = 192), saliva (N = 165) or lymphoblastoid cell lines (N = 105) using standard techniques. DNA quantification was performed using a Quantifiler® Trio Quantification Kit (ThermoFisher Scientific, Waltham, MA, USA).

Single nucleotide polymorphism genotyping

Overall, 62 SNPs of the MSY were analyzed in order to assign 462 Y chromosomes to specific binary haplogroups/paragroups. Most of the markers have been previously genotyped in haplogroup-specific population studies [13-17]. With the exception of V258, all the SNPs have been previously described [15,17-19]. V258 is an A to T transversion at position chrY: 2691574 (February 2009 assembly of the UCSC Genome Browser, http://genome.ucsc.edu/), which defines a sister clade of E-M200 within haplogroup E-M85. Most of the SNPs were genotyped according to a hierarchical approach based on the MSY phylogeny, using PCR amplified products and subsequent heteroduplex DHPLC analysis [18,20], RFLP analysis or Sanger sequencing.

Y-STR multiplex genotyping

Multiplex amplification of 27 Y-STRs was performed on an Applied Biosystems® GeneAmp® PCR System 9700 Silver Block Thermal Cycler (Thermo Fisher Scientific, Waltham, MA, USA) using the Yfiler® Plus PCR Amplification Kit (Yfiler® Plus, Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's protocol utilizing 1 ng of genomic DNA. Amplified DNAs were electrophoresed on a 3500 XL Genetic Analyzer and the fragment analysis was performed with a GeneMapper® IDX v.1.4

(Thermo Fisher Scientific, Waltham, MA, USA). The Authors followed ISFG recommendations for the analysis of the polymorphisms.

Haplotype data were submitted to the Y-chromosomal haplotype reference database (www.yhrd.org) (YHRD accession numbers YA004198-YA004207). The contributors successfully passed the quality control test.

Y-STR microvariant sequence analysis

Four loci that were found to carry micro-variant alleles (DYS449, DYS458, DYS627 and DYF387S1) were PCR-amplified and sequenced. PCR primers were designed on the basis of the MSY sequence reported in the UCSC Genome Browser web site (February 2009 assembly of the human genome; http://genome.ucsc.edu/) using Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) to obtain PCR products of 0.5-1.1 Kb in length (Supplementary Table 2). For the multilocus marker DYF387S1, which consists of two paralogous sequences located about 2.1 Mb apart, two different paralog-specific primers were used, which made it possible to amplify and sequence the two Y-STRs separately (Supplementary Table 2). Sequencing templates were obtained through PCR in a 50 μl reaction containing 20 ng of genomic DNA, 200 mM of each dNTP, 2.5 mM $MgCl_2$, 1 unit of Taq polymerase, and 10 pmol of each primer. A touchdown PCR program was used with an annealing temperature decreasing from 63 to 56 °C over 14 cycles, followed by 30 cycles with an annealing temperature of 56 °C. After DNA amplification, the PCR products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany).

Cycle sequencing was performed using the BigDye Terminator Cycle Sequencing Kit with Amplitaq DNA polymerase (Applied Biosystems, Foster City, CA, USA) and an internal or PCR primer. Cycle sequencing products were purified by ethanol precipitation and run on an ABI Prism 3730XL DNA sequencer (Applied Biosystems). Chromatograms

were aligned and analyzed for microvariant characterization using Sequencher 4.8 (Gene Codes Corporation, Ann Arbor, MI, USA).

Forensic and population genetic parameter estimation

The genetic diversity (GD) for each locus was calculated according to Nei and Tajima [21]. Haplotype frequencies were determined using the counting method. The haplotype diversity (HD) was calculated as $HD = n(1-\Sigma p_i^2)/(n-1)$, where n is the sample size and p_i the frequency of the i^{th} haplotype. The haplotype match probability (HMP) was estimated using the formula $HMP = \Sigma p_i^2$, whereas the discrimination capacity (DC) was calculated as the ratio between the number of different haplotypes and the total number of haplotypes in the dataset.

Analysis of molecular variance

Analysis of molecular variance (AMOVA, [22]) was performed using Arlequin, ver.3.5.1.3 [23]. Two hierarchical levels (individuals grouped into 20 populations and populations grouped into 4 countries) were considered. The analysis was performed either considering or not the molecular distances between Y-STR haplotypes (Φ-statistics and F-statistics, respectively). As a measure of molecular distance in the Φ-statistic analysis, we used the sum of the squared number of repeat difference between two Y-STR haplotypes [24]. Significance levels of both Φ- and F-statistics were obtained by comparing the actual values with the distribution of 10,000 values obtained by randomization. Both DYS385 and DYF387S1 multi-locus markers consist of two Y-STR loci which are coamplified using a single primer pair. For the purposes of our AMOVA analysis, these multi-locus markers were not considered, while null alleles and duplications were coded as missing data at the relevant locus in the input file. The number of complete repeats resulting from sequencing analysis were assigned to the microvariant alleles.

Network analysis

Phylogenetic relationships among haplotypes defined by the 27 STRs analyzed were depicted by means of median joining (MJ) networks [25], through use of the Network 4.0 program (Life Sciences and Engineering Technology Solutions Web site; [26]) with the epsilon value set to zero. When two peaks were detected in the multi-locus systems DYS385 and DYF387S1, the smaller allele and the larger allele were arbitrarily assigned to two different allelic series.

For all the statistical analyses, the alleles of the DYS389II locus were converted to the DYS389B nomenclature by subtracting the repeat number of the DYS389I locus from that of the DYS389II locus.

Results and Discussion

SNP-defined haplogroup distribution

Single nucleotide polymorphism genotyping was performed in order to obtain a stable low-resolution phylogenetic framework to be used for the interpretation of Y-STR diversity patterns. Overall, 62 Y-specific SNPs were selected for their ability to capture the main Y eastern African haplogroups. The SNPs analyzed identified 32 binary haplogroups or paragroups in the sample under study. The corresponding phylogenetic tree and the haplogroup frequency distribution in the four countries analyzed are shown in Fig. 2. Haplogroup frequencies for the 20 populations analyzed are reported in Supplementary Fig. 1. As to the phylogenetic structure within haplogroup E-Z827, it should be noted that there are incongruences between recently published trees [17, 27, 28] regarding the relative positions of markers V6, M293 and M34. After genotyping these markers and Z827

in a set of informative subjects, we confirmed that V6 defines a node within E-Z827 (as reported in [17]) rather than a sister clade of E-Z827 as suggested in [27,28].

