Genetic evidence for complexity in ethnic differentiation and history in East Africa

Poloni, Estella S.; Naciri, Yamana; Bucho, Rute; Niba, Régine; Kervaire, Barbara; Excoffier, Laurent Georges Louis; Langaney, André; Sanchez-Mazas, Alicia

How to cite

POLONI, Estella S. et al. Genetic evidence for complexity in ethnic differentiation and history in East Africa. In: Annals of human genetics, 2009, vol. 73, n° 6, p. 582–600. doi: 10.1111/j.1469-1809.2009.00541.x

This publication URL: https://archive-ouverte.unige.ch/unige:2714

Publication DOI: 10.1111/j.1469-1809.2009.00541.x

Annals of human genetics

doi: 10.1111/j.1469-1809.2009.00541.x

Genetic Evidence for Complexity in Ethnic Differentiation and History in East Africa

Estella S. Poloni[1*,#], Yamama Naciri[2#], Rute Bucho[1,2], Régine Niba[2], Barbara Kervaire[3], Laurent Excoffier[4,5], André Langaney[1,6] and Alicia Sanchez-Mazas[1]

[1]*Laboratoire d'Anthropologie, Génétique et Peuplements, Département d'Anthropologie et d'Écologie, Université de Genève, 1211 Geneva 4, Switzerland*

[2]*Laboratoire de Systématique Végétale et Biodiversité, Unité de Phylogénie et Génétique Moléculaires, Conservatoire et Jardin botaniques, 1292 Chambésy, Geneva, Switzerland*

[3]*Unité d'Immunologie de la Transplantation, Laboratoire National de Référence pour l'Histocompatibilité, Hôpitaux Universitaires de Genève, 1200 Geneva, Switzerland*

[4]*Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, Universität Bern, 3012 Bern, Switzerland*

[5]*Swiss Institute of Bioinformatics, http://www.isb-sib.ch/*

[6]*Muséum National d'Histoire Naturelle, 75281 Paris Cedex 05, France*

Summary

The Afro-Asiatic and Nilo-Saharan language families come into contact in Western Ethiopia. Ethnic diversity is particularly high in the South, where the Nilo-Saharan Nyangatom and the Afro-Asiatic Daasanach dwell. Despite their linguistic differentiation, both populations rely on a similar agripastoralist mode of subsistence. Analysis of mitochondrial DNA extracted from Nyangatom and Daasanach archival sera revealed high levels of diversity, with most sequences belonging to the L haplogroups, the basal branches of the mitochondrial phylogeny. However, in sharp contrast with other Ethiopian populations, only 5% of the Nyangatom and Daasanach sequences belong to haplogroups M and N. The Nyangatom and Daasanach were found to be significantly differentiated, while each of them displays close affinities with some Tanzanian populations. The strong genetic structure found over East Africa was neither associated with geography nor with language, a result confirmed by the analysis of 6711 HVS-I sequences of 136 populations mainly from Africa. Processes of migration, language shift and group absorption are documented by linguists and ethnographers for the Nyangatom and Daasanach, thus pointing to the probably transient and plastic nature of these ethnic groups. These processes, associated with periods of isolation, could explain the high diversity and strong genetic structure found in East Africa.

Keywords: Archival serum, mitochondrial, East Africa, genetic diversity, ethnogenesis

Introduction

Human societies whose members maintain marked cultural ties are commonly expected to display a sharp genetic differentiation from their neighbours. This expectation is partly met with the correlation between genetic differentiation and language affiliation that is found at both large (Cavalli-Sforza et al., 1988; Sokal, 1988; Excoffier et al., 1991; Chen et al., 1995; Poloni et al., 1997, 2005; Sanchez-Mazas et al., 2005; Wood et al., 2005; Belle & Barbujani, 2007) and fine scales (Friedlaender et al., 2007; Lansing et al., 2007; Hassan et al., 2008). However, examples of populations that do not fit this expectation exist, where the genetic diversity shows clear signatures of genetic exchange between culturally different groups (Langaney & Gomila, 1973; Blanc et al., 1990; Chaix et al., 2004). In this study we examine whether a similar situation is found among two East African ethnic groups, namely the Nyangatom and the Daasanach, who are found in a region central to current

*Corresponding author: Estella S. Poloni, Laboratoire d'Anthropologie, Génétique et Peuplements, Département d'Anthropologie et d'Écologie, Université de Genève, Rue Gustave-Revilliod 12, 1211 Geneva 4, Switzerland. Tel: +41 22 379 69 77. Fax: +41 22 379 31 94. E-mail: estella.poloni@unige.ch
#These two authors contributed equally to this work.

Sequence data from this study have been deposited in GenBank under accession nos. FJ887983 to FJ888153 and FJ888154 to FJ888343.

debates on human origins (Ramachandran et al., 2005). East Africa indeed harbours populations with a very large ethnic and genetic diversity assumed to be of very ancient origin (Tishkoff et al., 1996, 2007a; Watson et al., 1997; Quintana-Murci et al., 1999; Kivisild et al., 2004; Campbell & Tishkoff, 2008). Accordingly, some of the deepest branches of the Y chromosome and mitochondrial DNA phylogenies are found in this region (Passarino et al., 1998; Salas et al., 2002; Semino et al., 2002; Cavalli-Sforza & Feldman, 2003; Gonder et al., 2007). From a linguistic perspective, the region can be seen as the centre of an old expansion leading to language differentiation, since the four major African linguistic families (including the Khoisan-related languages of Tanzania) meet in East Africa. Accordingly, the Bantu expansion that occurred from East Africa towards the South has been genetically documented (Pereira et al., 2001; Cruciani et al., 2002), and a recent study evidenced the spread, in the same direction, of Y chromosome lineages, probably through the movements of pastoralist populations (Henn et al., 2008). However, East Africa could also represent a contact zone of already differentiated languages and populations, although these two processes are not mutually exclusive. Indeed, the original heartlands of the Nilo-Saharan and Afro-Asiatic linguistic families (or at least some of the primary Afro-Asiatic subdivisions) have both been tentatively placed in East Africa (Ehret, 2002; Militarev, 2002; Blench, 2006). This raises the question of whether the large cultural and genetic diversity found in East Africa is the footprint of an *in situ* evolutionary history or whether recurrent genetic exchanges between linguistically and genetically differentiated populations occurred, due to a higher mobility of human groups than usually assumed. Assuming an ancient *in situ* differentiation of present day East African populations, we expect to observe little genetic structure and extensive haplotype sharing if large effective population sizes have been maintained. Instead, with smaller effective sizes and gene flow restricted by linguistic barriers, we would expect to observe a stronger population structure, with low levels of diversity within populations and little sharing of haplotypes between populations.

To address this issue, we have sequenced the two mitochondrial hypervariable segments and genotyped four mitochondrial coding-region SNPs on DNA extracted from sera that were collected in the early seventies among the Nyangatom and the Daasanach. These two neighbouring East African populations dwell in the Lower Omo River Valley (LORV), the south-western corner of Ethiopia, bordering on Kenya and Sudan. A dozen different ethnic groups live in this area, most of whom are transhumant agripastoralists relying mainly on cattle rearing (Moseley & Asher, 1994; Tornay, 2001). These groups display strong mutual antagonist relationships, probably promoted by competition for the access to grazing areas (Almagor, 1978; Tornay, 2001). This ethnic diversity is translated into linguistic diversity, as two of the four African language families are found in the region (Moseley & Asher, 1994), i.e. Nilo-Saharan (e.g. Nyangatom, Mursi, Muguji, Turkana) and Afro-Asiatic, the latter being further represented by languages pertaining to two of its primary branches: Omotic (e.g. Karo) and Cushitic (e.g. Daasanach, Arbore). Interestingly, the Nyangatom and Daasanach, like the majority of the other ethnic groups in the area, are assumed to be mainly endogamous, although some rare cases of intermarriage and adoption have been reported (Tornay, 2001). Actually, relationships between the Nyangatom and the Daasanach, and more generally between all the populations in the area, are characterised by an opposition between a chronic state of inter-group warfare and strong personal ties of "friendship" among members of distinct groups (Almagor, 1978; Fukui & Markakis, 1994; Tornay, 2001; Tosco, 2001). This situation provides an opportunity to test whether their different linguistic affiliation constitutes a genetic barrier in a context of close neighbourhood.

Materials and Methods

Populations and Samples

The samples we have studied included 384 sera remaining from those that were collected in 1971–1972 for an epidemiological and serological study of arboviruses in the LORV (Rodhain et al., 1972, 1975). This study benefited from the presence in the field of Serge Tornay, who has been conducting ethnological research in the Nyangatom population for over three decades (Tornay, 2001). The remaining sera included 297 Nyangatom, 69 Daasanach and 17 Turkana samples (see Supporting Information, Text S1). The Daasanach (also referred to as Marille) are located directly north of Lake Turkana. The Daasanach language is a member of the Eastern Cushitic branch of Afro-Asiatic. The Nyangatom (also referred to as Donyiro or Buma) are the north-eastern neighbours of the Daasanach, with settlements following the Kibish and Omo rivers. Their Nilo-Saharan language is a member of the Teso-Turkana branch of the Nilotic group, and it is a sister-language of Turkana, the language of a Kenyan population whose settlements extend up to the western shore of Lake Turkana. Census sizes (Tosco, 2001; Gordon, 2005) are estimated as 35,000 Daasanach (1994 and 1998 censuses), 14,000 Nyangatom (1998 census) and 340,000 Turkana (1994 census). Individual interviews and samplings were carried out by François Rodhain, Serge Tornay and their colleagues in the early seventies (Text S1).

DNA Extraction, D-Loop Sequencing and SNPs Genotyping

DNA can be retrieved from sera (Merriwether, 1999), provided that the serum purification protocol was not efficient enough to completely separate the different blood components. This is often the case when separation is performed in the field (as

it was in our case), leading to sera being "contaminated" by white cells. We used Isoquick Nucleic Acid Extraction kits (Orca Research Inc., Bothell, Washington, USA) to extract DNA from the sera, following the manufacturer's instructions. Because this DNA was degraded and only available at very low concentrations, we used nested PCRs to selectively amplify the first (HVS-I) and second (HVS-II) hypervariable segments (Text S1). Internal PCR products (PCR2) were checked on agarose gels, and successful amplifications were purified using Prep-A-Gene beds (Biorad, Hercules, California, USA) before sequencing. Forward and reverse sequencing was performed using each of the two PCR2 primers pairs. We used Perkin-Elmer standard protocols for BigDye Terminator kits and the sequences were run on an ABI 377 Perkin Elmer automated sequencer (Perkin Elmer, Waltham, Massachusetts, USA). The forward and reverse sequences were then stored in files and assembled using the AutoAssembler software (Applied Biosystems, Foster City, California, USA).

Primers were designed to amplify short DNA fragments that included the four SNPs positions 3594, 10810, 10873 and 10400, commonly used to characterise haplogroups L1, L2, L3, M and N (Jobling et al., 2004). The PCR fragments were purified using ExoSAP IT (Amersham Biosciences, Little Chalfont, UK) following the manufacturer's protocol and then digested using appropriate restriction enzymes (*Hpa*I for SNP 3594, *Hinf* I for SNP 10810, *Mnl*I for SNP 10873 and *Alu*I for SNP 10400, Text S1).

Preventing Contamination

Risks of contamination between samples or from the staff in the laboratory were minimised as follows: staff wore gloves, white coat and caps, all material was UV irradiated before handling, PCRs were performed in a separate room under a flow hood, and sterile tubes and filter tips were used. No contamination from the lab staff was detected, and no signal of such contamination has been found in the negative controls used for HVS-I, HVS-II or the four SNPs. To further ensure that no contamination occurred between the samples, we systematically compared all the sequences by pairs of individuals that were amplified during the same amplification session (Text S1). Out of the 362 sequences of the final data set, our conservative estimate leads to approximately 0.8% of the total data set that could possibly reflect contamination (Text S1).

