Meta Gene 27 (2021) 100837

Contents lists available at ScienceDirect

Meta Gene

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

Mitochondrial HVRI and whole mitogenome sequence variations portray similar scenarios on the genetic structure and ancestry of northeast Africans

Maha M. Osman [a,b], Hisham Y. Hassan [a,1], Mohammed A. Elnour [a,1], Heeran Makkan [c], Eyoab Iyasu Gebremeskel [a,d], Thoyba Gais [a], Mahmoud E. Koko [a], Himla Soodyall [c], Muntaser E. Ibrahim [a,*]

[a] Institute of Endemic Diseases, University of Khartoum, Sudan
[b] Commission for Biotechnology and Genetic Engineering, National Center for Research, Sudan
[c] Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, National Health Laboratory Service and the University of the Witwatersrand, Johannesburg, South Africa
[d] Eritrea Institute of Technology, Mai-Nefhi, Eritrea

ARTICLE INFO

Keywords:
Mitochondrial DNA
mtDNA control region
Sudan
Haplogroups
Linguistic groups

ABSTRACT

A significant portion of the current human adaptive and demographic traits is believed to have originated in north-east Africa, the putative scene of early human evolution. However interesting such assumption might be, the genetic structure and phenotypic traits of populations in this part of the globe remains poorly studied, including the widely analyzed genome of the mitochondria. Mitochondrial hypervariable region and whole mitogenomes of Sudanese and South Sudanese were compared to regional and global sequences. Haplotypes, mismatch distribution, PCAs, F_{ST}, N_e, and phylogenetic trees were constructed and analyzed. The HVRI in particular and whole mitogenomes produced robust phylogenies that were greatly concordant. Observed haplotypes were found to belong mainly to the major mitochondrial macrohapolgroups L0, L1, L2, L4, L5, L3, M and N. Expectedly the L0 was confined to populations previously shown to occupy the deepest and most ancestral lineages in the human evolutionary tree. The observed regional variation and diversity, depicted in various metrics imply that the female lineages in this part of Africa are likely to have been shaped by a longer history of in-situ evolution.

1. Introduction

Both paleobiological and archaeological data indicate that modern humans may have originated in eastern Africa, with varying estimates in the range of 40–100,000 years ago (Mcbrearty and Brooks, 2000; McDougall et al., 2005; Tishkoff et al., 2009; Elhassan et al., 2014). Furthermore, the earliest migrations of modern humans out of Africa are equally shown to may have originated from eastern Africa(Elhassan et al., 2014; Tishkoff et al., 1996; Quintana-Murci et al., 1999; Kivisild et al., 2004). Sudan and South Sudan are in the heart of this historical milieu being endowed with some of the most spectacular cultural diversity in Africa, comprising approximately 90 discrete ethnic groups who collectively speak over 100 languages (Gordon, 2005). Many of these groups extend across to countries neighboring Sudan and southern Sudan often rendering the concept of assignment based on contemporary political borders somewhat irrelevant. Several questions pertaining to the pattern of succession of the different groups and cultures in early Sudan have been raised with the hope of acquiring clues into the history, state formation and main demographic and migration events in the country and the region. The overwhelming emphasis so far has been on the history and archaeology of the Nile Valley. Following the progressive desertification of the Sahara towards the end of the last glacial maxima around 10,000 BCE, the Nile has become the most important path not only in the north-south migrations (Fox, 1997), but also the east-west axis across the Sahel and Sahara which hitherto remain largely understudied.

Estimating the effective size of a population (N_e) is largely dependent on the amount and pattern of DNA sequence variation, decided by the interactions among evolutionary forces, as well as the population coalescence time (Elhassan et al., 2014; Wall, 2003; Gasca-Pineda et al.,

* Corresponding author at: Institute of Endemic Diseases, University of Khartoum, 11111 Khartoum, Sudan.
E-mail address: mibrahim@iend.org (M.E. Ibrahim).
[1] Contributed equally as first authors.

https://doi.org/10.1016/j.mgene.2020.100837
Received 27 August 2020; Received in revised form 19 November 2020; Accepted 2 December 2020
Available online 10 December 2020
2214-5400/© 2020 Elsevier B.V. All rights reserved.