Y-STR single-locus diversity

Genetic diversity values (GD) of the 25 Y-STR loci analyzed for each country and the whole eastern African sample are reported in Supplementary Table 3. Genotypes at the 25 loci under study were obtained for all but three subjects, which showed null alleles at three different loci (DYF387S1, DYS19, and DYS533). All the Y-STR loci analyzed were polymorphic in the four countries under study, with the exception of locus DYS391 in Djibouti. The highest GD values were observed at the two multi-copy loci DYS385 and DYF387S1 (Overall GD = 0.94 and 0.91, respectively). In the overall sample, when 23 single-copy loci were considered, RM Y-STRs showed significantly higher GD values (range: 0.79 – 0.86) than conventional Y-STRs (range: 0.27 – 0.83) (Mann-Withney U test, $p < 10^{-3}$). The number of different alleles (excluding diallelic patterns and microvariants) at single-copy loci ranged between 4 and 14, with a significant difference ($p < 10^{-3}$) between RM Y-STRs and conventional Y-STRs (ranges: 10 - 14 and 4 - 10 respectively). A significant positive correlation was observed when the mutation rate of single-copy loci was compared to both GD (Pearson's r = 0.58, $p = 0.003$) and number of alleles (Pearson's r = 0.76, $p < 10^{-4}$) (Fig. 3).

Molecular characterization of microvariant alleles and evidence of gene conversion at DYF387S1

Microvariant alleles were observed at loci DYS449 (4 subjects), DYS627 (5 subjects), DYS458 (23 subjects) and DYF387S1 (9 subjects) (Table 1). With the exception of DYS458, all these loci are complex RM Y-STRs which contain multiple Y-STR motifs within a single amplicon.

Four microvariants observed at DYS449 occurred on three different binary haplogroup backgrounds (A-M28, E-V32 and B-M109), each characterized by a different indel located outside the repeat motifs (Table 1). Microvariant alleles at this Y-STR have been previously observed and sequenced [29], but were characterized by a distinct molecular structure and were associated with other binary haplogroups (A-P97, K-M9, NO-M214, R-M198 and R-M269). The resequencing of DYS449 confirms the high sequence diversity of this locus due to the presence of different elements of variable length, and which can be particularly relevant for analyses based on new NGS-based forensic platforms. We also noted that, contrary to the observation of Mulero et al. [30], the allelic nomenclature for the DYS449 marker in the Yfiler® Plus kit seems to correspond to the number of variable (TTCT) repeats according to Ballantyne et al. [3] rather than to the number of variable (TTTC) repeats according to Redd et al. [31] (Table 1).

Microvariant alleles at the locus DYS627 also occurred in three different backgrounds (A-M13, A-M28 and E-V258). The DYS627.2 microvariants found on haplogroups A-M13 and A-M28 were due to the variation in the number of repeats of a short $(GA)_n$ dinucleotide repeat sequence upstream to the main variable (AAAG) tetranucleotide motif (Table 1). Variation in the number of dinucleotide repeats at this $(GA)_n$ repeat could explain, at least in part, the recurrent observation of DYS627.2 microvariants at this locus reported in previous studies [10, 12, 32, 33, 34]. The presence of multiple variable elements within this amplicon (as well as within DYS449) could lead to alleles equal in length but different in sequence, which cannot be distinguished with standard STR genotyping techniques.

Resequencing of five out of 23 subjects carrying different DYS458.2 microvariants (repeat number ranging from 16.2 to 20.2) revealed that the microvariant pattern was due to a dinucleotide insertion/deletion within the GAAA repeat motif of the Y-STR. Such a microvariant structure has been previously described [35] and associated predominantly

with haplogroup J-M267 [35, 36]. Consistently, in our dataset, all the DYS458.2 alleles were found on chromosomes carrying the M267 mutation and *vice-versa*.

The multi-locus marker DYF387S1 consists of two Y-STR loci about 2.1 Mb apart (chrY:25.93 Mb and chrY:28.03 Mb, February 2009 assembly of the UCSC Genome Browser, http://genome.ucsc.edu/) and located within two nearly-identical inverted arms of the MSY palindrome P1 [37]. In order to amplify and sequence the two loci separately, two different locus-specific primers were designed taking advantage of a T-G paralogous sequence variant (PSV) located at position chrY: 25930757 (T base on the left arm) and chrY: 28031622 (G base on the right arm), about 680 bp away from the STRs. In all cases, the DYF387S1.2 microvariant allele was found to be the consequence of a two base indel 5' to the $(AAAG)_n$ repeat, and was found on the Y-STR located on the left arm of the palindrome (Table 1). In a single sample from Kenya (D19-457), both DYF387S1 loci carried the two base indel (DYF387S1 = 39.2, 40.2). All the 9 chromosomes carrying microvariant allele/s at the DYF387S1 marker turned out to belong to haplogroup B-M182, while a single B-M182 subject from Kenya (D17-403, Supplementary Table 4) showed no microvariants. Based on this information and on the phylogenetic relationships among the chromosomes analyzed in the present study (Fig. 4), we hypothesize it is likely that the microvariant DYF387S1.2 arose only once on the T-bearing paralogous sequence, before the divergence of B-M150 and B-M112, but after the coalescence of B-M60 chromosomes. The two "unusual" B-M182 chromosomes, carrying either two microvariants (D19-457) or no microvariants (D17-403), probably experienced unidirectional opposite gene conversion events between DYF387S1 paralogs followed by a single-step mutation in one of the two STR loci. This is consistent with the observation that D19-457 and D17-403 display less difference in repeat length between DYF387S1 loci (genotype 39.2,40.2 and 33,34, respectively) than the other B-M182 chromosomes (difference in repeat number ranging from 3 to 10). Our findings support the hypothesis that non-allelic gene conversion has a

widespread effect on the dynamics and diversity of duplicated Y chromosome STRs contained within palindromes [37] such as the multicopy loci DYS385 and DYF387S1.

Excess of diallelic/triallelic genotypes observed for RM Y-STRs in cell lines

Overall, 22 diallelic/triallelic genotypes were observed at 13 different STRs (Table 2). In all cases, the two “alleles” (or the two closest alleles for triallelic patterns) differed for a single repeat unit, consistent with a model involving single step mutations at tetranucleotide repeats. A balanced diallelic pattern (peak height ratio 1.0 - 1.2) at DYS460 was observed in four subjects, all belonging to the uncommon eastern African haplogroup E-M41. These findings are consistent with a germline duplication-mutation model [38], with a single duplication event in the germline of an E-M41 common ancestor followed by DYS460 single-step mutation/s.

Diallelic/triallelic patterns at loci DYS576, DYS627 and DYS385 were also observed in multiple individuals, but in different haplogroup backgrounds (Table 2), suggesting multiple independent events of duplication and/or mutation [38, 39].

Interestingly, most of the diallelic/triallelic patterns (18 out of 22, or 17 out of 18 if we exclude the above-mentioned monophyletic germ-line duplication at DYS460) were observed in DNAs extracted from cell lines, which represent less than one quarter of the subjects in our dataset (105 out of 462 subjects). The observed excess of genotypes with diallelic patterns in DNA from cell lines was statistically significant (Fisher’s exact test, $p < 10^{-6}$), and was more pronounced for RM Y-STR loci (10 independent mutations observed in 630 single-locus genotypes) than for conventional Y-STR loci (7 mutations in 1995 genotypes) (Fisher’s exact test, $p < 10^{-2}$). The high number of observed diallelic genotypes may be the consequence of increased somatic STR mutability (with or without locus duplication), which is known to occur in cultured cell lines [40] and which one would expect

to be more pronounced for RM Y-STRs. The observation of several unbalanced diallelic patterns (Table 2, last columns) is consistent with the hypothesis of somatic mutations. Our observation is in line with the findings of [3], who also reported an increased number of multiallelic genotypes at RM Y-STRs in HGDP-CEPH cell lines compared to blood-derived DNA.