Alignment and Classification of LORV Sequences

All LORV sequences were manually aligned in PAUP (Swofford, 1991), and subsequently checked with ClustalW in BioEdit (Hall, 1999). All mutations relative to the revised Cambridge Reference Sequence (rCRS, Andrews et al., 1999) were scored. For HVS-I, the stretch between np 16182 and np 16193 was not taken into account in most analyses, because of uncertainty in the correct alignment (Bendall & Sykes, 1995). For HVS-II, all indel positions embedded in poly-C, poly-A and poly-T tracks (np 308, 309, 309.1, 309.2, 315.1 and 356.1) were also discarded for the same reason.

Sequences were classified into haplogroups on the basis of HVS-I, HVS-II and the four SNPs, following Kivisild et al. (2004) and references therein, as well as the erratum to that publication (Table S1). For this purpose, a parsimonious alignment of the 16182–16193 stretch was used. Haplogroup distributions in the LORV populations were compared with the data from Kivisild et al. (2004) and Tishkoff et al. (2007a). Tentative dating of haplogroups was computed as fully explained in Text S1.

Database of HVS-I Sequence Variation

To investigate the geographic context of Southern Ethiopian mtDNA diversity, we established a database of HVS-I sequences from Africa and neighbouring areas in the Middle East, West Asia, and South Europe, including populations that had been sequenced for at least 20 individuals. The resulting database contained 6711 aligned sequences from 136 populations (4119 from Africa, 1404 from Southern Europe, and 1188 from the Middle East and West Asia, Table S2), spanning 264 bp between positions 16090 to 16365, after discarding positions 16182 to 16193, as explained above. Following Salas et al. (2002) and Sanchez-Mazas & Poloni (2008), the 136 samples were allocated to seven geographic regions (Fig. S2), i.e. East Africa (EA), Central and Southwest Africa (CSWA), West Africa (WA), Southeast and South Africa (SESA), North Africa (NA), the Middle East and West Asia (MEWA), and South Europe (SE).

Diversity and Structure of Populations

Arlequin ver. 3.11 (Excoffier et al., 2005) was used to estimate several indices of diversity within populations (gene diversity, nucleotide diversity and mean number of pairwise differences) and to perform selective neutrality tests based on Tajima's D and Fu's F_s. A Fu's F_s value was considered significant at the 5% level when its associated P-value was below 2%, as suggested by Fu (1997) and further confirmed in Arlequin documentation. We also computed allelic richness ($R_{(g)}$, Petit et al., 1998) based on the rarefaction index of Hurlbert (1971), setting the value of g equal to the smallest sample size in the dataset. We also tested the fit of the observed mismatch distributions with each of the two models of population expansion implemented in Arlequin, i.e. a sudden stepwise demographic expansion (model 1: Schneider & Excoffier, 1999) and a spatial expansion (model 2: Ray et al., 2003; Excoffier, 2004).

The Arlequin software was also used to compute Reynolds' genetic distances (Reynolds et al., 1983) between pairs of populations. These distances were either based on frequency distributions only (conventional F_{ST} indices, Weir & Cockerham, 1984), or weighted by an evolutionary distance between haplotypes (Φ_{ST} indices, Excoffier et al., 1992). For HVS-I sequence data, the Kimura-2P model was used, with a Gamma correction of 0.4, a transition/transversion ratio of 10/1, and without considering indels. For haplogroup data, a molecular distance was designed by simply counting the minimum number of mutational steps separating lineages, therefore accounting for their phylogenetic relationships. A hierarchical AMOVA framework was applied to infer the proportion of the total genetic variation due to

differences between groups of populations (Φ_{CT}) and between populations within these groups (Φ_{SC}). Pairwise Reynolds' genetic distances were submitted to Principal Coordinates Analysis (PCA) using GenAlEx ver. 6 (Peakall & Smouse, 2006) and Multidimensional Scaling Analysis (MDS) using NTSYSpc (Rohlf, 2007).

Results

HVS Sequences, Coding-Region SNPs and Genetic Diversity in the LORV Samples

Out of the 379 Nyangatom, Daasanach and Turkana serum-extracted DNA samples from the LORV, we obtained 171 HVS-I sequences (112 Nyangatom, 49 Daasanach and 10 Turkana, 45% global success rate) of about 400 bp, and 190 HVS-II sequences (136 Nyangatom, 47 Daasanach and 7 Turkana, 50% global success rate) of about 460 bp in the final data set (Text S1). All HVS-I and HVS-II sequences were submitted to GenBank (accession numbers FJ887983 to FJ888153 for HVS-I and FJ888154 to FJ888343 for HVS-II). With respect to the 4 SNPs (3594, 10400, 10810 and 10873), at least two of them were successfully genotyped in 84 LORV DNA samples, 64 of which were also successfully sequenced in HVS-I, 11 in HVS-II but not in HVS-I, and the remaining nine were neither sequenced in HVS-I nor in HVS-II (see Text S1 for detailed success scores). High and overall similar levels of diversity were observed in the LORV samples (Text S1).

Haplogroups Distributions in the LORV Samples

The complete classification of the 171 HVS-I sequences from the LORV is shown in Table S1 and Figure S1 (provided as Supporting Information), and we describe in detail the salient features of this classification in the additional information (Text S1). Interestingly, we observed two apparently new star-like clades of haplotypes, that we tentatively renamed L0g and L3i1. The L0g clade appears as a sister-clade of L0a that lacks the np 16188 transversion (Fig. S1a). The L3i1 sub-clade is defined by HVS-I motif 16153–16174-16223–16319 (Fig. S1d).

We found that the main basal branches of the mitochondrial phylogeny were represented in the LORV populations, i.e. the L0, L1, L5, L2, L6, L4 and L3 haplogroups, albeit at quite different frequencies (Fig. 1; the Turkana sample was not considered here given its very small size). Notable differences between the Nyangatom and the Daasanach included haplogroups L0f, L5* and L4g (Figs. 1b, 1c and 1d), all of them being more frequent in the latter than in the former, whereas the Nyangatom were found more diversified in the L2, L6 and L3 lineages than the Daasanach (Figs. 1d and 1e). Despite these sharp contrasts in haplogroups frequencies, we found a substantial level of haplotype sharing among populations, in that most star-like clades in the networks included both Nyangatom and Daasanach haplotypes along with haplotypes from other Ethiopian populations (Fig. S1).

A notable characteristic of these networks is the substantial diversity of LORV haplotypes found in several clades. For instance, four distinct L0a2 haplotypes were found in the Nyangatom, two of which were shared with the Daasanach (and the Turkana, Figure S1a), whereas the L0a2 haplotypes found among the neighbouring Northern Ethiopian populations studied by Kivisild et al. (2004) appear as more derived (Figure S1a). We could not verify however if the LORV haplotypes bear the characteristic COII/$tRNA^{lys}$ 9-bp deletion. We also observed several haplotypes of supposedly more western origins (Salas et al. 2002), such as the four L2b Nyangatom haplotypes that are apparently less derived than those found in Northern Ethiopians (Fig. S1a), as well as several L3b, L3d and L3e haplotypes (Fig. S1c).

Assuming that all the sequence diversity observed in a clade had accumulated *in situ*, three distinct configurations can be evidenced among these clades (see Table S3 for estimated TMRCAs): first, young clades that seemed restricted to one population (such as the Daasanach sub-clade stemming from L0f or the single Daasanach L5* haplotype, Fig. S1a), with TMRCA confidence intervals below 20 thousand years (kyrs); then young clades including haplotypes from at least two out of the three LORV populations (such as L0a2, L5a1 and the sub-clade at the tip of L5a2 in Fig. S1a; the sub-clade of L4g whose central node is represented by two Nyangatom sequences in Fig. S1b; the sub-clade at one tip of L3h, the L3i1 clade and the sub-clade in L3x1 in Figure S1d) with TMRCA confidence intervals below 30 kyrs; finally, older clades that did also include haplotypes from at least two out of the three LORV populations (such as L0a1 and L5a2 in Fig. S1a, and other sub-clades in L4g, Fig. S1c). Given that uncertainty in the mutation rate was not accounted for in the TMRCA estimations, the confidence intervals in Table S3 are only associated with the level of diversity. This calls for extreme caution in their interpretation. For instance, the estimated age of the L0a2 clade in the Nyangatom was 23 kyrs (with a 95% confidence interval of 12 to 33 kyrs) which is inconsistent with the postulated spread of L0a2 sequences through the Bantu expansion (Salas et al., 2002 Kivisild et al., 2004).

North-South Genetic Differentiation in Ethiopia

About 95% of the Nyangatom and Daasanach sequences belong to the L lineage (Fig. 1, the Turkana sample was not

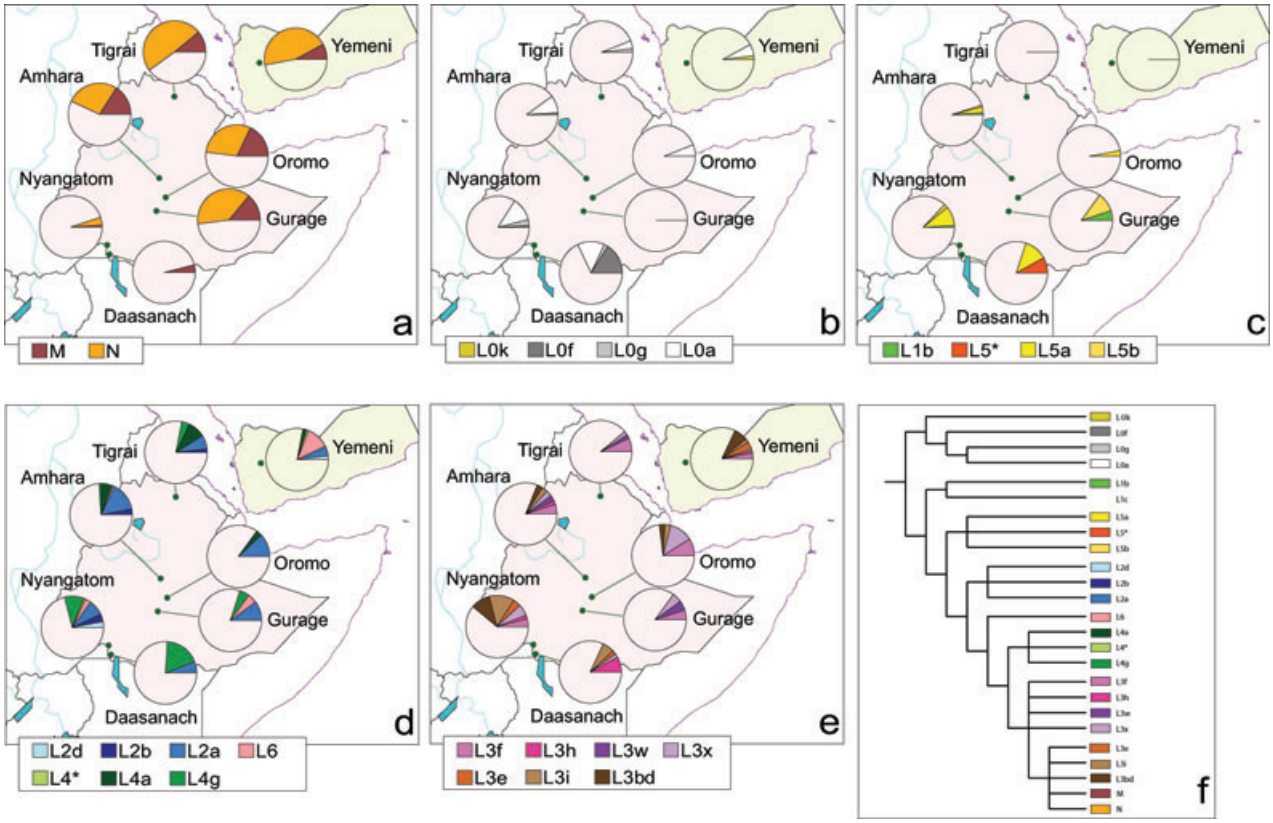

Figure 1 Frequency distributions of the mtDNA haplogroups in the Nyangatom, Daasanach (this study) and five other Ethiopian and Yemeni populations (data from Kivisild et al., 2004). (a) L0 sequences. (b) L1 and L5 sequences. (c) L2, L6 and L4 sequences. (d) L3 sequences. (e) M and N sequences. (f) Schematic representation of the mtDNA haplogroups detected in this study.

considered here given its very small size). This is in sharp contrast with what is found in neighbouring Northern Ethiopian populations (the Ethiopian Oromo, Amhara, Tigrai and Gurage populations studied by Kivisild et al. (2004), all referred to as "Northern Ethiopians" hereafter) where the frequency of L sequences varies between 40% (Tigrai) and nearly 60% (Amhara). Conversely, M and N lineages were much more frequent in the Northern Ethiopians (from 43% to 60% of the sequences, Fig. 1a) than in either the Nyangatom (5%) or the Daasanach (4%).