2013).

We have recently reported east African populations to possess the highest N_e worldwide; based on MT-CO2, and autosomal microsatellite markers (Elhassan et al., 2014), a robust outcome for two different classes of markers.

The mitochondrial genome (Soares et al., 2009; Rito et al., 2013) in itself displays a wide range of sequence variation patterns between its different regions. The hypervariable control region displays one of the highest known mutation rate, and is still among the most useful tools in assessing recent episodes of population divergence within the context of the female lineages, providing enough information to reconstruct the demographic history even of closely related human populations (González-martín et al., 2015). The presence of ancient haplogroup lineages like L0a, L0f, L5 accompanied by a large long-term effective population size and/or large degree of long-term population structure suggested eastern Africa might be the source of origin of many other African mtDNA haplogroup lineages (Kivisild et al., 2004; Gonder et al., 2007; Watson et al., 1997; Chen et al., 2000).

With the availability of global samples of complete mtDNA genome sequences, it is now possible to precisely make phylogenetic inferences from mtDNA haplogroups (Ingman et al., 2000; Torroni et al., 2001; Ingman and Gyllensten, 2003; Mishmar et al., 2004; Ruiz-Pesini, 2004; Macaulay, 2005) of previously non-represented population and make better insights on early modern human dispersals(Gonder et al., 2007). Deciphering mtDNA haplogroups of populations living at the epicenter of these ancient human dispersals makes the endeavor more interesting and is also key to understanding the prehistoric genetic landscape of Africa in general and east Africa in particular.

In this study we employ mtDNA sequences from the control region (non-coding hypervariable region (HVR)) as well as whole mitogenomes of Sudanese, South Sudanese and other regional and global populations in an attempt to examine how females have contributed to shaping the gene pool in Sudan within a wider context of human demography and cultural evolution in the region.

2. Materials and methods

2.1. Samples and mitochondrial control regions sequencing

A total of 164 mitochondrial DNA samples from maternally unrelated individuals sampled by Hassan et al. (Hassan et al., 2008), belonging to twelve different ethnic groups in Sudan and South Sudan (for details see Table 4 and Fig. S1) were amplified. HVRI and HVRII were amplified as one fragment using primers L15996 and H408 Vigilant et al., 1989) DNA extraction and purification was performed according to the manufacturer instructions (NucleoSpin®, MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany). HVRI and HVRII PCR products were sequenced with the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, CA, USA) using primer L15996(Vigilant et al., 1989). Purification was carried out with QIAGEN DyeEx 2.0 Spin Kit (QIAGEN®, Hilden, Germany). Sequencing reactions were resolved on ABI 3700 Genetic Analyzer (Applied Biosystems, CA, USA). For those sequences containing a homopolymeric cytosine stretch, reverse sequencing was performed using primers H16401(Vigilant et al., 1989). Sequences were deposited in NCBI gene bank [KC764460 - KC764910].

A total of 9 whole mitogenomes sequences from Sudanese had been extracted from whole genome and exome sequences, compared to deeply ancestral sequences selected from mtDB -Human Mitochondrial Genome Database (http://www.mtdb.igp.uu.se/)- which include Pygmy (Mbuti and Biaka), San, Hausa, Khwei, Ethiopian, and Fulani. Chimps and Neanderthals were used as an outgroup.

2.2. Bioinformatics and sequence analysis

DNA Sequences were aligned using the Clustal W algorithm (Thompson et al., 1994) implemented in BioEdit(Hall, 1999). The sequences were assigned into different mitochondrial haplogroups (supplementary **Table S1**), following the nomenclatures previously published by multiple authors(Quintana-Murci et al., 1999; Kivisild et al., 2004; Chen et al., 2000; Huoponen et al., 1996; Macaulay et al., 1999; Bandelt et al., 2001; Underhill et al., 2001; Salas et al., 2002; Behar et al., 2008).

Whole mitochondrial genome sequences of 9 samples were extracted and aligned to rCRS from whole genome and exome sequences using the software MitoSeek (Guo et al., 2013) from samples sequenced on IlluminaHiSeq platform at Beijing Genomic Institute (BGI, Hong Kong). Variants were called using Freebayes v0.9.21-19-gc003c1e (Garrison and Marth, 2012).