Low levels of intra-population Y-STR haplotype diversity

The list of 462 Y-STR haplotypes observed in eastern Africa is reported in Supplementary Table 4, along with the binary haplogroup affiliation for each sample. Forensic parameters (haplotype diversity, haplotype match probability and discrimination capacity) for the 20 populations analyzed in the present study are reported in Table 3. Overall, the analysis of 462 Y-chromosomes led to the identification of 410 different haplotypes: 374 were unique haplotypes, while the other were shared by two subjects (28 haplotypes), three subjects (5 haplotypes), four subjects (2 haplotypes), and 9 subjects (1 haplotype) (Supplementary Table 5). In all cases, haplotype sharing was limited to subjects from the same country carrying the same binary haplogroup. Furthermore, apart from two Ethiopians (Oromo and Amhara), the subjects who shared the same Y-STR haplotype were from the same ethnic group. Overall, these findings suggest that, when using Y-STRs from the Yfiler® Plus kit, chromosome sharing among subjects is (mainly) due to hidden recent paternal co-ancestry rather than identity by state. The relatively low number of different haplotypes observed mirrors the relatively low discrimination capacity (DC) figures for some of the populations under study (Table 3). These values ranged from 0.83 in Eritrea to 0.95 in Ethiopia (Table 3). At population level, the Cushitic–speaking Saho from Eritrea were characterized by the lowest level of haplotype diversity (HD = 0.986) and DC (DC = 0.73) yet reported. Some of the observed DC figures were in the lower range of DC values that have, to date, been reported for other populations analyzed with the same 27 Y-STRs multiplex [7-12, 33]. It is likely that these differences are due, at

least in part, to different sampling strategies between the present study, where most samples came from small rural areas, and previous studies, which seems to comprise mostly samples from cities (Bergamo, Salzburg), large regions (Northern, Central and southern Italy, Upper Austria, Henan in China) or whole countries (Denmark). However, Somalian immigrates in Denmark, analyzed in [8], which are expected to come from different areas of Somalia, also show low intra-population diversity, suggesting that low diversity could be a characteristic of eastern African ethnic groups. Interestingly, three different groups of Somali analyzed in the present study (Somali from Ethiopia, Djibouti and Kenya) also showed reduced haplotype diversity and/or average gene diversity (Table 3). Both the Saho and the Somali are known to be organized in strict patrilineal and patrilocal clans. It has been hypothesized that this kind of social structure can explain patterns of variability characterized by low Y-chromosome diversity within groups and large differences between groups [41]. In the present study, the impact of clan organization on the reduction of intra-population Y chromosome diversity can be fully appreciated in the Somali from Djibouti, the only population for which we have information regarding clan affiliation. The Somali ethnic group, which totals about 10 million people in Somalia, Ethiopia, Djibouti and Kenya, is organized in six major clan families (Darod, Dir, Isaaq, Hawiye, Digil and Rahanwein) [42]. In our sample of Somali from Djibouti, subjects belonging either to the Dir (N = 24) or Hawiye (N = 1) clan family share the T-M70 haplogroup, but are quite distinct regarding the Y-STR haplotype (Fig. 5), while all the subjects belonging to the Isaak clan (N = 9) share the haplogroup E-V32. Similar differences in haplogroup distribution among patrilineal clans belonging to other African ethnic groups have been previously reported [43,44]. Conversely, two Nilo-Saharan populations included in our survey (Nara and Cunama from Eritrea), which represent rare examples of eastern African matrilineal groups [45], were characterized by high DC and

HD, and presented the highest average gene diversity across loci among the populations surveyed (Table 3).

High level of inter-population Y-STR haplotype diversity

We observed statistically significant pair-wise differences in haplotype frequencies among all the countries analyzed, both considering molecular differences among haplotypes (Φ_{ST} range: 0.075 - 0.373, Supplementary Table 6) or not (F_{ST} range: 0.002 – 0.007, Supplementary Table 7). Significant differences ($p < 0.01$) were also observed within each country, when the ethnic affiliation was considered (Φ_{ST} ranging from 0.078 in Ethiopia to 0.438 in Eritrea; F_{ST} range: 0.009 – 0.013, Table 4). Population pairwise Φ_{ST} and F_{ST} values are reported in Supplementary Tables 8 and 9, respectively. When molecular differences between haplotypes were taken into account, most of the pair-wise population comparisons (116/190) were found to be statistically significant. Not surprisingly, the highest Φ_{ST} values were observed for pair-wise comparisons involving the ethnic groups characterized by the lowest intra-population diversity parameters, i.e. the Saho from Eritrea and the Somali from Djibouti (pairwise Φ_{ST} values range 0.547 – 0.713 and 0.074 – 0.713, respectively, Supplementary Table 8). Both these groups are characterized by high frequencies of binary haplogroups which are relatively uncommon in other African populations (Supplementary Fig. 1, [13, 15, 17, 18, 46]), such as haplogroup E-V22 in the Saho (88%) and T-M70 in the Somali from Djibouti (74% in the population, 100% in the Dir clan), with Y-STRs presenting a low degree of within-population molecular differentiation (Supplementary Fig. 2B and 3). Overall, the observed high level of inter-population diversity highlights the fact it is important to take into account ethnic composition and, possibly, the clan structure of the populations under study when making forensic and evolutionary inferences.

Comparison of the discrimination capacity of Yfiler® and Yfiler® Plus multiplexes

Overall, the addition of ten new Y-STRs in the Yfiler® Plus multiplex resulted in a 15% increase in the power of discrimination compared to the 17 Y-STR Yfiler® multiplex (DC = 0.887 and 0.771, respectively, Table 3 and Supplementary Table 10). The DC increase was particularly pronounced for the Somali from Djibouti (0.941 *vs.* 0.471, 100%), where the additional loci allowed the almost complete resolution in singletons of four "common" 17-loci haplotypes shared by 3, 4, 6 and 9 subjects (Supplementary Table 5 and Supplementary Fig. 2). The increase in the power of discrimination was also pronounced for the Saho from Eritrea (41%), although, as mentioned above, it remained the lowest (DC = 0.734) so far reported for the Yfiler® Plus multiplex. Thus, although the 27-loci Yfiler® Plus was comparatively very effective in increasing the power of discrimination of its 17-loci predecessor, it is still not powerful enough to adequately distinguish individuals from populations characterized by a strong patrilineal structure such as some of the eastern African groups here analyzed.