The frequency distributions of the 24 haplogroups shown in Figure 1 were first used to compute conventional population pairwise F_{ST} indices. The Nyangatom and the Daasanach were significantly differentiated, at the 1% level, both from each other ($F_{ST} = 2.2\%$, $P < 0.003$) and from the other five populations considered here. Conversely, among the five populations studied by Kivisild et al. (2004), most population pairs were found undifferentiated at the 5% level. Interestingly, the Nyangatom were less differentiated from the Northern Ethiopians (average $F_{ST} = 7.8 \pm 3.3\%$) than the Daasanach were (average $F_{ST} = 11.2 \pm 3.6\%$). Comparable results were observed when Φ_{ST} indices were computed by weighting the haplogroup frequencies by their phylogenetic relationships. The Nyangatom and the Daasanach were significantly differentiated ($\Phi_{ST} = 4.3\%$, $P < 0.002$), while the Northern Ethiopian populations formed a homogeneous group that was less differentiated from the Nyangatom (average $\Phi_{ST} = 6.2 \pm 2.8\%$) than from the Daasanach (average $\Phi_{ST} = 18.5 \pm 3.8\%$). Finally, we also computed Φ_{ST} indices on the basis of sequence variation in HVS-I (sequence-based Φ_{ST}) and found results that reproduced again the patterns obtained with haplogroups, both when unweighted (conventional F_{ST}) and when weighted by the phylogenetic distance separating haplogroups (haplogroup-based Φ_{ST}). Therefore, these analyses confirmed that genetic distances between populations computed on sequence variation reproduced with high accuracy the genetic relationships between populations that are inferred from haplogroup distributions.

Genetic Structure in East Africa

We compared the haplogroup frequency distributions observed in Ethiopia with those found in the Tanzanian populations studied by Tishkoff et al. (2007a), using the level of phylogenetic resolution adopted in that study (Fig. 2). This

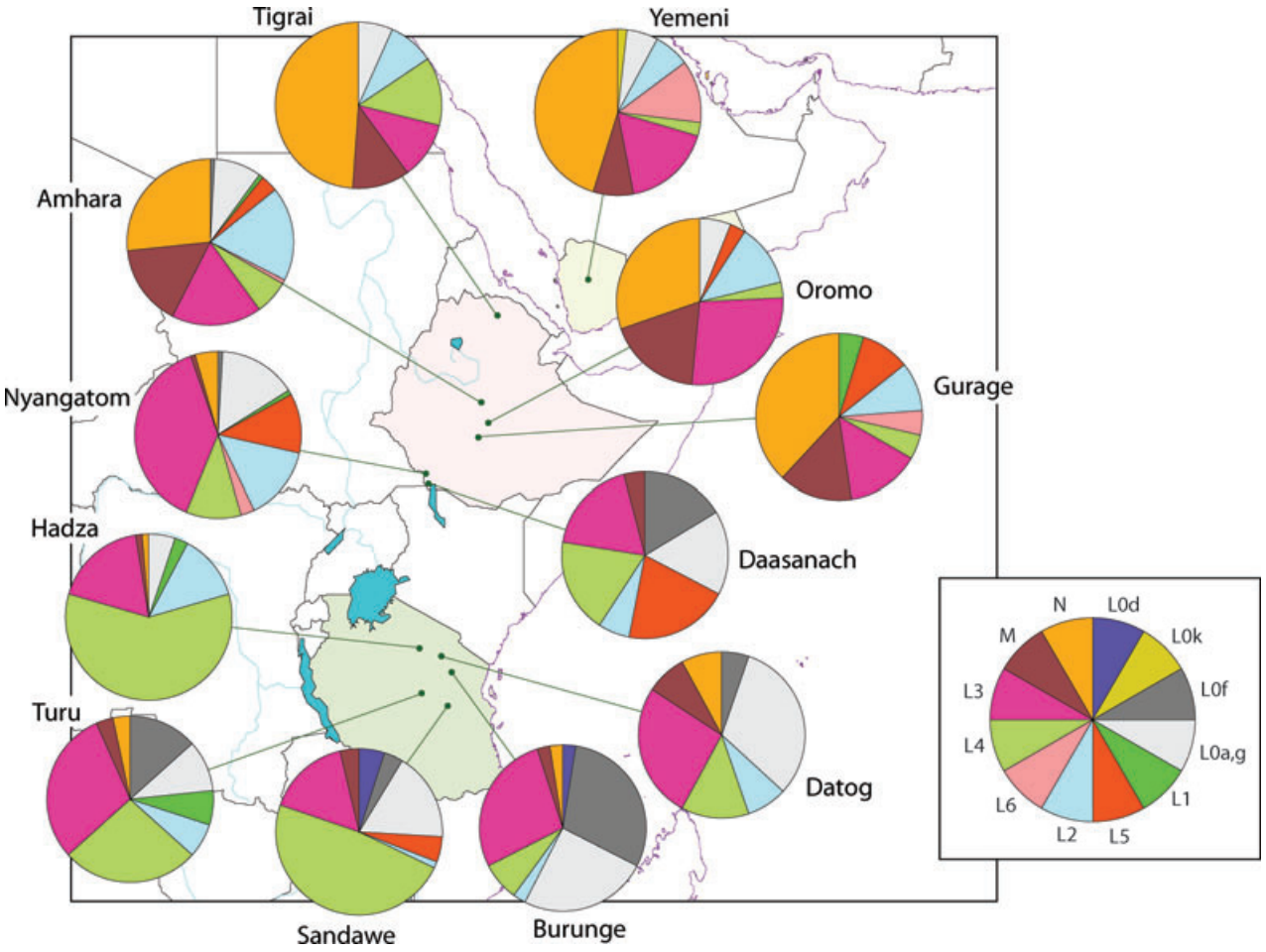

Figure 2 Frequency distributions of the major mtDNA haplogroups in the Nyangatom, Daasanach (this study) and ten other East African populations (data from Kivisild et al., 2004 and Tishkoff et al., 2007a).

comparison revealed that the LORV populations were more similar in their mtDNA diversity to the Tanzanian populations than to the Northern Ethiopians. Besides the low frequencies of M and N lineages that also characterised the Tanzanian populations (totaling between 2 and 11% of the sequences), L0a and L0g sequences also reached high frequencies here (up to 25% and 32% in the Datog and Burunge, respectively) and L0f sequences, commonly observed among the Daasanach (16%), were found to be particularly frequent in the Burunge (30%) and Turu (13%).

Despite these similarities, haplogroup-based Φ_{ST} indices indicated that the Nyangatom were significantly differentiated, at the 5% level, from most of the Tanzanian populations, with Φ_{ST} values ranging from 3.9% (Datog) to 17.6% (Burunge). On the other hand, non-significant Φ_{ST} values between the Daasanach and several Tanzanian populations were observed (Turu, Datog, and Burunge). A multidimensional scaling (MDS) analysis of the genetic distances computed on the haplogroup-based Φ_{ST} indices summarises this complex pattern of relationships observed in East Africa (Fig. 3). The Northern Ethiopian populations clustered in the upper part of the MDS plot, forming a homogeneous group (AMOVA $\Phi_{ST} = 1.2\%$, $P = 0.054$). The LORV and Tanzanian populations were found scattered in the lower part of the plot and displayed a significant level of genetic structure (AMOVA $\Phi_{ST} = 6.2\%$, $P < 0.0001$). As expected (Gonder et al., 2007; Tishkoff et al., 2007a), the Hadza, located to the right of the plot, were found to diverge significantly from all other populations. An MDS analysis of genetic distances among 20 East African and three Arabian populations computed on HVS-I sequence variation produced very similar results (Fig. 4). Again, the Khoisan-related Hadza samples were found to be isolated, whereas the Northern Ethiopians clustered in a homogeneous group, together with the Nubians from Sudan and the Somalis. This analysis based on HVS-I sequences also revealed substantial genetic structure among the other East Africans, which included populations from Southern Sudan (Dinka) and Kenya (Kikuyu), in addition to the LORV and

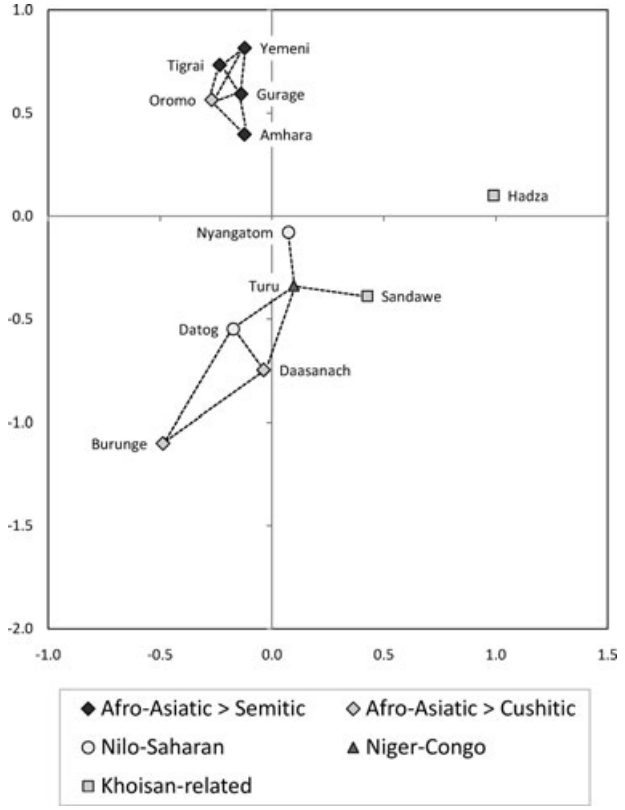

Figure 3 Multidimensional scaling analysis of genetic distances between 20 East African and 3 Arabian population samples, computed from haplogroup-based Φ_{ST} indices. A very good fit to the original distances was achieved (stress value = 0.031). Symbols indicate linguistic affiliation of the populations (language families shown in caption). Broken lines link population pairs that were not differentiated at the 5% level.

Tanzanian populations. This MDS plot evidenced the fact that the genetic diversity of East Africans was almost entirely covered by that found among the Nilo-Saharan populations on the one hand, and among the Afro-Asiatic Cushitic populations on the other (shaded areas in Fig. 4).