2.3. Phylogenetic analyses and Bayesian skyline plot

Bayesian evolutionary phylogenetic trees with a relaxed molecular clock were constructed using Markov chain Monte Carlo (MCMC) algorithms implemented in BEAST v1.8.0 package(Drummond et al., 2012). The best fitting substitution models for each region has been estimated using MEGA v5.05.(Tamura et al., 2011). TN93 as a substitution model (Hazkani-Covo and Graur, 2007) is a suitable one for whole mitochondrial genomes, while KHY model was used for different selected genes (control regions, HVRI and HVR2) and (coding regions include; MT-CO1, MT-CO2, ND2). Trees were rooted using a chimpanzee mtDNA sequences as an outgroup in addition to Neanderthal.

The estimated effective population size based on each of control non-coding regions, whole mitochondrial genome, and coding regions using the mutation rates as reported in Soares et al., (Soares et al., 2009), 1.6×10^{-7}, 2.4×10^{-8}, and 3.6×10^{-8} substitution per site per year, respectively. In order to assess past changes in female effective population size (N_e) based on HVRI control region, we used a Bayesian Skyline coalescent tree prior with 10 groups under a piecewise-constant model. The analysis was run for 10 million generations with parameters logged every 1000 generations, and Tracer 1.6 (http://tree.bio.ed.ac.uk/software/ tracer/) was used to inspect chain convergence and conduct the skyline plot construction.

2.4. Haplotypes frequencies and median-joining network

To portray haplotypes variation and frequencies of mitochondrial haplogroups among Sudanese populations, sequences of mtDNA HVRI and HVRII were used to draw a median-joining network (Bandelt et al., 1999) for the different haplogroups using Network 4.6.11 software (available at http://www.fluxus-engineering.com). Arlequin version3.11. (Excoffier et al., 2005) was used to calculate haplotype and nucleotide diversity, F_{ST} distances between pairs of populations, and analysis of molecular variance (AMOVA).

2.5. Mismatch distribution and PCA

Mismatch distributions for Sudanese population were computed using Arlequin 3.11 for Sudanese population after pooling small populations with sample sizes less than 4 into larger groups according to their linguistic affiliation.

To assess the genetic affinities among Sudanese population and other global populations an HVRI PCA was plotted based on F_{ST} matrices using PAST software (Kot and Daniel, 2008). (available at http://folk.uio.no/ohammer/past).

3. Results

3.1. Haplogroups and haplotype frequencies in Sudanese population

A total of 149 haplotypes were found in the Sudanese sequences of the control region. The number of shared haplotypes between different populations was different (Table 1A and B). The observed haplogroups

and their frequencies for each haplotype observed in Sudanese are given in (Fig. S1) and supplementary Table 1 (Table S1). The haplotypes were found to belong to the major mitochondrial macrohapolgroups L0, L1, L2, L4, L5, L3, M and N. The L0 macrohaplogroup was confined to populations occupying the deepest and most ancestral lineages in the human evolutionary tree particularly Nilo-Saharan speaking groups who appear to display the highest frequency of haplogroup L0. The number of haplotypes was generally high for each sampled population consistent with a high effective size of populations in this part of the world. Although few populations had markedly higher number of haplotypes; haplotype sharing was insightful to the population structure and history. Hausa for example had large number of haplotypes but they tend to share some with the Fulani an expected outcome of a documented recent common history. Hausa and Beja had the highest number of haplotypes shared with other populations indicating a central role of this group in the history of the Sudan and the region (Table 1: A and B).

3.2. Inference on evolutionary history from Mitogenome and partial mtDNA sequences

The corollary in Inference of evolutionary history from mitogenome and partial mtDNA sequences and its robustness was verified in the current set of data. The patterns in the tree topology and clustering of the HVRI and whole mitochondrial genome using different methods and algorithms were greatly concordant (Fig. 1) and supplementary (Fig. S2), indicating that the HVRI has the upper hand in depicting the overall information content of the mtDNA sequence variation expectedly due to the amount of polymorphic sequences embedded in the HVR and the relative lack of evolutionary constrains (Figs. 1). Generally speaking some genes in the coding region sequences like MT-CO1 and ND2 were less informative than the MT-CO2, as shown in supplementary Figure (Fig. S3 (a), (b), (c), and (d)). In both Bayesian and maximum likelihood analysis, two main clusters were resolved with Chimp and Neanderthal sequences as outgroups, both clusters contained member population of northeast Africa in addition to SAN and or Pygmies and Fulani/Hausa (Fig. S2).