Acknowledgements

We are grateful to all donors for contributing samples. Some of the samples were provided by the Coriell Institute Repository (44 Maasai and 44 Luhya from Kenya) and the National Laboratory for the Genetics of Israeli Populations at Tel-Aviv University (17 Ethiopian Jews). This work was supported by the Sapienza University of Rome (grant number C26A153PCN to FC).

References

[1] L. Roewer, Y chromosome STR typing in crime casework, Forensic Sci. Med. Pathol. 5 (2009) 77–84.

[2] K.N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, et al., Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications, Am. J. Hum. Genet. 87 (2010) 341–353.

[3] K.N. Ballantyne, V. Keerl, A. Wollstein, Y. Choi, S.B. Zuniga, A. Ralf, et al., A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages, Forensic Sci. Int. Genet. 6 (2012) 208–218.

[4] K.N. Ballantyne, A. Ralf, R. Aboukhalid, N.M. Achakzai, M.J. Anjos, Q. Ayub, et al., Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats, Hum. Mutat. 35 (2014) 1021-1032.

[5] A. Adnan, A. Ralf, A. Rakha, N. Kousouri, M. Kayser, Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan, Forensic Sci. Int. Genet. In press.

[6] S. Gopinath, C. Zhong, V. Nguyen, J. Ge, R.E. Lagacé, M.L. Short et al. Developmental validation of the Yfiler® Plus PCR amplification kit: an enhanced Y-STR multiplex for casework and database applications, Forensic Sci. Int. Genet. 24 (2016) 164-175.

[7] O. García, I. Yurrebaso, I.D. Mancisidor, S. López, S. Alonso, L. Gusmão, Data for 27 Y-chromosome STR loci in the Basque Country autochthonous population, Forensic Sci. Int. Genetic. 20 (2015) e10-e12.

[8] J.K. Olofsson, H.S. Mogensen, A. Buchard, C. Borsting, N. Morling, Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit, Forensic Sci. Int. Genetic. 16 (2015) 232–236.

[9] E. Ottaviani, S. Vernarecci, P. Asili, A. Agostino, P. Montagna, Preliminary assessment of the prototype Yfiler® Plus kit in a population study of Northern Italian males, Int. J. Legal Med. 129 (2015) 729–730.

[10] I. Pickrahn, E. Müller, W. Zahrer, B. Dunkelmann, J. Cemper-Kiesslich, G. Kreindl, et al., Yfiler® Plus amplification kit validation and calculation of forensic parameters for two Austrian populations, Forensic Sci. Int. Genetic. 21 (2016) 90-94.

[11] C. Rapone, E. D'Atanasio, A. Agostino, M. Mariano, M.T. Papaluca, F. Cruciani, et al., Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler® Plus kit) in an Italian population sample, Forensic Sci. Int. Genetic. 21 (2016) e1-e5.

[12] L. Wang, F. Chen, B. Kang, H. Zheng, Y. Zhao, L. Li, et al., Genetic population data of Yfiler Plus kit from 1434 unrelated Hans in Henan Province (Central China), Forensic Sci. Int. Genetic. 22 (2016) e25-e27.

[13] F. Cruciani, P. Santolamazza, P. Shen, V. Macaulay, P. Moral, A. Olckers, et al., A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes, Am. J. Hum. Genet. 70 (2002) 1197-1214.

[14] F. Cruciani, R. La Fratta, P. Santolamazza, D. Sellitto, R. Pascone, P. Moral, et al., Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa, Am. J. Hum. Genet. 74 (2004) 1014-1022.

[15] F. Cruciani, R. La Fratta, B. Trombetta, P. Santolamazza, D. Sellitto, E.Beraud Colomb, et al., Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12, Mol. Biol. Evol. 24 (2007) 1300-1311.

[16] F. Cruciani, B. Trombetta, A. Massaia, G. Destro-Bisol, D. Sellitto, R. Scozzari, A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa, Am. J. Hum. Genet. 88 (2011) 814-818.

[17] B. Trombetta, E. D'Atanasio, A. Massaia, M. Ippoliti, A. Coppa, F. Candilio, et al., Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the African continent, Genome Biol. Evol. 7 (2015) 1940-50.

[18] P.A. Underhill, P. Shen, A.A. Lin, L. Jin, G. Passarino, W.H.Yang, et al., Y chromosome sequence variation and the history of human populations, Nat. Genet. 3 (2000) 358-361.

[19] T.M. Karafet, F.L. Mendez, M.B. Meilerman, P.A. Underhill, S.L. Zegura, M.F. Hammer, New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree, Genome Res. 5 (2008) 830-838.

[20] P.A. Underhill, L. Jin, A. Lin, S.Q. Mehdi, T. Jenkins, D. Vollrath et al., Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography, Genome Res. 7 (1997) 996-1005.

[21] M. Nei, F. Tajima, DNA polymorphism detectable by restriction endonucleases, Genetics 97 (1981) 145–163.

[22] L. Excoffier, P.E. Smouse, J.M. Quattro, Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data, Genetics 131 (1992) 479-491.