In keeping with the MDS analysis of Figure 4, heterogeneous levels of genetic diversity within East African populations were observed (Table 1). In terms of allelic richness, gene diversity and mean number of pairwise differences (MNPD), the sharpest contrast was between the low diversity values found in the two Hadza samples, compared to the other populations. Conversely, the high level of differentiation found among the Afro-Asiatic Cushitics and among the Nilo-Saharans (Fig. 4) was not associated with low genetic diversities within populations (Table 1). For instance, the highest MNPD values in East Africa were observed among the Sukuma and Datog from Tanzania, the Turkana from Kenya, and the Nyangatom and Daasanach from the LORV (values ranged from 7.9 to 8.4, Table 1). Such high values were only seldom observed elsewhere in Africa (Table S4).

In East Africa, only four populations out of 20 displayed significantly negative Tajima's D values at the 5% level (Dinka, Tigrai, Amhara1, Amhara2, Table 2). Conversely, the Fu's F_s were generally significantly negative at the 5% level, with the exception of those displayed by four Tanzanian populations (Burunge, Turu, Sandawe, and Hadzabe). However, as shown in Table 2 (and Table S4), these non significant F_s values were globally not in agreement with the tests of population expansion, neither under the demographic expansion model (model 1), nor under the spatial expansion model (model 2). As a single case of rejection of the spatial expansion model was observed (Beta Israel) out of 20 tests performed on the East African data, the type-I error threshold was not exceeded. Moreover, considering the entire dataset of 136 populations, only four populations rejected the null hypothesis out of 136 tests at the 5% level (Table S4), a proportion that is smaller than the type-I error rate, so that overall the model of a spatial expansion was conserved.

Patterns and Levels of Genetic Structure in Africa and Surrounding Regions

To gain better insights into the pattern of genetic relationships among populations, we analyzed the matrix of pairwise Reynolds' genetic distances among the 136 populations of the HVS-I database both by means of MDS analysis and Principal Coordinates Analysis (PCA). The plot of the MDS analysis on is shown in Figure 5. Three populations appeared as clear outliers in this plot, the Mbenzele (as evidenced by Destro-Bisol et al., 2004), the Ju|'hoansi and the Herero, and this pattern of extreme genetic differentiation was confirmed by PCA (Fig. S3). Besides these three outlier populations, which were also remarkable in the low level of internal genetic diversity they displayed (both in gene diversity and mean number of pairwise differences between sequences, Table S4), the general pattern of variation among populations was found to be geographically structured, with a clear north-south orientation. Indeed, when the three outliers were not taken into account, the correlation between the latitudinal location of the populations and the PCA's first axis coordinates was both very high and significant ($r = 0.882$, 95% C.I. = [0.84; 0.91], Fig. S3). As can be seen in the MDS plot of Figure 5a, populations from Southern Europe were located in the uppermost section of the plot, followed then by Middle Easterners, West Asians, and North Africans, whereas all Sub-Saharan populations were located below these groups. A geographic structure of the genetic variability was also observed among Sub-Saharan Africans, with most West Africans clustering towards the left of the plot, most East Africans towards the right, and the populations from the South of the continent mainly

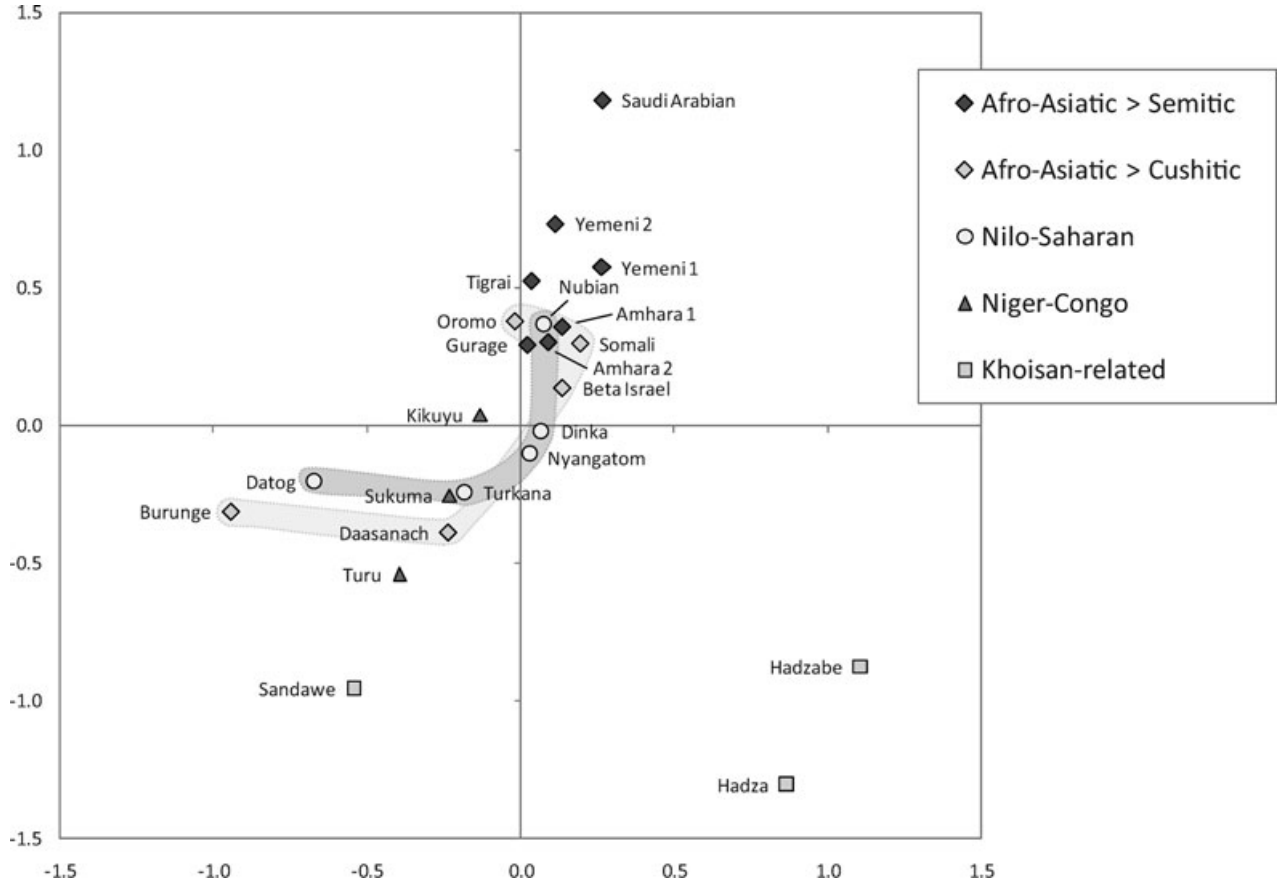

Figure 4 Multidimensional scaling analysis of genetic distances between 20 East African and 3 Arabian population samples, computed from HVS-I sequence-based Φ_{ST} indices. The populations are color-coded according to linguistic affiliation, as shown in the caption. A very good fit to the original distances was achieved (stress value = 0.079). The shaded areas highlight the positions in the plot of, respectively, the Cushitic (Afro-Asiatic) and the Nilo-Saharan populations.

below these groups. Thus, populations from the same geographic group tended to cluster together, although no clear cuts between the different geographic origins were observed. Actually, the less clustered geographic group in this analysis was the East African group.

The level of genetic structure (Φ_{ST}) inferred from sequence variation in the 136 populations of the HVS-I database was almost 16% (Table 3), but the three genetically divergent populations outlined in the multivariate analyses (i.e. the Mbenzele, Ju|'hoansi and Herero) contributed substantially to this level of differentiation, as testified by the steady decrease in Φ_{ST} values when these samples were sequentially removed from the computations, from 15.6% to 13.8%. In Africa, Φ_{ST} values within geographic groups were all significant, and ranged from 2.6% for West Africa (WA) to 14.6% for South-eastern and South Africa (SESA). Here again, Φ_{ST} values within some geographic groups were inflated by a few very divergent populations, such as the Ju|'hoansi and Herero in the Southeast and South Africa (SESA) group, or the Mbenzele in the Central and Southwest Africa (CSWA) group. As shown in Table 3, Φ_{ST} estimations obtained without considering those divergent populations in their respective geographic groups were consistently below 4%, except for the East Africa (EA) and the SESA groups. However, whereas at least 70% of the genetic distances between populations were significant in EA, only 53% of the genetic distances were significant in SESA (and dropped to 37% when the Ju|'hoansi and Herero were not considered). Thus, as shown in Table 3, the level of genetic structure in a given group was only marginally associated with the proportion of significant distances between populations in the group. The simultaneous comparison of these two estimations between geographic groups indicated that EA shelters the highest level of population structure in the continent.

An AMOVA design was performed to compare the geographic groups by pairs (Table 4). Among almost all the comparisons involving two Sub-Saharan geographic groups, the level of genetic differentiation within groups exceeded the level between those groups even when the most divergent populations were discarded from the analyses. This result

Table 1 HVS-I genetic diversity indices[a] in 20 East African populations and averages for the six other geographic regions considered in this study[b].

Population (Country)	Language Family[c]	Sample size	Number of haplotypes	Allelic richness[d]	Gene diversity	MNPD[e]	Reference
East Africa (EA), 20 samples, 1024 sequences							
Nubian (Sudan)	NS	42	29	15.9	0.970	6.9	Krings et al., 1999
Dinka (Sudan)	NS	46	40	18.8	0.993	7.2	Krings et al., 1999
Tigrai (Ethiopia)	AA>Semitic	44	38	18.7	0.993	7.2	Kivisild et al., 2004
Amhara 1 (Ethiopia)	AA>Semitic	74	61	18.9	0.994	7.2	Thomas et al., 2002
Amhara 2 (Ethiopia)	AA>Semitic	120	84	18.5	0.991	7.5	Kivisild et al., 2004
Gurage (Ethiopia)	AA>Semitic	21	21	20.0	1.000	7.1	Kivisild et al., 2004
Beta Israel (Ethiopia)	AA>Cushitic	48	26	14.7	0.961	7.5	Thomas et al., 2002
Oromo (Ethiopia)	AA>Cushitic	33	30	18.8	0.992	7.2	Kivisild et al., 2004
Daasanach (Ethiopia)	AA>Cushitic	49	34	16.7	0.979	8.2	This study
Nyangatom (Ethiopia)	NS	112	64	17.5	0.984	7.9	This study
Somali	AA>Cushitic	27	23	17.8	0.989	5.9	Watson et al., 1996, 1997
Turkana (Kenya)	NS	47	37	17.6	0.984	8.0	This study and Watson et al., 1996, 1997
Kikuyu (Kenya)	NC	24	21	17.9	0.989	6.7	Watson et al., 1996, 1997
Burunge (Tanzania)	AA>Cushitic	38	22	13.7	0.937	7.5	Tishkoff et al., 2007a
Datog (Tanzania)	NS	39	29	17.0	0.981	8.1	Tishkoff et al., 2007a
Turu (Tanzania)	NC	29	18	13.9	0.951	6.6	Tishkoff et al., 2007a
Sukuma (Tanzania)	NC	21	21	20.0	1.000	8.4	Knight et al., 2003
Sandawe (Tanzania)	Kh	82	26	10.7	0.827	6.3	Tishkoff et al., 2007a
Hadza (Tanzania)	Kh	79	26	9.2	0.713	4.7	Tishkoff et al., 2007a
Hadzabe (Tanzania)	Kh	49	7	5.3	0.724	4.3	Knight et al., 2003
Average ± standard deviation		*51.2 ± 28.3*	*32.9 ± 18.0*	*16.1 ± 3.9*	*0.948 ± 0.087*	*7.0 ± 1.1*	
Range		*21–120*	*7–84*	*5.3–20.0*	*0.713–1.000*	*4.3–8.4*	
North Africa (NA), 18 samples, 950 sequences[f]							
Average ± standard deviation		*52.8 ± 20.7*	*29.8 ± 9.2*	*14.9 ± 2.5*	*0.951 ± 0.032*	*4.6 ± 0.9*	
Range		*23–115*	*15–44*	*9.8–18.5*	*0.886–0.990*	*3.1–6.3*	
Central and Southwest Africa (CSWA), 24 samples, 977 sequences[f]							
Average ± standard deviation		*40.7 ± 19.3*	*26.9 ± 11.7*	*16.1 ± 2.7*	*0.968 ± 0.041*	*6.4 ± 1.0*	
Range		*20–110*	*10–68*	*7.1–19.0*	*0.798–0.995*	*4.2–8.3*	
West Africa (WA), 15 samples, 722 sequences[f]							
Average ± standard deviation		*48.1 ± 32.4*	*30.8 ± 14.9*	*16.2 ± 1.4*	*0.972 ± 0.015*	*6.3 ± 0.6*	
Range		*22–121*	*14–68*	*13.2–18.5*	*0.942–0.992*	*5.0–7.0*	
Southeast and South Africa, 15 samples, 446 sequences[f]							
Average ± standard deviation		*29.7 ± 22.8*	*14.7 ± 8.9*	*11.7 ± 3.3*	*0.877 ± 0.142*	*6.0 ± 1.5*	
Range		*20–109*	*6–45*	*4.9–16.4*	*0.399–0.981*	*2.3–7.5*	
Middle East and West Asia (MEWA), 18 samples, 1188 sequences[f]							
Average ± standard deviation		*66.0 ± 42.7*	*44.0 ± 29.3*	*16.6 ± 2.5*	*0.973 ± 0.023*	*4.9 ± 0.8*	
Range		*20- 191*	*17–121*	*11.5–19.1*	*0.923–0.995*	*3.4–6.4*	
South Europe (SE), 26 samples, 1404 sequences[f]							
Average ± standard deviation		*54.0 ± 18.6*	*32.6 ± 9.4*	*14.8 ± 1.3*	*0.933 ± 0.032*	*3.3 ± 0.5*	
Range		*25–106*	*20–59*	*13.2–17.2*	*0.882–0.975*	*2.4–4.7*	