A phylogenetic tree constructed using BEAST, employing HVRI sequences, attested to the above but furthermore showed few sequences namely: D04 (Dinka- Nilo-Saharan), DM4 and G02 (Jamoeya and Gaalien, both Afro-Asiatic speakers) diverging earlier from the common ancestor and represents Haplogroup L5c compared to those who had deep ancestral history from mtDB - Human Mitochondrial Genome Database. These samples, however, were not available for further comparison with the mitogenome due to the scanty amount of DNA recovered upon sampling (Fig. 1).

The whole mitogenome sequences result showed, samples to cluster according to their geographic areas, the earliest cluster contains the majority of Africans including some of the Sudanese.

Skyline plots based on HVRI control region show expansion dates to alter when the north eastern sequences (Sudanese) were included (Fig. 2 (A)) as compared to the plot of the world population (Fig. 2B), indicating the pivotal contribution of east Africans to human key evolutionary events. We conclude that the HVRI sequences in east Africans reflects not only the large effective size (N_e) of this group as attested by the sequence and haplotype variation but also in most cases the ancestral position of these sequences.

3.3. Population structure based on median joining networks and principal component analysis (PCA)

Median Joining Networks supplementary Figures (Fig. S4 (a) L0, (b) L1 (c) L2 (d) L3/M/N (e) L4/L5) were constructed for each macro-haplogroup observed in major Sudanese populations. Expectedly the L0 macro-haplogroup was confined to populations like Nilotics, Beja and Nuba previously found to occupy the deepest and most ancestral lineages in the human evolutionary tree. The same applies to L5 which is recognized as one of the early clades branching from this tree. The L1 has a limited distribution almost exclusively confined to closely related groups like Hausa and Fulani (in addition to Nilotics and Meseria) all except Nilotics presently identified as recent migrants from West Africa, suggesting that this macro-haplogroup might particularly be associated with a westward dimension of the Sahel and great Sahara.