[23] 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. Res. 10 (2010) 564–567.
[24] M. Slatkin, A measure of population subdivision based on microsatellite allele frequencies, Genetics 139 (1995) 457-462.
[25] H.J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1999) 37-48.
[26] H.J. Bandelt, Y.G. Yao, C.M. Bravi, A. Salas, T. Kivisild, Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies, Am. J. Hum. Genet. 54 (2009) 174-181.
[27] M. van Oven, A. Van Geystelen, M. Kayser, R. Decorte, M.H. Larmuseau, Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome, Hum. Mutat. 35 (2014) 187-191.
[28] M. Karmin, L. Saag, M. Vicente, M.A Wilson Sayres, M. Järve, U.G. Talas, et al., A recent bottleneck of Y chromosome diversity coincides with a global change in culture, Genome Res. 25 (2015) 459-466.
[29] N.M. Myres, K.H. Ritchie, A.A. Lin, R.H. Hughes, S.R. Woodward, P.A. Underhill, Y-chromosome short tandem repeat intermediate variant alleles DYS392.2, DYS449.2, and DYS385.2 delineate new phylogenetic substructure in human Y-chromosome haplogroup tree, Croat. Med. J. 50 (2009) 239-249.
[30] J. Mulero, J. Ballantyne, K. Ballantyne, B. Budowle, M. Coble, L. Gusmão, et al., Nomenclature update and allele repeat structure for the markers DYS518 and DYS449, Forensic Sci. Int. Genetic. 13 (2014) e3.
[31] A.J. Redd, A.B. Agellon, V.A. Kearney, V.A. Contreras, T. Karafet, H. Park, et al., Forensic value of 14 novel STRs on the human Y chromosome, Forensic Sci. Int. Genet. 130 (2002) 97-111.
[32] C. Robino, A. Ralf, S. Pasino , M.R. De Marchi, K.N. Ballantyne, A. Barbaro, et al., Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI collaborative exercise, Forensic Sci. Int. Genetic. 15 (2015) 56-63.
[33] R. Bai, Y. Liu, J. Zhang, M. Shi, H. Dong, S. Ma, et al., Analysis of 27 Y-chromosomal STR haplotypes in a Han population of Henan province, Central China, Int. J. Legal Med. 130 (2016) 1191-1194.
[34] Y. Wang, Y.J. Zhang, C.C. Zhang, R. Li, Y. Yang, X.L. Ou, et al., Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong Province, Southern China, Forensic Sci. Int. Genet. 21 (2016) 5-9.
[35] N.M. Myres, J.E. Ekins, A.A. Lin, L.L. Cavalli-Sforza, S.R. Woodward, P. Underhill, Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405, Croat. Med. J. 48 (2007) 450-459.
[36] S. Tofanelli, G. Ferri, K. Bulayeva, L. Caciagli, V. Onofri, L. Taglioli, et al., J1-M267 Y lineage marks climate-driven pre-historical human displacements, Eur J. Hum Genet. 17 (2009) 1520-1524.
[37] P. Balaresque, T.E King, E.J Parki, E. Heyer, D. Carvalho-Silva, T. Kraaijenbrink, et al., Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats, Hum. Mutat. 35 (2014) 609-617.
[38] P. Balaresque, G.R. Bowden, E.J. Parkin, G.A. Omran, E. Heyer, L. Quintana-Murci, et al., Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis, Hum. Mutat. 29 (2008) 1171-1180.
[39] B. Rolf, P. Wiegand, B. Brinkmann, Somatic mutations at STR loci - a reason for three-allele pattern and mosaicism, Forensic Sci. Int. 126 (2002) 200-202.
[40] J.L. Weber, C. Wong, Mutation of human short tandem repeats, Hum. Mol. Genet. 8 (1993) 1123-1128.
[41] H. Oota, W. Settheetham-Ishida, D. Tiwawech, T. Ishida, M. Stoneking, Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence, Nat Genet. 29 (2001) 20-21.
[42] J. Gundel (2009). Clans in Somalia. Vienna: Austrian Red Cross. http://www.ecoi.net/file_upload/90_1261130976_accord-report-clans-in-somalia-revised-edition-20091215.pdf.
[43] C. Ottoni, M.H.D. Larmuseau, N. Vanderheyden, C. Martínez-Labarga, G. Primativo, G. Biondi, et al., Deep into the roots of the Libyan Tuareg: A genetic survey of their paternal heritage, Am. J. Phys. Anthropol. 145 (2011) 118-124.
[44] H. Sanchez-Faddeev, J. Pijpe, T. van der Hulle, H.J. Meij, K.J. van der Gaag, P.E. Slagboom et al., The influence of clan structure on the genetic variation in a single Ghanaian village, Eur. J. Hum. Genet. 21 (2013) 1134–1139.
[45] L. Favali, R. Pateman, Blood, Land, and Sex: legal and political pluralism in Eritrea, Indiana University Press (2003).
[46] E.T. Wood, D.A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, et al., Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur. J. Hum. Genet. 13 (2005) 867–876.

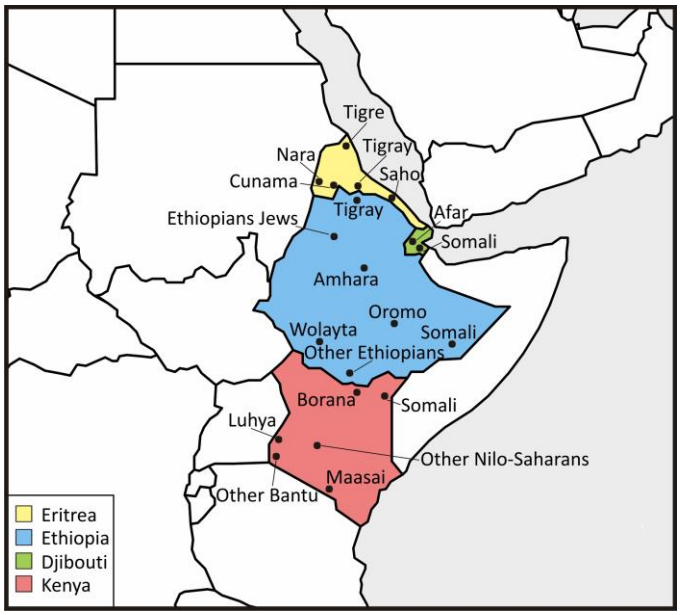
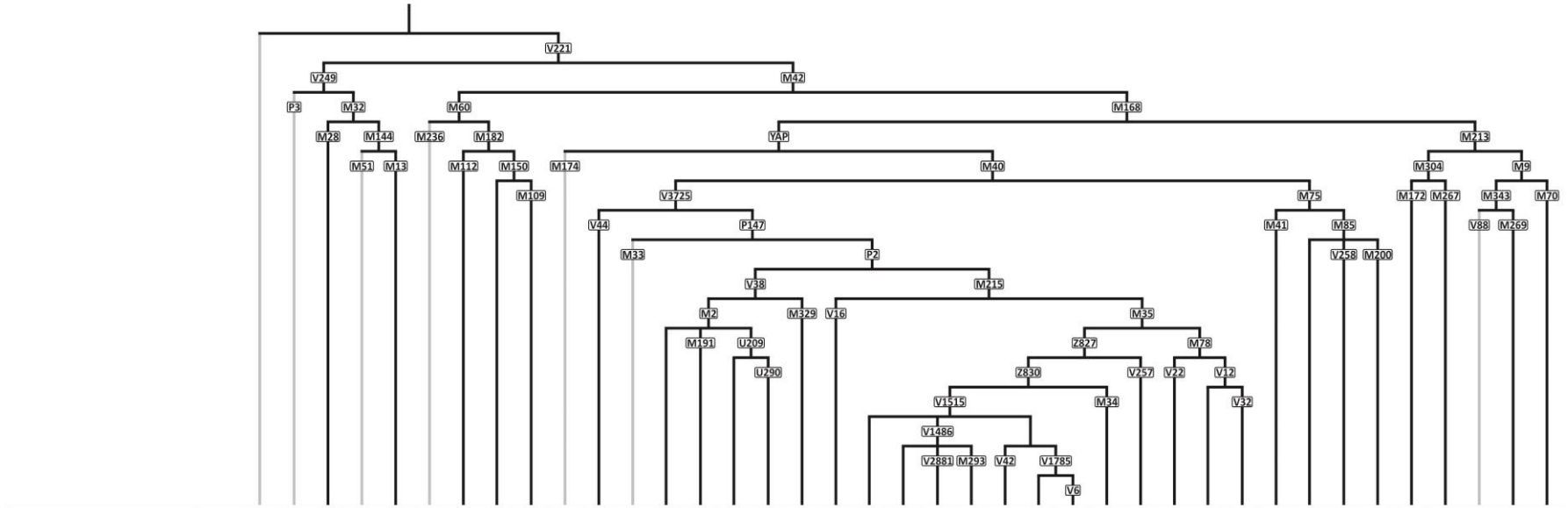


Figure 1: Geographic location of the eastern African populations analyzed. Populations are coded as reported in Supplementary Table 1.