[a]All sites with less than 5% of undetermined nucleotides per position and per sample were used in computations, thus 264 nucleotides from np 16090 to np 16365, excluding nps 16182 to 16193. All statistics were estimated by giving equal weight to transitions and transversions, and excluding indels.

[b]Geographic groups are defined in Figure S2 and Table S2.

[c]NS, Nilo-Saharan. AA, Afro-Asiatic. AA>Semitic, Semitic branch of Afro-Asiatic. AA>Cushitic, Cushitic branch of Afro-Asiatic. NC, Niger-Congo. Kh, Khoisan-related click-languages of Tanzania.

[d]The allelic richness index ($R_{(g)}$) is the number of haplotypes expected in a subsample of size $g = 20$.

[e]MNPD, mean number of pairwise differences between sequences (insertions and deletions not considered).

[f]Individual samples values are reported in Table S4.

Table 2 Results of the tests of selective neutrality and population equilibrium and of the spatial expansion model in 20 East African populations and averages for the six other geographic regions considered in this study[a].

Population	Tajima's D (P)[b]	Fu's Fs (P)[c]	P(SSD-2)[d]	τ [C.I. 95%][e]
East Africa (EA), 20 samples, 1024 sequences				
Nubian	−1.40 (0.064)	−15.26 (<0.001)	0.631	5.9 [3.7; 9.4]
Dinka	−1.50 (0.041)	−25.03 (<0.001)	0.915	5.6 [3.8; 8.7]
Tigrai	−1.48 (0.045)	−25.03 (<0.001)	0.183	7.6 [5.2; 8.7]
Amhara 1	−1.52 (0.039)	−24.98 (<0.001)	0.191	7.5 [5.4; 8.3]
Amhara 2	−1.63 (0.020)	−24.77 (<0.001)	0.166	7.8 [5.8; 8.6]
Gurage	−1.14 (0.123)	−16.97 (<0.001)	0.490	7.3 [4.3; 8.8]
Beta Israel	−0.56 (0.329)	−7.62 (0.018)	0.041	−
Oromo	−1.32 (0.079)	−23.81 (<0.001)	0.120	7.6 [5.0; 8.8]
Daasanach	−0.51 (0.351)	−17.57 (<0.001)	0.389	8.9 [5.9; 10.4]
Nyangatom	−1.17 (0.102)	−24.72 (<0.001)	0.594	8.0 [5.7; 9.6]
Somali	−1.38 (0.072)	−15.78 (<0.001)	0.258	6.2 [3.6; 7.5]
Turkana	−1.15 (0.115)	−24.38 (<0.001)	0.223	8.8 [5.9; 10.2]
Kikuyu	−1.37 (0.066)	−12.69 (<0.001)	0.258	5.2 [3.0; 8.6]
Burunge	−0.59 (0.316)	−5.76 (0.032)	0.488	8.5 [5.0; 10.9]
Datog	−1.07 (0.138)	−14.46 (<0.001)	0.306	8.0 [5.5; 10.9]
Turu	−0.88 (0.198)	−4.98 (0.034)	0.429	7.0 [4.1; 9.8]
Sukuma	−1.04 (0.150)	−15.20 (<0.001)	0.454	9.0 [5.8; 10.7]
Sandawe	−0.82 (0.222)	−3.54 (0.161)	0.706	8.0 [4.7; 12.2]
Hadza	−1.20 (0.101)	−8.51 (0.012)	0.619	6.6 [3.3; 10.9]
Hadzabe	0.02 (0.568)	3.60 (0.906)	0.073	5.8 [2.7; 9.8]
Average ± standard deviation[f]				*7.3 ± 1.2*
Range				*5.2–9.0*
North Africa (NA), 18 samples, 950 sequences[g]				
Average ± standard deviation[f]				*3.9 ± 1.4*
Range				*1.9–6.1*
Central and Southwest Africa (CSWA), 24 samples, 977 sequences[g]				
Average ± standard deviation[f]				*5.5 ± 1.3*
Range				*3.4–7.5*
West Africa (WA), 15 samples, 722 sequences[g]				
Average ± standard deviation[f]				*6.0 ± 1.3*
Range				*4.0–8.0*
Southeast and South Africa (SESA), 15 samples, 446 sequences[g]				
Average ± standard deviation[f]				*6.5 ± 2.0*
Range				*1.2–9.1*
Middle East and West Asia (MEWA), 18 samples, 1188 sequences[g]				
Average ± standard deviation[f]				*4.8 ± 1.0*
Range				*3.1–6.4*
South Europe (SE), 26 samples, 1404 sequences[g]				
Average ± standard deviation[f]				*2.7 ± 0.7*
Range				*1.2–4.4*

[a]Geographic groups are defined in Figure S2 and Table S2.
[b]Tajima's D statistic (and statistical significance based on 10'000 simulated random neutral samples at population equilibrium).
[c]Fu's Fs statistic (and statistical significance based on 10'000 simulated random neutral samples at population equilibrium).
[d]P(SSD-2), Probability value of the null hypothesis of the spatial expansion model (model 2), based on 10'000 simulated random samples under the null model.
[e]τ: moment estimator of the time since expansion [95% confidence interval based on 10'000 simulated random samples under the null model]. This estimator is not provided if the probability of the model is < 5%.
[f]Average is computed on values of τ with associated probability ≥5%.
[g]Individual samples values are reported in Table S4.

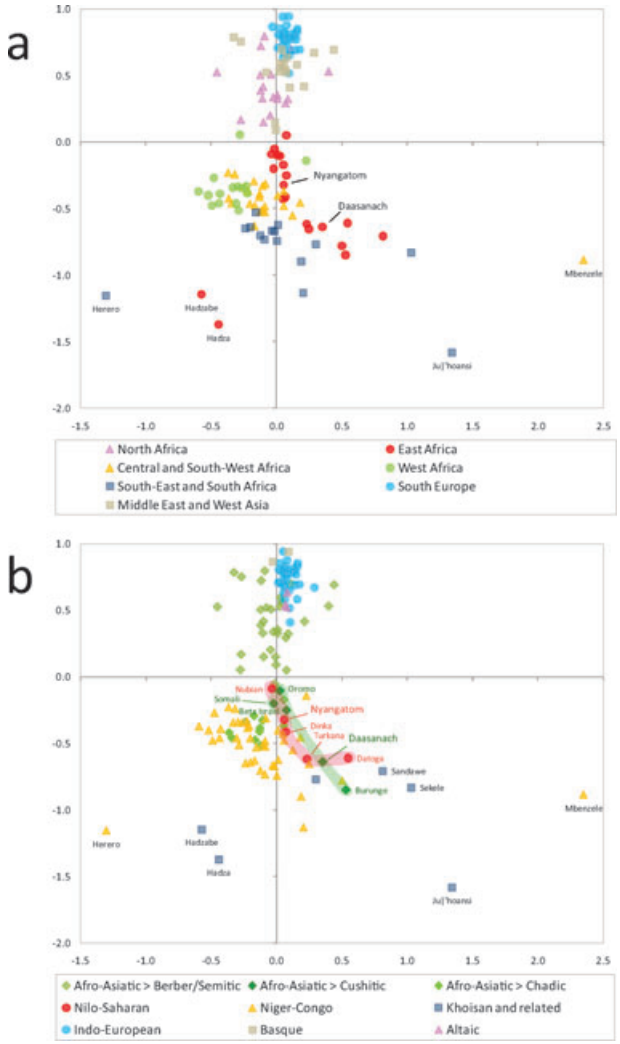

Figure 5 Multidimensional scaling analysis of genetic distances between 136 populations from Africa and surrounding regions, computed from HVS-I sequence-based Φ_{ST} indices. A good fit to the original distances was achieved (stress value = 0.101). As shown in the captions, the populations are color-coded according to geographic location in a, and according to linguistic affiliation in b. The green- and red-shaded areas in b highlight the positions in the plot of, respectively, the Cushitic (Afro-Asiatic) and the Nilo-Saharan populations.

was expected in view of both the generally high levels of genetic structure inferred for the geographic groups (Table 3) and the substantial overlapping of these groups observed in the MDS (Fig. 5a). The only exceptions to this trend were the comparison of East and West Africa (EA versus WA) and that of West Africa with the southern group (WA versus SESA), but only when divergent populations were excluded. The rather low level of genetic structure found among West Africans (Φ_{ST} = 2.6%, Table 3) clearly contributed to these observations. The same trend of similar levels of differentiation between groups and between populations within groups was also observed in the three comparisons involving the neighbouring regions around the Mediterranean area (NA, MEWA and SE). Inversely, all AMOVA analyses involving one such group and a Sub-Saharan group led to higher Φ_{CT} than Φ_{SC} values, in keeping with their differentiation in the upper and lower half-sections of the MDS plot (Fig. 5a). This trend was not due to the few highly divergent populations found in some groups, since the major effect of their exclusion on the Φ-statistics was that of notably decreasing the Φ_{SC} values.

Since each language family has a distinct geographic distribution, a genetic structure associated to linguistic differentiation was expected and was indeed observed (Fig. 5b). In particular, Afro-Asiatic populations displayed a pattern of genetic differentiation developing roughly into three directions, with speakers of Berber and/or Semitic languages mainly clustering in the upper half-section of the MDS plot, speakers of Chadic languages towards the left, and speakers of Cushitic languages towards the right (Fig. 5b). However, given the sampling distribution of populations included in the HVS-I database, zones of geographic contact between distinct linguistic stocks were also represented in the analysis. While the MDS plot displayed a complete overlapping of Chadic (Afro-Asiatic) and Niger-Congo populations from Western, Central and South-western Africa, a more complex pattern was observed in East Africa. Here, both Afro-Asiatic Cushitic and Nilo-Saharan populations differentiated in parallel, along a north-south oriented axis, as highlighted by the shaded areas in Figure 5b.