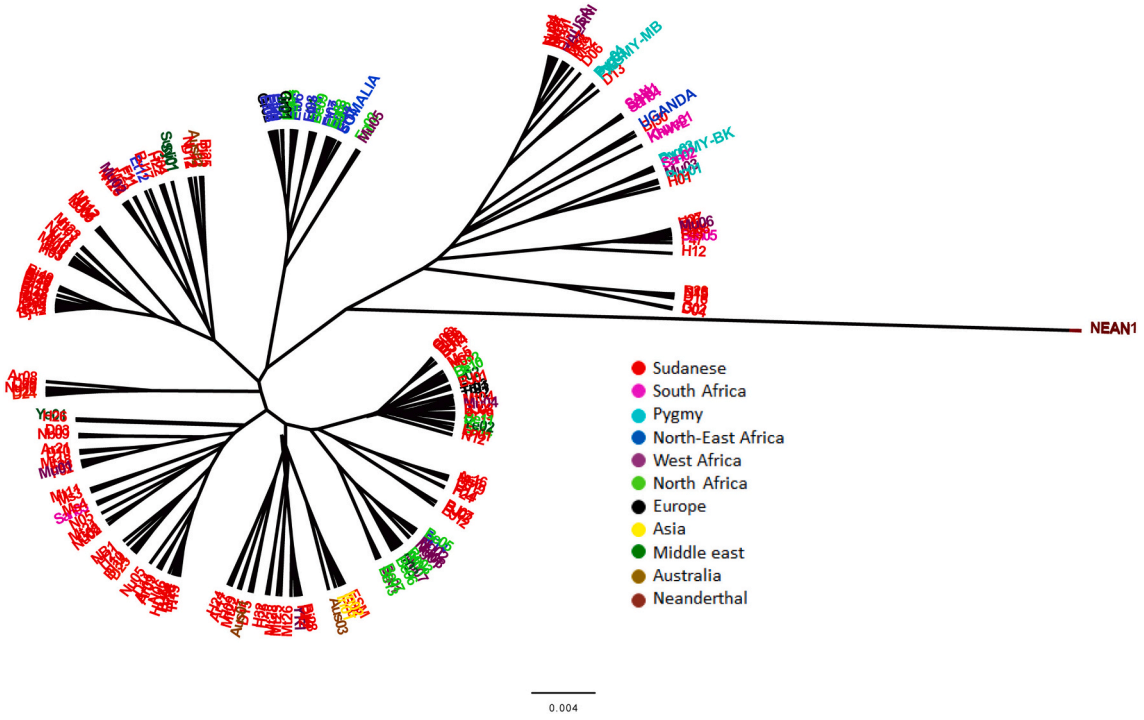

Fig. 1. Phylogenetic tree based on MCMC algorithm has been constructed from control region-*HVRI* sequences, relating different ancestral African genomes. Neanderthal (NEAN) used as an outgroup.

Table 1
Shared haplogroups and haplotypes between different Sudanese groups are represented by colors for each mtDNA control region, HVRI (Table 1(A)) and HVRII (Table 1(B)). A number of low frequency HVR unique haplogroups are not represented here that include: J1b, J2b, K, L0a, L0a2, L0f2, L1b, L1c2, L1c3, L 2a2, L2b, L2b1, L2c, L2c2, L2d, L2e, L3d, L3e1, L3e2b, L3e4, L3h, L3h1,L3h2, L3x2, L5a, L5c, M-M10, M1, L1a, L1b, T1, U3, and U3a. Although haplogroup L2a1 has the largest number of haplotypes (17), only one haplotype was shared between the Afro- Asiatic; Ar (1 Arakien), Bj (Beja), H (Hausa), and Nilo-Saharan Mt. (Masalit). Beja and Nilotics showed the highest number of the HVRII shared haplotypes.

Haplogroup	Haplotypes	Groups / Shared haplotypes	Ar (N=12)	Bj (N=42)	F (N=12)	G (N=8)	H (N=21)	Ms (N=7)	Mt (N=20)	N (N=29)	Nb (N=4)	Nu (N=9)
L0a1	6	I										
		II										
L1b1	4											
L2a	7											
L2a1	17											
L3b	3											
L3f	3											
L3f1b	4	I										
		II										
L4b2	3											
L5a2	3											
M	6											
R0	2											
H	4											

Haplogroup	Haplotypes	Groups / Shared Haplotypes	Ar (N=12)	Bj (N=42)	F (N=12)	G (N=8)	H (N=21)	Ms (N=7)	Mt (N=20)	N (N=29)	Nb (N=4)	Nu (N=9)
L0a1	5	I										
		II										
		III										
L2a	6											
L2a1	4	I										
		II										
L2a2	3											
L3b	2	I										
		II										
L3d	5											
L3f	2											
L3f1b	4											
L4b2	3											
M	6											
R0	5											
H	4											
J1b	2											

The first component of the PCA plot (Fig. 3) which represent 65.3% of the variation differentiates between Sudanese and South Africans, Pygmy, Australia, possibly denoting a deep genealogical event that occurred early in human history. The second component which explains 13.9% of variations, on the other hand, differentiated the Sudanese among themselves and from Northeast Africans, Europeans, Middle Eastern and West Africans. This may account for demographic events that are yet to be fully explained possibly in the expansion of early Afro-Asiatic speakers or agro-pastoralists.

3.4. Genetic diversity and neutrality tests

Analysis of Molecular variance (AMOVA) shows that when populations are grouped according to linguistic affiliation the level of maternal genetic among-group variation (FCT = 0.02257) is low but still higher than when populations are grouped according to geographic location (FCT = 0.00653). The low values of variation among geographic and linguistic groups indicate that both geography and language have no significant role in shaping the genetic structure of the Sudanese females unlike previous Y-chromosome analysis(Hassan et al., 2008) which showed high genetic – geographic –linguistic correlates between and within groups. AMOVA results for the whole dataset are given in (Table 2).