Country	N	Y*(xV221)	A-P3	A-M28	A-M51	A-M13	B-M236	B-M112	B-M150*	B-M109	D-M174	E-V44	E-M33	E-M2*	E-M191	E-U209*	E-U290	E-M329	E-V16	E-V1515*	E-V1486*	E-V2881	E-M293	E-V42	E-V1785*	E-V6	E-M34	E-V257	E-V22	E-V12*	E-V32	E-M41	E-M85*	E-V258	E-M200	J-M172	J-M267	R-V88	R-M269	T-M70
Eritrea	161	-	-	3	-	6	-	-	3	-	-	-	-	-	-	-	-	1	-	3	2	1	-	5	12	13	2	-	85	1	14	2	-	-	-	-	7	-	-	1
Ethiopia	119	-	-	1	-	21	-	-	-	-	-	2	-			-	-	10	3	-	-	9	1	2	-	8	15	-	5	-	17	-	-	-	-	2	14	-	1	8
Djibouti	54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	9	-	-	-	-	2	-	-	30
Kenya	128	-	-	-	-	10	-	6	-	1	-	-	-	1	43	8	7	3	-	-	-	1	15	-	-	12	-	1	5	-	12	4	1	2	6	-	-	-	-	-
Total	462	-	-	4	-	37	-	6	3	1	-	2	-	1	43	8	7	14	3	3	2	11	16	7	12	35	18	1	95	1	52	6	1	2	6	2	23	-	1	39

Figure 2: MSY phylogenetic tree and SNP-defined haplogroup distribution in four eastern African countries. Haplogroups not observed in our dataset are represented in grey.

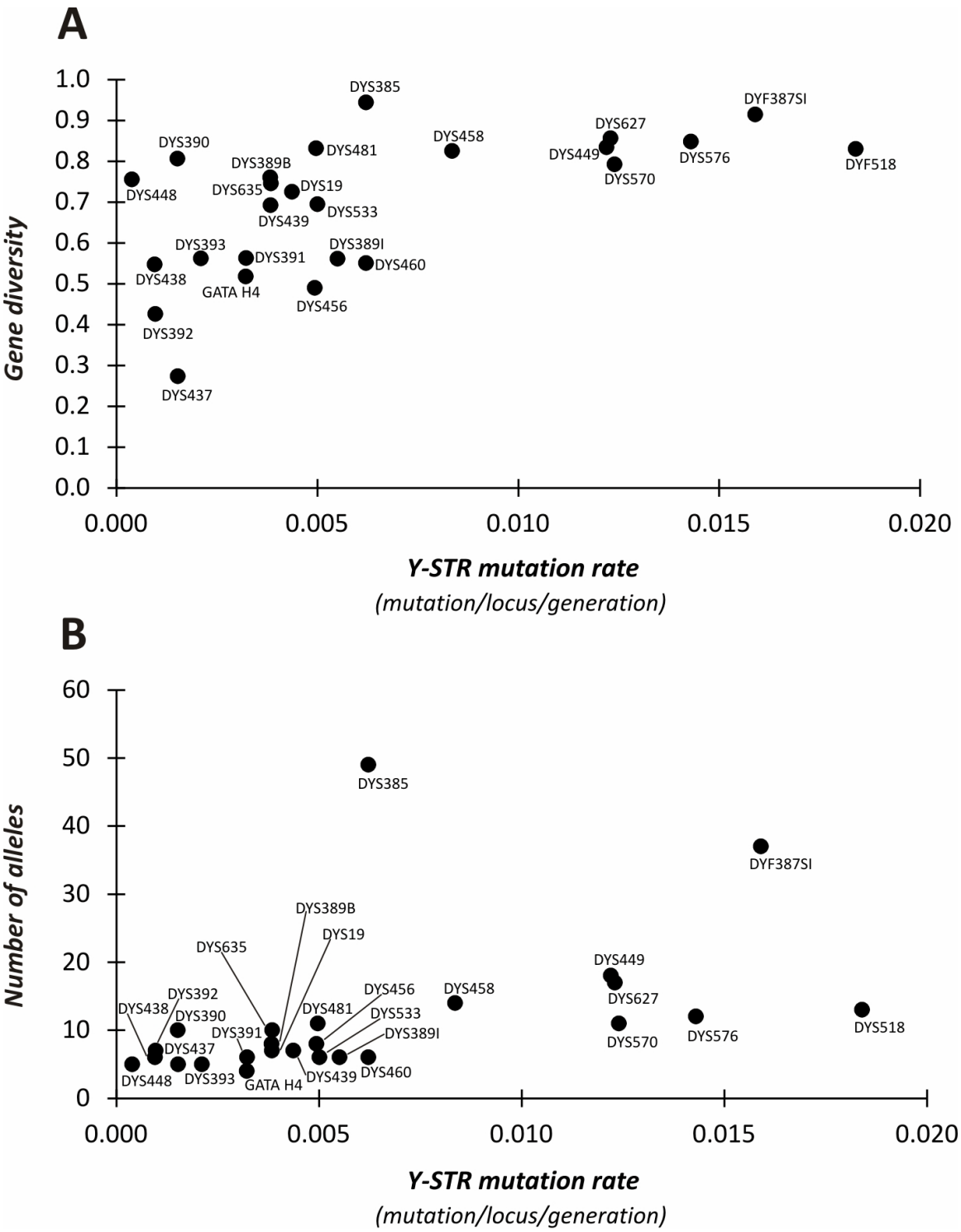
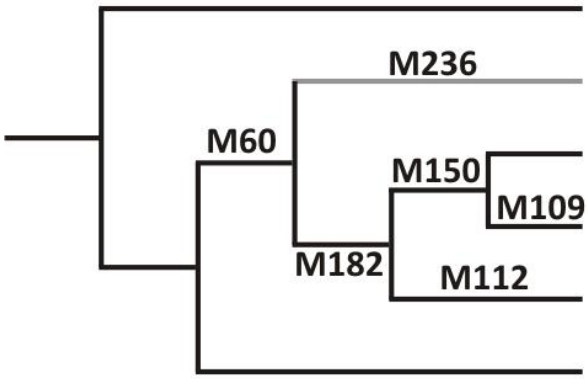


Figure 3: Correlation between Y-STR mutation rate (x-axis) and: (A) gene diversity, (B) number of alleles. Note that DYS385 and DYF387S1 are multi-copy loci.



Y haplogroup	DYF387S1.2 interallele	consensus allele
Y* (xBT)	–	41
B-M236	–	1
B-M150*(xM109)	3	–
B-M109	1	–
B-M112	6	1
CT	–	411

Figure 4: SNP-based phylogenetic relationships among chromosomes carrying the consensus or the microvariant allele at the DYF387S1 locus. The sample harboring the B-M236 haplogroup is from western Africa (in grey).