Distinct patterns of genetic variability were observed in the linguistically-defined African groups (Table 5). Notably, the Khoisan group was characterised by the highest level of genetic structure between populations, with an associated Φ_{ST} value of more than 20%, but also by a low level of molecular diversity within populations (with a mean number of pairwise differences between sequences, MNPD, smaller than 6 on average). The highest levels of molecular diversity within populations were found for both the Afro-Asiatic Cushitic and the Nilo-Saharan groups (average MNPD > 7). But despite this similarity, genetic differentiation between Cushitic populations (Φ_{ST} = 7%) was notably higher than that between Nilo-Saharan populations (Φ_{ST} = 3.9%).

We then tested the significance of genetic differentiation between linguistically-defined groups of Sub-Saharan African populations (Table 6). Because of the extreme differentiation among Khoisan populations, the AMOVA design was restricted to Afro-Asiatic, Nilo-Saharan and Niger-Congo populations. In keeping with the MDS analyses (Figs. 4 and 5b), the Nilo-Saharans were not found differentiated from the Cushitics (Φ_{CT} = 0). All other pairwise comparisons led to significant Φ_{CT} values, except for the comparison involving the Chadic and Niger-Congo groups. However, similarly

Table 3 Levels of genetic structure (Φ_{ST}) and proportion of non significant genetic distances (in %) in geographic groups.

Group[a]	Number of populations in group	Φ_{ST}[b]	Proportion of significant genetic distances at the 5% level
Entire dataset	136 (133)[c]	15.62 (13.83)[c]	86.80 (86.19)[c]
Africa	92 (89)[d]	13.22 (10.70)[d]	84.45 (83.38)[d]
Outside Africa	44	2.54	49.05
North Africa (NA)	18	3.83	73.86
East Africa (EA)	20 (18)[e]	8.66 (5.83)[e]	75.79 (69.93)[e]
Central and Southwest Africa (CSWA)	24 (23)[f]	10.81 (2.87)[f]	49.64 (45.06)[f]
West Africa (WA)	15	2.57	44.76
Southeast and South Africa (SESA)	15 (13)[g]	14.58 (7.84)[g]	53.33 (37.18)[g]
Middle East and West Asia (MEWA)	18	3.05	60.13
South Europe (SE)	26	0.59	22.15

[a]Geographic groups (and codes for) are defined in Table S2 and Figure S2.
[b]Kimura-2P was used as a molecular distance, with a gamma correction of 0.4, a transition/transversion ratio of 10/1 and without considering indels. All Φ_{ST} values are given in percent, and were significant ($P < 0.001$).
[c]In brackets: computations based on 133 populations, i.e. without the Mbenzele, Ju|'hoansi, and Herero samples (see text).
[d]In brackets: computations based on 83 African populations, i.e. without the Mbenzele, Ju|'hoansi, and Herero samples.
[e]In brackets: computations based on 18 EA populations, i.e. without the two Hadza samples.
[f]In brackets: computations based on 23 CSWA populations, i.e. without the Mbenzele sample.
[g]In brackets: computations based on 13 SESA populations, i.e. without the Ju|'hoansi and Herero samples.

to the results of the AMOVA analyses involving geographic groups, we found that the Φ_{CT} values, although significant, were consistently close or inferior to the levels of genetic structure observed between populations within the groups (Φ_{SC} values, Table 6). Thus, similarly to the geographic factor, no strong association between linguistically-defined and genetically differentiated groups emerged from this analysis.

Discussion

Archival Blood Fractions as a Valuable Source of Information

This study demonstrates that archival blood fractions can be a good source of DNA for human polymorphism screening, as has already been reported for numerous nuclear markers in DNA extracted from old red blood cell fractions (Weiss et al., 1994; Buchanan et al., 2000). It is currently more difficult to characterise nuclear polymorphisms on DNA extracted from archival sera because nDNA is rarer than mtDNA in terms of copy numbers. Another important disadvantage of molecular studies based on old plasma sera is that DNA is often degraded into small pieces such that only small segments of DNA (usually <400 bp) can be amplified (Torroni et al., 1993a, 1993b; Merriwether, 1999; Lie et al., 2007). This implies that very special care has to be taken to prevent contamination of the extracted DNA and PCR amplifications, especially when nested PCRs are used (this study). Despite the difficulties associated with the use of archival blood fractions, we were able to amplify about 50% of the samples available with very weak contamination suspicion (less than 1%). Moreover, the same serum-extracted DNAs have been successfully used to characterise the HLA variation (Sanchez-Mazas et al., in preparation). Furthermore, we have shown that genetic distances based on HVS-I sequence variation, even though restricted to the non-coding fraction of the mitochondrial genome, were as informative as genetic distances inferred from haplogroups to uncover genetic differentiation patterns among populations.

Inferences on the Nyangatom and Daasanach Past History

In this study, we have found that the Nilo-Saharan Nyangatom and the Afro-Asiatic Daasanach, two agripastoralist semi-nomadic populations settled in close vicinity in the Lower Omo River Valley, exhibit similarly high levels of genetic diversity, a result that has been repeatedly found for other populations in East Africa, either for mtDNA (Watson et al., 1996, 1997; Salas et al., 2002; Kivisild et al., 2004) or for other markers (Lovell et al., 2005; Sanchez-Mazas & Poloni, 2008). This study however showed that although very diverse, the Nyangatom and the Daasanach are also significantly differentiated from each other. This result could be expected on the basis of their respective linguistic affinities. However, the Turkana, who are the Nyangatom and Daasanach

AMOVA design[a]		Φ-statistics[b]	
Group 1 versus	Group 2	Φ_{CT}[c]	Φ_{SC}[c]
Within Sub-Saharan Africa			
EA	CSWA	2.51 (2.67)	9.68 (4.40)
EA	WA	5.18 (5.66)	6.47 (4.57)
EA	SESA	3.27 (3.90)	10.29 (6.36)
CSWA	WA	2.04 (1.94)	7.69 (2.76)
CSWA	SESA	2.23 (2.56)	11.92 (4.31)
WA	SESA	6.30 (6.40)	7.46 (4.54)
Between Sub-Saharan Africa and neighbouring regions around the Mediterranean area			
EA	NA	11.47 (10.76)	6.98 (5.08)
CSWA	NA	11.98 (12.17)	8.31 (3.32)
WA	NA	12.90	3.19
SESA	NA	17.96 (18.33)	8.42 (5.46)
EA	MEWA	13.91 (12.92)	6.32 (4.61)
CSWA	MEWA	15.04 (15.15)	7.43 (3.04)
WA	MEWA	15.90	2.87
SESA	MEWA	20.73 (21.10)	7.08 (4.62)
EA	SE	20.87 (19.96)	5.73 (3.78)
CSWA	SE	22.04 (22.49)	7.00 (2.05)
WA	SE	24.08	1.59
SESA	SE	30.48 (31.18)	6.41 (3.43)
Between neighbouring regions around the Mediterranean area			
NA	MEWA	0.92	3.36
NA	SE	2.28	2.16
MEWA	SE	1.20	1.93

Table 4 AMOVA analyses of the level of genetic structure between (Φ_{CT}) and within (Φ_{SC}) geographic groups compared two by two.

[a]Codes for geographic groups are defined in Table S2 and Figure S2.
[b]Kimura-2P was used as a molecular distance, with a gamma correction of 0.4, a transition/transversion ratio of 10/1 and without considering indels. All Φ-statistics are given in percent, and were significant ($P < 0.005$).
[c]The resulting Φ-statistics for computations performed after excluding genetically divergent populations from each group (i.e. the two Hadza samples from the EA group, the Mbenzele from the CSWA group, and the Ju|'hoansi and Herero samples from SESA, see text) are given in brackets.

closest neighbours to the South were found differentiated from the former but not from the latter. The Nyangatom and the Turkana languages belong to the Teso-Turkana subgroup of the Nilo-Saharan family and are mutually intelligible, whereas the Daasanach speak a Cushitic language of the Afro-Asiatic family. Our results suggest that genetic exchanges (at least for women) are more common between the Daasanach and the Turkana than between these two populations and the Nyangatom.

We have additionally found that the Nyangatom and the Daasanach are also sharply differentiated from the other Ethiopian populations studied so far (which we have coined as "Northern Ethiopians" as most are located in the Ethiopian highlands), which are, in contrast, found to be genetically rather homogeneous. Instead, the Nyangatom and the Daasanach present genetic affinities with some Tanzanian populations, although these populations are presently separated by at least one thousand kilometres.

The general pattern of genetic relationships among East African populations observed in this study does neither match clearly with geography nor with linguistic classification, in keeping with the results of Wood et al. (2005) that evidenced a weak correlation of mtDNA variability with either language or geography at the continental level. Our results on mtDNA are also in agreement with the patterns of genetic differentiation observed for the GM immunoglobulin allotypes (Sanchez-Mazas & Poloni, 2008; Sanchez-Mazas et al., in preparation), although the comparison is restricted by differing samplings of populations available for the two genetic systems, since no Cushitic populations of Kenya or Tanzania, in particular, were represented in the GM dataset. As for mtDNA, the Nyangatom and the Daasanach are also

Table 5 Levels of genetic structure (Φ_{ST}) between populations and of genetic diversity (mean number of pairwise differences) within populations in African linguistic groups.

Group[a]	Number of populations in group	Φ_{ST}[b]	Average MNPD[c] ± standard deviation
Afro-Asiatic	48 (38)[d]	8.59 (8.76)[d]	5.5 ± 1.3 (5.6 ± 1.3)[d]
Afro-Asiatic > Berber/Semitic	34 (24)[d]	5.31 (5.34)[d]	5.1 ± 1.2 (5.1 ± 1.3)[d]
Afro-Asiatic > Cushitic	5	7.01	7.3 ± 0.8
Afro-Asiatic > Chadic	9	1.86	6.1 ± 0.6
Nilo-Saharan	5	3.86	7.6 ± 0.5
Niger-Congo	43 (41)[e]	10.15 (5.06)[e]	6.4 ± 1.1 (6.6 ± 0.8)[e]
Khoisan and related	6 (3)[f]	26.38 (21.57)[f]	5.4 ± 1.6 (5.7 ± 2.2)[f]

[a]Language groups are defined in Table S2 and Figure S2. For the Afro-Asiatic language family, we also considered primary subdivisions represented in the dataset, i.e. Afro-Asiatic > Berber/Semitic (speakers of Berber and/or Semitic languages), Afro-Asiatic > Cushitic (speakers of Cushitic languages), and Afro-Asiatic > Chadic (speakers of Chadic languages).
[b]Kimura-2P was used as a molecular distance, with a gamma correction of 0.4, a transition/transversion ratio of 10/1 and without considering indels. All Φ_{ST} values are given in percent, and were significant ($P < 0.001$).
[c]MNPD, mean number of pairwise differences between sequences, averaged over group.
[d]In brackets: computations based on populations from Africa only, i.e. without Afro-Asiatic populations from the Middle East.
[e]In brackets: computations based on 41 Niger-Congo populations, i.e. without the Mbenzele and Herero samples.
[f]In brackets: computations for the proper Khoisan group, i.e. based on the 3 Khoisan populations from Southern Africa.

Table 6 Genetic differentiation between Sub-Saharan African populations whose languages belong to the Afro-Asiatic, Nilo-Saharan and Niger-Congo linguistic stocks: AMOVA analyses of the level of genetic structure between (Φ_{CT}) and within (Φ_{SC}) linguistic groups compared two by two.