F_{ST} for the control region is 0.04273 and 0.03672, when populations are placed in linguistic and geographic groups, respectively. When populations are grouped according to linguistic and geographic affiliation; the proportion of among populations within group variance (FSC) of the control region sequenced is 0.02063 and 0.03039, respectively. FSC values are slightly higher than when populations are grouped according to geographic regions, indicating a fairly high degree of haplotype sharing and that mitochondrial DNA do in fact exhibit some subtle clustering on the basis of languages and geography. This may be partly due to the fact that most of the immigrant Afro-Asiatic speaking groups have lost their mtDNA due to genetic drift as suggested by results. All populations, except Nuba and Nubians, scored negative Tajima's D values (Table 3); all with non-significant 0.05 levels. These values were consistent with an excess of low-frequency mutations characteristic of a large population size (see discussion section).

3.5. Pairwise mismatch distribution

Pairwise mismatch distributions (Fig. 4) plots are not significantly different in all Sudanese populations from values expected under a model of population expansion (Rogers, 1995), or an excess of low-frequency mutations, reflected in a negative values Tajima's D. All populations showed smooth distribution with no significant difference between the observed and simulated distribution an evidence of normal population expansion. Nuba is excluded from mismatch analysis due to the small sample size.

4. Discussion

Interestingly, the presence of high frequencies of sub-clades of haplogroup L0 among certain Sudanese groups may represent a relic of a former wider distribution of these sub-haplogroups in this part of East Africa and may suggest that some Sudanese populations do relate to the