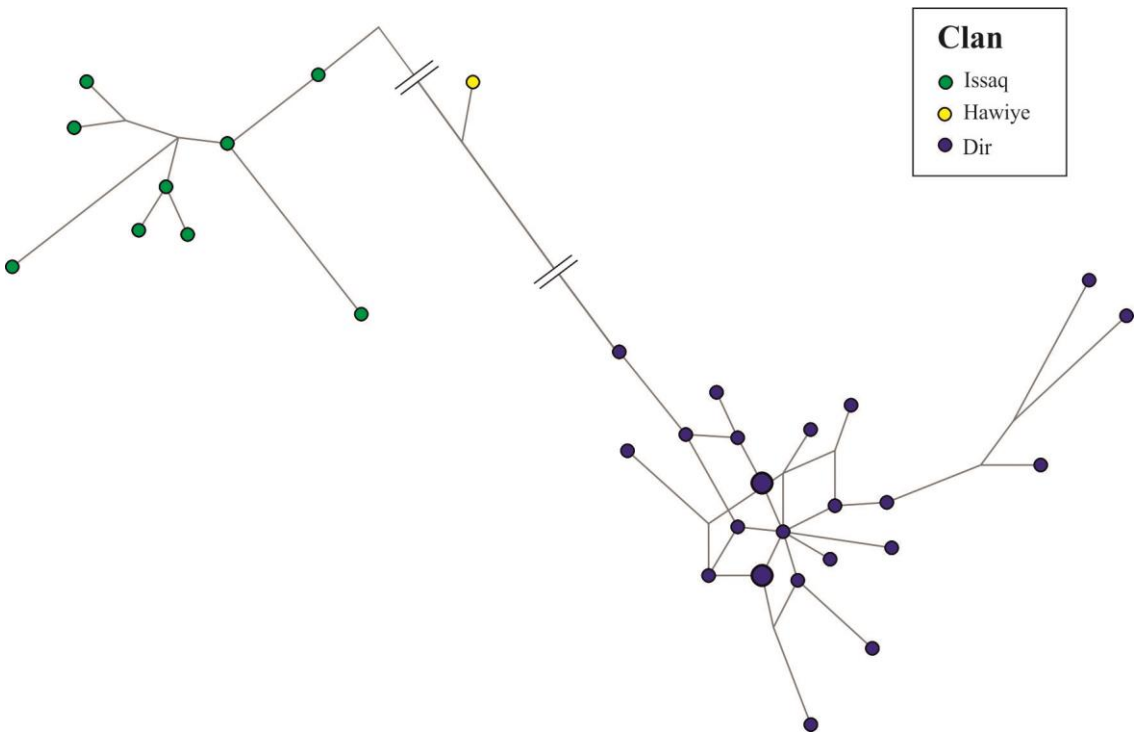


Figure 5: Network of Y-STR haplotypes in the Somali from Djibouti. Haplotypes are represented by circles, with areas proportional to the number of subjects harboring the haplotype. Branch lengths are proportional to the number of one-repeat mutations separating two haplotypes. Haplotypes are colour-coded on the basis of clan affiliation.

Table 1: Sequence structure of microvariant alleles at four Y-STRs

Y-STR	Sample code	Genotype	Haplogroup	Structure
DYS449	A1-007	31.1	A-M28	(TTCT)16 CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTT--- (TTCT)16
DYS449	A2-025	31.1	A-M28	(TTCT)16 CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTT--- (TTCT)16
DYS449	B8-232	32.1	E-V32	(TTCT)15 CTCTCTCCTC--- N9 (TTCT)3 TTTCCTCTTTCC (TTCT)18
DYS449	D18-443	33.2	B-M109	(TTCT)15 TT CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTTTCC (TTCT)18
DYS449	Control 1	33	A-M28	(TTCT)15 CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTTTCC (TTCT)18
DYS449	Control 2	32	R-M269	(TTCT)17 CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTTTCC (TTCT)15
DYS449	Reference	30		(TTCT)15 CTCTCTCCTCCTC N9 (TTCT)3 TTTCCTCTTTCC (TTCT)15
DYS627	A1-003	21.2	A-M13	(AGAA)3 N16 (AG)5 (AAAG)19 AAAAA GA GAAA N71 (AAGG)3
DYS627	A1-006	22.2	A-M13	(AGAA)3 N16 (AG)5 (AAAG)20 AAAAA GA GAAA N71 (AAGG)3
DYS627	A2-031	20.2	A-M28	(AGAA)2 AGAG N16 (AG)9 (AAAG)16 AAAAA GA GAAA N71 (AAGG)3
DYS627	D17-427	22.2	E-V258	(AGAA)3 N16 (AG)6 (AAAG)20 AAAAA -- GAAA N71 (AAGG)3
DYS627	D17-433	22.2	E-V258	(AGAA)3 N16 (AG)6 (AAAG)20 AAAAA -- GAAA N71 (AAGG)3
DYS627	Control 1	22	R-M269	(AGAA)3 N16 (AG)6 (AAAG)19 AAAAA GA GAAA N71 (AAGG)3
DYS627	Control 2	20	A-V148	(AGAA)1 N16 (AG)6 (AAAG)19 AAAAA GA GAAA N71 (AAGG)3
DYS627	Reference	21		(AGAA)3 N16 (AG)6 (AAAG)18 AAAAA GA GAAA N71 (AAGG)3
DYF387 (25.9 MB)	D18-443	38,41.2	B-M109	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)11 AA (AAAG)17
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)11 (AAAG)14
DYF387 (25.9 MB)	D19-457	39.2,40.2	B-M112	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 AA (AAAG)16
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)11 AA (AAAG)16
DYF387 (25.9 MB)	D16-378	35.2,38	B-M112	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 AA (AAAG)13
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 (AAAG)15
DYF387 (25.9 MB)	D17-394	33,43.2	B-M112	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)12 AA (AAAG)18
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 (AAAG)11
DYF387 (25.9 MB)	D17-396	36,40.2	B-M112	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 AA (AAAG)17
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)8 (AAAG)15
DYF387 (25.9 MB)	D17-409	36,40.2	B-M112	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 AA (AAAG)17
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)8 (AAAG)15
DYF387 (25.9 MB)	A1-001	37,41.2	B-M150*	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 AA (AAAG)19

DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)12 (AAAG)12
DYF387 (25.9 MB)	A1-004	37,41.2	B-M150*	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 AA (AAAG)19
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)12 (AAAG)12
DYF387 (25.9 MB)	A1-009	37,41.2	B-M150*	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 AA (AAAG)19
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)12 (AAAG)12
DYF387 (25.9 MB)	Control 1	35,36	R-M269	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 (AAAG)13
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 (AAAG)13
DYF387 (25.9 MB)	Control 2	35,37	B-M236	(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 (AAAG)13
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N16 (GAAG)10 (AAAG)14
DYF387 (25.9 MB)	Reference	35,36		(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)9 (AAAG)13
DYF387 (28.0 MB)				(AAAG)3 (GTAG)1 (GAAG)4 N20 (GAAG)10 (AAAG)13
DYS458	5 selected samples[a]	16.2 to 20.2	J-M267	(GAAA)14-18 AA (GAAA)2 GGAGGG
DYS458	Control	19	R-M269	(GAAA)19 GGAGGG
DYS458	Reference	19		(GAAA)16 GGAGGG

[a]Only five out of 23 observed DYS458.2 microvariants have been sequenced

Table 2: List of samples showing diallelic or triallelic patterns at Y-STRs

Sample code	DNA Source	Haplogroup	Locus[a]	Alleles	RFU	Peak ratio[b]
D18-446	Blood	E-M191	DYS385	15,16,18	150;233;256	1.7
D18-447	Blood	E-M41	DYS460	10,11	4511;4818	1.1
D19-451	Blood	E-M41	DYS460	10,11	744;877	1.2
A4-140	Saliva	E-M41	DYS460	10,11	7843;7569	1.0
D17-408	Cell line	E-M41	DYS460	11,12	4142;3365	1.2
D17-408	Cell line	E-M41	DYS389II	31,32	743;1373	1.8
D16-362[c]	Cell line	E-M191	DYS627	18,19	1385;1902	1.4
D16-365[d]	Cell line	E-M191	DYS576	17,18	5580;14023	2.5
D16-371	Cell line	E-U209	DYS627	20,21	845;918	1.1
D16-378	Cell line	B-M112	DYS390	21,22	796;830	1.0
D17-399	Cell line	E-M293	DYS439	11,12	1554;2007	1.3
D17-437	Cell line	E-V22	DYS635	21,22	2191;880	2.5
D17-415	Cell line	E-U209	DYS438	11,12	303;165	1.8
D17-419	Cell line	E-U290	DYS576	15,16	4933;3611	1.4
D17-419	Cell line	E-U290	DYS518	38,39	683;803	1.2
D17-424	Cell line	A-M13	DYS449	35,36	1998;1188	1.7
D17-428	Cell line	E-V32	DYS385	12,17,18	1254;927;395	3.2
D17-428	Cell line	E-V32	DYS481	22,23	950;820	1.2
D16-383[d]	Cell line	E-M191	DYS576	15,16	8610;6518	1.3
D17-432[c]	Cell line	E-M191	DYS627	18,19	893;1399	1.6
B7-214	Cell line	E-M329	DYF387S1	35,37,38	2212;2798;2905	1.3
B7-213	Cell line	A-M13	DYS576	17,18	1598;1928	1.2

[a] RM Y-STRs are underlined.

[b] Major-to-minor allele ratio.

[c] Although found in the same Y-STR (DYS627) and in the same haplogroup (E-M191), mutations at samples D16-362 and D17-432 are probably independent because they are found on two haplotypes showing a large number of STRs differing in size (see Supplementary Table 4).

[d] Although found in the same Y-STR (DYS576) and in the same haplogroup (E-M191), mutations at samples D16-365 and D16-383 are probably independent because they are found on two haplotypes showing a large number of STRs differing in size (see Supplementary Table 4).

Table 3: Forensic parameters for 20 eastern African populations analyzed in this study using Yfiler® Plus

Country	Ethnic group	Sample size	Number of haplotypes	Haplotype diversity	Haplotype match probability	Discrimination capacity	Mean number of pairwise differences	Average gene diversity over loci
Eritrea		161	133	0.9950	0.0111	0.8261	14.6	0.5839
	Cunama	19	18	0.9942	0.0582	0.9474	18.1	0.7230
	Nara	15	15	1.0000	0.0667	1.0000	16.1	0.6816
	Saho	94	69	0.9860	0.0244	0.7340	8.0	0.3188
	Tigray	28	27	0.9974	0.0383	0.9643	16.2	0.6470
	Tigre	5	4	0.9000	0.2800	0.8000	12.7	0.5080
Ethiopia		119	113	0.9990	0.0094	0.9496	17.1	0.6853
	Amhara	34	34	1.0000	0.0294	1.0000	16.8	0.6741
	Ethiopians Jews	22	19	0.9870	0.0579	0.8636	16.9	0.6764
	Oromo	28	28	1.0000	0.0357	1.0000	16.4	0.6578
	Other Ethiopians	9	9	1.0000	0.1111	1.0000	15.5	0.6189
	Wolayta	11	11	1.0000	0.0909	1.0000	17.1	0.6829
	Somali	10	8	0.9333	0.1600	0.8000	14.8	0.5920
	Tigray	5	5	1.0000	0.2000	1.0000	16.4	0.6560
Djibouti		54	51	0.9979	0.0206	0.9444	13.9	0.5547
	Afar	20	19	0.9947	0.0550	0.9500	11.7	0.4699
	Somali	34	32	0.9964	0.0329	0.9412	10.6	0.4236
Kenya		128	113	0.9979	0.0099	0.8828	16.6	0.6653
	Borana	7	7	1.0000	0.1419	1.0000	16.1	0.6457
	Luhya	51	43	0.9929	0.0265	0.8431	14.2	0.5663
	Maasai	45	38	0.9919	0.0301	0.8444	16.8	0.6708
	Other Bantu	11	11	1.0000	0.0909	1.0000	15.5	0.6189
	Other Nilo-Saharans	9	9	1.0000	0.1111	1.0000	17.2	0.6889
	Somali	5	5	1.0000	0.2000	1.0000	14.7	0.5880
Overall		462	410	0.9991	0.0030	0.8874	17.3	0.6927

Table 4: Analysis of molecular variance for Y-STR haplotypes in eastern Africa

	Number of populations	Number of groups [a]	Φ-statistics			F-statistics		
			Φ_{ST} (P^b)	Φ_{CT} (P^b)	Φ_{SC} (P^b)	F_{ST} (P^b)	F_{CT} (P^b)	F_{SC} (P^b)
All eastern Africa	20	1	0.338 (0.000)			0.011 (0.000)		
Country-level analysis								
Eritrea	5	1	0.438 (0.000)			0.013 (0.000)		
Ethiopia	7	1	0.078 (0.000)			0.009 (0.000)		
Djibouti	2	1	0.299 (0.000)			0.013 (0.013)		
Kenya	6	1	0.138 (0.000)			0.009 (0.002)		
Overall	20	4	0.361 (0.000)	0.168 (0.037)	0.232 (0.000)	0.012 (0.000)	0.001 (0.188)	0.011 (0.000)

[a] Populations grouped into countries

[b] The *P*-value indicates the fraction of cases in which a Φ-value greater than the quoted value is obtained in a permutation test of samples across populations (Φ_{ST} /F_{ST} and Φ_{SC} /F_{SC}) and populations across groups (Φ_{CT}/F_{CT}).