AMOVA design[a]		Φ-statistics[b]	
Group 1 versus	Group 2	Φ_{CT}	Φ_{SC}
Afro-Asiatic > Semitic	Afro-Asiatic > Cushitic	1.74 ($P = 0.026$)	3.94
Afro-Asiatic > Semitic	Afro-Asiatic > Chadic	2.91	1.70
Afro-Asiatic > Cushitic	Afro-Asiatic > Chadic	3.55	4.08
Afro-Asiatic > Semitic	Nilo-Saharan	2.36	2.77
Afro-Asiatic > Cushitic	Nilo-Saharan	*−0.64* (*P = 0.764*)	5.16
Afro-Asiatic > Chadic	Nilo-Saharan	2.53	2.79
Afro-Asiatic > Semitic	Niger-Congo	2.95	4.61
Afro-Asiatic > Cushitic	Niger-Congo	3.01	5.24
Afro-Asiatic > Chadic	Niger-Congo	*0.44* (*P = 0.100*)	4.68
Nilo-Saharan	Niger-Congo	2.40	4.86

[a]The AMOVA design was restricted to Sub-Saharan African populations (Table S2 and Figure S2). Thus, the Afro-Asiatic > Semitic group did not include North African and Middle Eastern populations. Also, the divergent Mbenzele and Herero samples were excluded from computations involving the Niger-Congo group.
[b]Kimura-2P was used as a molecular distance, with a gamma correction of 0.4, a transition/transversion ratio of 10/1 and without considering indels. All Φ-statistics are given in percent, and were significant ($P < 0.005$), unless specified by italics and associated *P*-value given in brackets.

significantly differentiated from each other for the GM system. Furthermore, the sharp contrast between the LORV populations and the "Northern Ethiopians" is also evidenced by this nuclear system, in that the latter are genetically close to populations speaking Semitic and Berber languages, whereas the Nyangatom and the Daasanach tend to cluster nearer to Sub-Saharan populations affiliated to the Niger-Congo and Nilo-Saharan phyla.

Evidence for Admixture Among East African Populations

The pattern observed in East Africa (with the exception of the Khoisan-related Hadza and Sandawe populations), which combines a high level of within-population diversity with strong genetic structure among populations, suggests the occurrence of periodical episodes of admixture in

these populations, separated by periods of isolation and genetic drift. Indeed, the observation of high levels of diversity within populations could be due to long-term large effective population sizes maintained in East Africa. In this case, however, little genetic structure between populations should be expected, since there would be little opportunity for genetic drift to act. Alternatively, gene flow can produce high within-population diversity, and in the present case, it could also account for the extensive sharing of haplotypes and haplogroups observed between the Nyangatom and the Daasanach, as well as with other populations.

The intermediate linkage disequilibrium (LD) found in East Africa (Tishkoff et al., 1996) in contrast with Europe (high LD) and Sub-Saharan Africa (low LD, Tishkoff & Kidd, 2004; Conrad et al., 2006), could be due to such admixture events, more frequently occurring in this region compared to other Sub-Saharan populations. Substantial levels of gene flow among Nilo-Saharan, Afro-Asiatic and Niger-Congo populations from Tanzania have already been inferred by Tishkoff et al. (2007a) and our results suggest that these gene flows could have occurred in a larger region extending up to Southern Ethiopia.

If gene flow was constrained by the distance separating present-day population locations, it would produce a geographically structured pattern of genetic differentiation (e.g. isolation-by-distance, Wright, 1943). If gene flow was constrained by a cultural barrier such as language, then a linguistically structured pattern of genetic differentiation would be expected. However, as stated above, neither one of these two patterns was found. This suggests that populations have moved in the past, so as to come into contact and experience gene flow (from the linguistic perspective, it is also possible that some populations have experienced language shifts, due to these contacts). Thus it is probable that the current geographic location of several East African populations reflects a rather recent situation.

Indirect evidence supports the hypothesis of extensive mobility and gene flows in East Africa. A shared origin of the Nyangatom with other populations (such as the Toposa, Turkana and Jiye) in the Nilo-Saharan Karimojong group in North Uganda has been postulated on the basis of ethnologic and linguistic evidence. From this earlier group, expansions would have occurred in the early 18th century, which would have brought pastoral populations into Sudan (Toposa), Kenya (Turkana), West Uganda (Jiye), and into Ethiopia (Nyangatom). Studies of the generational system do support such a recent ethnogenesis of the Nyangatom (Tornay, 2001). This example of extensive mobility of the pastoral groups in East Africa suggests that migrations might have been a significant factor promoting gene flow among groups.

More direct and independent evidence of gene flows is provided by ethno-linguistic studies of the Daasanach, which have led scholars to consider this population as a "conglomerate of peoples of different tribal affiliations" (Tosco, 2001, p. 10). Besides the presence, in the Daasanach language, of items considered to be of non-Cushitic origin, several other lines of evidence in the culture of this population have also fuelled this view (Tosco, 2001 and references therein). Among these is the fact that some of the tribal sections that compose the Daasanach society bear names that correspond to other populations (e.g. the Nilo-Saharan Samburu and the Afro-Asiatic Rendille) from which members are supposed to have joined the Daasanach. The hypothesis exists that the Daasanach was a Nilo-Saharan group, sharing a common origin with the Pokot, and would have migrated into its current settlement area as late as during the 19th century, where it would have shifted to its current language through the absorption of a Cushitic group into the population. The Pokot would have migrated in parallel towards the South, into Kenya and Uganda. Actually, rapid linguistic changes during the last four centuries through migrations and subsequent contacts, involving Nilo-Saharan, Afro-Asiatic and Niger-Congo languages, are documented for other East African populations (Nurse, 2000; several contributions in Nurse, 2001), which points to the fact that such events might have been very common in this area.

The pattern observed for the Nyangatom and the Daasanach thus seems to correspond to a general one in East Africa. The fact that this area harbours predominantly transhumant populations suggest that a high mobility could be the reason for the postulated gene flow between populations and the associated diversity found in these populations. Concomitantly, pastoral populations are assumed to have lower effective sizes than farmers (Cavalli-Sforza, 1996), which allows for genetic drift, and thus differentiation between populations, to occur. Our results therefore suggest that the history of East African populations is characterised by an alternation of contact events between highly mobile populations leading to gene flow, and periods of isolation favouring genetic drift and differentiation. Besides the Y-chromosome study of Henn et al. (2008), a high mobility of pastoralist populations could also be illustrated by the recent migration from East Africa towards Central-West Africa of present day Afro-Asiatic Chadic populations that Cerny et al. (2009) postulated on the basis of the distribution of mtDNA L3f haplogroups. Isolation might have been reinforced at times by cultural practices that encourage both strong group cohesion and opposition to other groups (Barth, 1969/1994), which could also be promoted by competition for access to grazing areas (Tornay, 2001). This marked group cohesion is thus in sharp contrast with the apparently plastic and transient ethnic composition of East African populations, as further evidenced by the lack of significant genetic differentiation (null Φ_{CT}) between Cushitic and Nilo-Saharan groups of populations in this region.

Inferences on East Africa History

A probable East African origin of modern humans is currently inferred from a variety of evidence among which are the higher diversity found there compared to other African regions, and the presence of old Y and mtDNA lineages in its populations. Accordingly a geographic pattern of decay has been found in the genetic diversity of autosomal microsatellites from East Africa towards the rest of the world (Prugnolle et al., 2005; Ramachandran et al., 2005). The high diversity in East Africa was interpreted as a sign of an ancient origin. However, our results might indicate that this high diversity could also come from a particular history of recent migrations and admixture promoted by the pastoralist societies that dominate in the region. The recent spread of a lactase-persistence allele (the C-14010 allele) in East Africa lends support to this hypothesis (Tishkoff et al., 2007b). According to current knowledge, pastoralism as a food producing way of life is probably not dated much further back than the Holocene, and the complexity of the ethnogenesis process might have been a common phenomenon since then. This is, however, in no way as ancient as the global history of our species, which lasts for at least 150,000 to 190,000 years, as attested from the East African fossil record (Clark et al., 2003; White et al., 2003; McDougall et al., 2005).

Addendum

While the present work was under review, a genome-wide study on the structure of African populations was published (Tishkoff et al., 2009). Similar to our own results, this study concluded that extensive gene flow has occurred among East African populations, particularly so between Nilo-Saharan and Afro-Asiatic Cushitic groups, and this within the last 5,000 years or so. Tishkoff et al. (2009) also refer to linguistic evidence for recent immigration of Nilo-Saharan, Afro-Asiatic and Niger-Congo groups in the area, a process that probably resulted in numerous language shifts among East African ethnic groups.

Acknowledgements

We wish to express our gratitude to Serge Tornay and François Rodhain, who besides putting the LORV serum collection as well as their field notes at our disposal, kindly provided also their time and expertise. Our thanks also go to Mathias Currat for helpful discussions on the project and critical reading of the manuscript, to Jean-Marie Tiercy for technical assistance in his lab at the University Hospital of Geneva, to Eric Huysecom and Anne Mayor for advice and to the Conservatory and Botanical Garden of Geneva for its support. We also wish to thank two anonymous reviewers for their constructive comments on a former version of the manuscript. This study was funded by the Swiss National Foundation grant 31-59375.99 to ASM.

References

Almagor, U. (1978) *Pastoral partners: affinity and bond partnership among the Dassanetch of South-West Ethiopia*. Manchester: Manchester University Press.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.

Barth, F. (1969/1994) Introduction. In: *Ethnic groups and boundaries: The social organization of culture difference* (ed. F. Barth). Oslo: Pensumtjeneste.

Belle, E. M. & Barbujani, G. (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *Am J Phys Anthropol* **133**, 1137–1146.

Bendall, K. E. & Sykes, B. C. (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet* **57**, 248–256.

Blanc, M., Sanchez-Mazas, A., Van Blyenburgh, N. H., Sevin, A., Pison, G. & Langaney, A. (1990) Interethnic genetic differentiation: GM polymorphism in eastern Senegal. *Am J Hum Genet* **46**, 383–392.

Blench, R. (2006) *Archaeology, language and the African past*. Lanham, MD: Altamira Press.

Buchanan, A. V., Risch, G. M., Robichaux, M., Sherry, S. T., Batzer, M. A. & Weiss, K. M. (2000) Long DOP-PCR of rare archival anthropological samples. *Hum Biol* **72**, 911–925.

Campbell, M. C. & Tishkoff, S. A. (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* **9**, 403–433.

Cavalli-Sforza, L. (1996) The spread of agriculture and nomadic pastoralism: insights from genetics, linguistics and archaeology. In: *The origins and spread of agriculture and pastoralism* (ed. D. R. Harris). London: UCL Press.

Cavalli-Sforza, L. L. & Feldman, M. W. (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet* **33**(Suppl), 266–275.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci U S A* **85**, 6002–6006.

Cerny, V., Fernandes, V., Costa, M. D., Hajek, M., Mulligan, C. J. & Pereira, L. (2009) Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. *BMC Evol Biol* **9**, 63.

Chaix, R., Austerlitz, F., Khegay, T., Jacquesson, S., Hammer, M. F., Heyer, E. & Quintana-Murci, L. (2004) The genetic or mythical ancestry of descent groups: lessons from the Y chromosome. *Am J Hum Genet* **75**, 1113–1116.

Chen, J., Sokal, R. R. & Ruhlen, M. (1995) Worldwide analysis of genetic and linguistic relationships of human populations. *Hum Biol* **67**, 595–612.

Clark, J. D., Beyene, Y., Woldegabriel, G., Hart, W. K., Renne, P. R., Gilbert, H., Defleur, A., Suwa, G., Katoh, S., Ludwig, K. R., Boisserie, J. R., Asfaw, B. & White, T. D. (2003) Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**, 747–752.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A. & Pritchard, J. K. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251–1260.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., Wallace, D. C., Oefner, P. J., Torroni, A., Cavalli-Sforza, L. L., Scozzari, R. & Underhill, P. A. (2002) 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**, 1197–1214.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G. & Capelli, C. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol* **21**, 1673–1682.

Ehret, C. (2002) Language family expansions: Broadening our understandings of cause from an African perspective. In: *Examining the farming/language dispersal hypothesis* (eds. P. Bellwood & C. Renfrew). Cambridge: McDonald Institute for Archaeological Research.

Excoffier, L. (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol* **13**, 853–864.

Excoffier, L., Harding, R. M., Sokal, R. R., Pellegrini, B. & Sanchez-Mazas, A. (1991) Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. *Hum Biol* **63**, 273–307.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50.

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

Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G., Zhadanov, S., Schurr, T. G. & Merriwether, D. A. (2007) Melanesian mtDNA complexity. *PLoS ONE* **2**, e248.

Fu, Y. X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Fukui, K. & Markakis, J. (1994) Introduction. In: *Ethnicity and conflict in the Horn of Africa* (eds. K. Fukui & J. Markakis). London: James Currey.

Gonder, M. K., Mortensen, H. M., Reed, F. A., De Sousa, A. & Tishkoff, S. A. (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* **24**, 757–768.

Gordon, R. G., Jr. (ed.) (2005) *Ethnologue: languages of the world*, Fifteenth edition, Dallas, TX, SIL International. Online version: http://www.ethnologue.com/.

Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**, 95–98.

Hassan, H. Y., Underhill, P. A., Cavalli-Sforza, L. L. & Ibrahim, M. E. (2008) Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. *Am J Phys Anthropol* **137**, 316–323.

Henn, B. M., Gignoux, C., Lin, A. A., Oefner, P. J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S. A., Mountain, J. L. & Underhill, P. A. (2008) Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A* **105**, 10693–10698.

Hurlbert, S. H. (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577–586.

Jobling, M. A., Hurles, M. E. & Tyler-Smith, C. (2004) *Human evolutionary genetics: origins, peoples and disease*. London, New York: Garland Science.

Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E. & Villems, R. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am J Hum Genet* **75**, 752–770.

Knight, A., Underhill, P. A., Mortensen, H. M., Zhivotovsky, L. A., Lin, A. A., Henn, B. M., Louis, D., Ruhlen, M. & Mountain, J. L. (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr Biol* **13**, 464–473.

Krings, M., Salem, A. E., Bauer, K., Geisert, H., Malek, A. K., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Utermann, G., Sajantila, A., Paabo, S. & Stoneking, M. (1999) mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration? *Am J Hum Genet* **64**, 1166–1176.

Langaney, A. & Gomila, J. (1973) Bedik and Niokholonko intra and inter-ethnic migration. *Hum Biol* **45**, 137–150.

Lansing, J. S., Cox, M. P., Downey, S. S., Gabler, B. M., Hallmark, B., Karafet, T. M., Norquest, P., Schoenfelder, J. W., Sudoyo, H., Watkins, J. C. & Hammer, M. F. (2007) Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci U S A* **104**, 16022–16026.

Lie, B. A., Dupuy, B. M., Spurkland, A., Fernandez-Vina, M. A., Hagelberg, E. & Thorsby, E. (2007) Molecular genetic studies of natives on Easter Island: evidence of an early European and Amerindian contribution to the Polynesian gene pool. *Tissue Antigens* **69**, 10–8.

Lovell, A., Moreau, C., Yotova, V., Xiao, F., Bourgeois, S., Gehl, D., Bertranpetit, J., Schurr, E. & Labuda, D. (2005) Ethiopia: between Sub-Saharan Africa and western Eurasia. *Ann Hum Genet* **69**, 275–287.

McDougall, I., Brown, F. H. & Fleagle, J. G. (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733–736.

Merriwether, D. A. (1999) Freezer anthropology: new uses for old blood. *Philos Trans R Soc Lond B Biol Sci* **354**, 121–129.

Militarev, A. (2002) The prehistory of a dispersal: the Proto-Afrasian (Afroasiatic) farming lexicon. In: *Examining the farming/language dispersal hypothesis* (eds. P. Bellwood & C. Renfrew). Cambridge: McDonald Institute for Archaeological Research.

Moseley, C. & Asher, R. E. (eds.) (1994) *Atlas of the world's languages*. London/New York, Routledge Reference.

Nurse, D. (2000) *Inheritance, contact, and change in two East African languages*. Köln: Rüdiger Köppe Verlag.

Nurse, D. (ed.) (2001) *Historical language contact in Africa*. Köln, Rüdiger Köppe Verlag.

Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M. & Santachiara-Benerecetti, A. S. (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* **62**, 420–434.

Peakall, R. & Smouse, P. E. (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288–295.

Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M. J. & Amorim, A. (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* **65**, 439–458.

Petit, R. J., El Mousadik, A. & Pons, O. (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation biology* **12**, 844–855.

Poloni, E. S., Sanchez-Mazas, A., Jacques, G. & Sagart, L. (2005) Comparing linguistic and genetic affinities among East Asian populations: a study of the Rh and Gm polymorphisms. In: *The peopling of East Asia: putting together archaeology, linguistics and genetics* (eds. L. Sagart, R. Blench & A. Sanchez-Mazas). London, New York: Routledge Curzon.

Poloni, E. S., Semino, O., Passarino, G., Santachiara-Benerecetti, A. S., Dupanloup, I., Langaney, A. & Excoffier, L. (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* **61**, 1015–1035.

Prugnolle, F., Manica, A. & Balloux, F. (2005) Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**, R159–R160.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., Mcelreavey, K. & Santachiara-Benerecetti, A. S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23**, 437–441.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**, 15942–15947.

Ray, N., Currat, M. & Excoffier, L. (2003) Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* **20**, 76–86.

Reynolds, J., Weir, B. S. & Cockerham, C. C. (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767–779.

Rodhain, F., Ardoin, P., Metselaar, D., Salmon, A. M. & Hannoun, C. (1975) An epidemiologic and serologic study of arboviruses in Lake Rudolf basin. *Trop Geogr Med* **27**, 307–312.

Rodhain, F., Hannoun, C. & Metselaar, D. (1972) Epidemiological and serological study of the arboviruses in the lower valley of the Omo (southern Ethiopia). *Bull World Health Organ* **47**, 295–304.

Rohlf, F. J. (2007) *NTSYSpc: Numerical Taxonomy System, ver. 2.10*. Setauket, NY: Exeter Publishing, Ltd.

Salas, A., Richards, M., De La Fe, T., Lareu, M. V., Sobrino, B., Sanchez-Diz, P., Macaulay, V. & Carracedo, A. (2002) The making of the African mtDNA landscape. *Am J Hum Genet* **71**, 1082–1111.

Sanchez-Mazas, A. & Poloni, E. S. (2008) Genetic diversity in Africa. In: *Encyclopedia of life sciences*. Chichester: John Wiley & Sons.

Sanchez-Mazas, A., Poloni, E. S., Jacques, G. & Sagart, L. (2005) HLA genetic diversity and linguistic variation in East Asia. In: *The peopling of East Asia: putting together archaeology, linguistics and genetics* (eds. L. Sagart, R. Blench & A. Sanchez-Mazas). London, New York: Routledge Curzon.

Sanchez-Mazas, A., Ries, F., Nunes, J. M., Dugoujon, J. M., Kervaire, B., Naciri, Y., Excoffier, L., Langaney, A., Poloni, E. S. & Tiercy, J. M. (in preparation) Classical and molecular markers typed on old sera reveal a close genetic relationship between Nilo-Saharans and Niger-Congo in sub-Saharan Africa.

Schneider, S. & Excoffier, L. (1999) Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.

Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L. & Underhill, P. A. (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* **70**, 265–268.

Sokal, R. R. (1988) Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci U S A* **85**, 1722–1726.

Swofford, D. L. (1991) *PAUP: phylogenetic analysis using parsimony, Macintosh version 3.0r*. Champaign, IL: Illinois Natural History Survey.

Thomas, M. G., Weale, M. E., Jones, A. L., Richards, M., Smith, A., Redhead, N., Torroni, A., Scozzari, R., Gratrix, F., Tarekegn, A., Wilson, J. F., Capelli, C., Bradman, N. & Goldstein, D. B. (2002) Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am J Hum Genet* **70,** 1411-1420.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P. & Krings, M. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387.

Tishkoff, S. A., Gonder, M. K., Henn, B. M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T. B., Ramakrishnan, U., Reed, F. A. & Mountain, J. L. (2007a) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol* **24**, 2180–2195.

Tishkoff, S. A. & Kidd, K. K. (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* **36**, S21–27.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A. & Deloukas, P. (2007b) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31–40

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L. & Williams, S. M. (2009) The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044.

Tornay, S. (2001) *Les Fusils jaunes. Générations et politique en pays nyangatom (Éthiopie)*. Paris-Nanterre: Société d'ethnologie.

Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., Smith, D. G., Vullo, C. M. & Wallace, D. C. (1993a) Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet* **53**, 563–590.

Torroni, A., Sukernik, R. I., Schurr, T. G., Starikorskaya, Y. B., Cabell, M. F., Crawford, M. H., Comuzzie, A. G. & Wallace, D. C. (1993b) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am J Hum Genet* **53**, 591–608.

Tosco, M. (2001) *The Dhaasanac Language. Grammar, texts and vocabulary of a Cushitic language of Ethiopia*. Köln: Rüdiger Köppe.

Watson, E., Bauer, K., Aman, R., Weiss, G., Von Haeseler, A. & Paabo, S. (1996) mtDNA sequence diversity in Africa. *Am J Hum Genet* **59**, 437–444.

Watson, E., Forster, P., Richards, M. & Bandelt, H. J. (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**, 691–704.

Weir, B. S. & Cockerham, C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Weiss, K. M., Buchanan, A. V., Daniel, C. & Stoneking, M. (1994) Optimizing utilization of DNA from rare or archival anthropological samples. *Hum Biol* **66**, 789–804.
White, T. D., Asfaw, B., Degusta, D., Gilbert, H., Richards, G. D., Suwa, G. & Howell, F. C. (2003) Pleistocene homo sapiens from Middle Awash, Ethiopia. *Nature* **423**, 742–747.
Wood, E. T., Stover, D. A., Ehret, C., Destro-Bisol, G., Spedini, G., Mcleod, H., Louie, L., Bamshad, M., Strassmann, B. I., Soodyall, H. & Hammer, M. F. (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* **13**, 867–876.
Wright, S. (1943) Isolation by Distance. *Genetics* **28**, 114–138.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Text S1 Supporting Information on Materials and Methods and on Results. References in Supporting Texts, Tables and Figures.
Figure S1 Median joining networks of mitochondrial DNA haplotypes, modified from Kivisild *et al.* (2004), so as to include the Nyangatom, Daasanach and Turkana sequences from the LORV sample.
Figure S2 Map with approximate geographic location of the 136 population samples included in the HVS-I database.
Figure S3 Three successive Principal Coordinates Analyses (PCA) of pairwise Reynolds' genetic distances among populations computed from HVS-I sequence-based ΦST indices.
Table S1 Scored mutated positions in the LORV sequences and classification into haplogroups.
Table S2 The 136 published and new population samples of the mtDNA HVS-I sequences database used in this study (from Africa, the Middle East and West Asia, and the South of Europe).
Table S3 TMRCA estimates for sub-haplogroups observed in the LORV samples that included at least five sequences.
Table S4 HVS-I genetic diversity indices, results of the tests of selective neutrality and population equilibrium, and results of the tests of population expansion performed on the 136 population samples of the HVS-I database.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 22 April 2009
Accepted: 17 July 2009