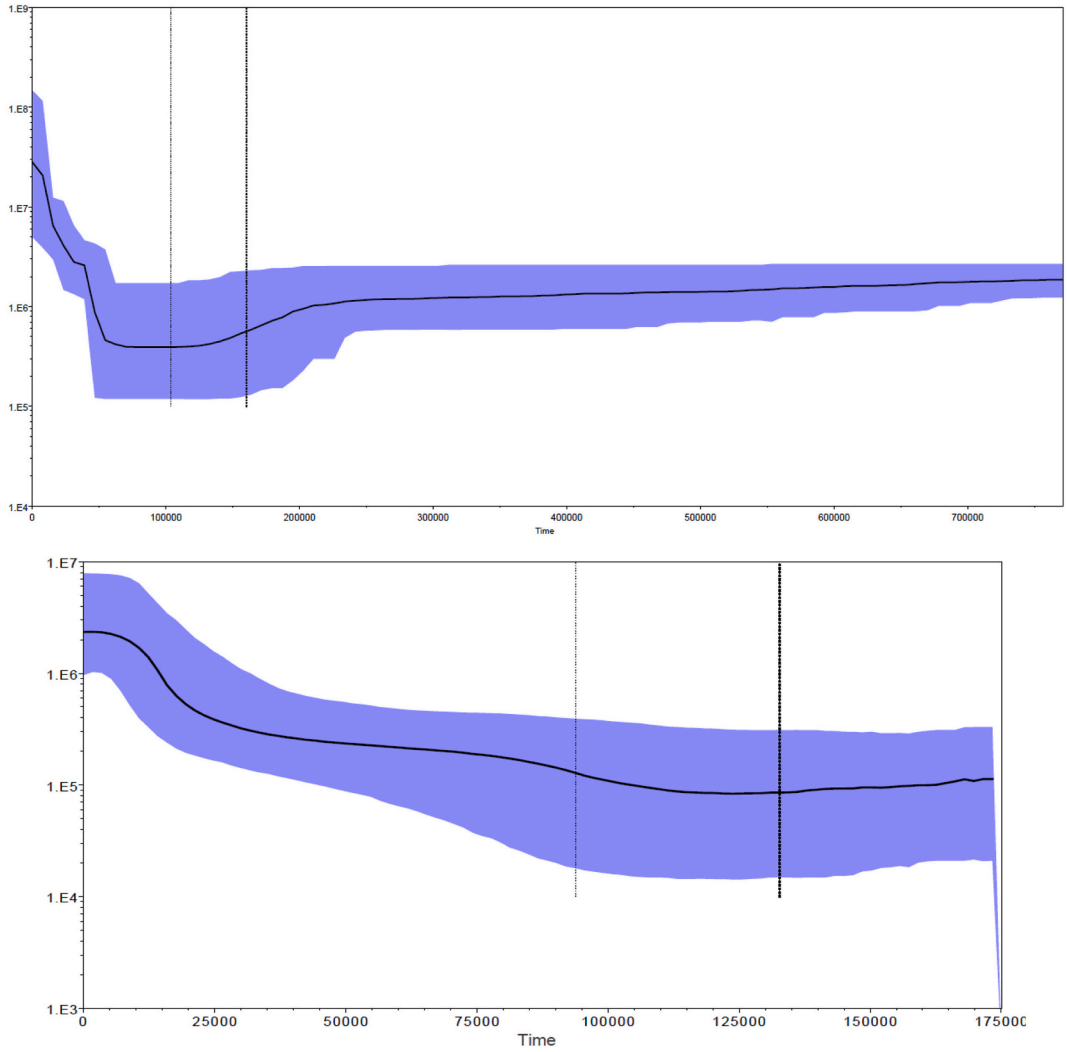

Fig. 2. (3A and 3B) Bayesian Skyline Plot (BSP). BSP based on mitochondrial DNA control HVRI region. The graph was constructed merging all populations as global population including east African (Sudanese), (Fig. 2A) as compared to the plot of the world population (Fig. 2B). The plot displays changes in world female effective population size (N_e) through time, a 1.6×10^{-7} sub/site/year rate.

deepest lineages of the mtDNA family tree. This is in Tally with evidence by coding sequence of the mtDNA (Elhassan et al., 2014) and Y-chromosome analysis has demonstrated that Khwe and San of South Africa as well as Nilotics and Nuba of East Africa display some of the deepest Y-chromosome clades of the human family tree (Hassan et al., 2008; Rogers, 1995; Underhill et al., 2000; Underhill et al., 2001; Cruciani et al., 2002; Semino et al., 2002; Knight et al., 2003), and equally manifested in a microsatellite genome-wide analysis (Tishkoff et al., 2009).

Particular groups such as Nubians, Nuba, Nilotics, and Beja are known to represent a continuum of settlement in the area known today as Sudan. Albeit the continuity of these groups that is evident from phylogenetic and cladistic analysis shown in the networks, the nature, and sequel of subsequent demographic and genetic events is not entirely clear, and the frequency of various mtDNA haplogroup in populations not included in the sample is yet to be determined.

The L2 and L3 clades attested by a plethora of branches and haplotypes might have coincided with a major expansion of our species possibly taking place around 50,000 YBP(Elhassan et al., 2014). The L4/L5 which is quite low in frequency almost restricted to Nilo-Saharan speaker particularly Nilotic groups, with the exception of two samples/haplotypes of Gaalien and Beja. The former currently Arabic speaking group has been shown by Y chromosome to possess a comparatively higher frequency of M13 suggesting a common history at some point with Nilotics (Hassan et al., 2008).

The position of L3/M and N haplogroups in different sub-clades may suggest this macrohaplogroup have originated in east Africa prior to the migration of the first *Homo sapiens* to Asia. The L3/M/N conventionally associated with the mitochondrial out-of-Africa scenario is widely distributed in Sudanese groups unknown of having history of migration or admixture with Asian or European groups, consistent with our suggestion that the major mitochondrial haplotype differentiation occurred well prior to the exodus event (Elhassan et al., 2014), then propagating the N carriers through migration and drift. For example the presence of Haplogroups R0 in populations like Beja, Arabs Nubians and Masalit, could hardly be taken as a result of migration from Asia to East Africa as suggested by similar presence in the Arabian peninsula (Abu-Amero et al., 2008; Gandini et al., 2016), given the cultural, genetic ethnic and demographic background of these groups. Some authors argue that this haplogroup, as well as other Eurasian haplogroups including U3 and T1 among Beja, may have been brought to Sudan by two routes one from North Africa, and the other through the present Ethiopia and Eritrea (Kivisild et al., 2004; Krings et al., 1999). On the other hand, an African origin of these macro haplogroups has been argued for in the literature (Quintana-Murci et al., 1999). The latter argument may find support from the analysis of slowly evolving genes of the coding region where

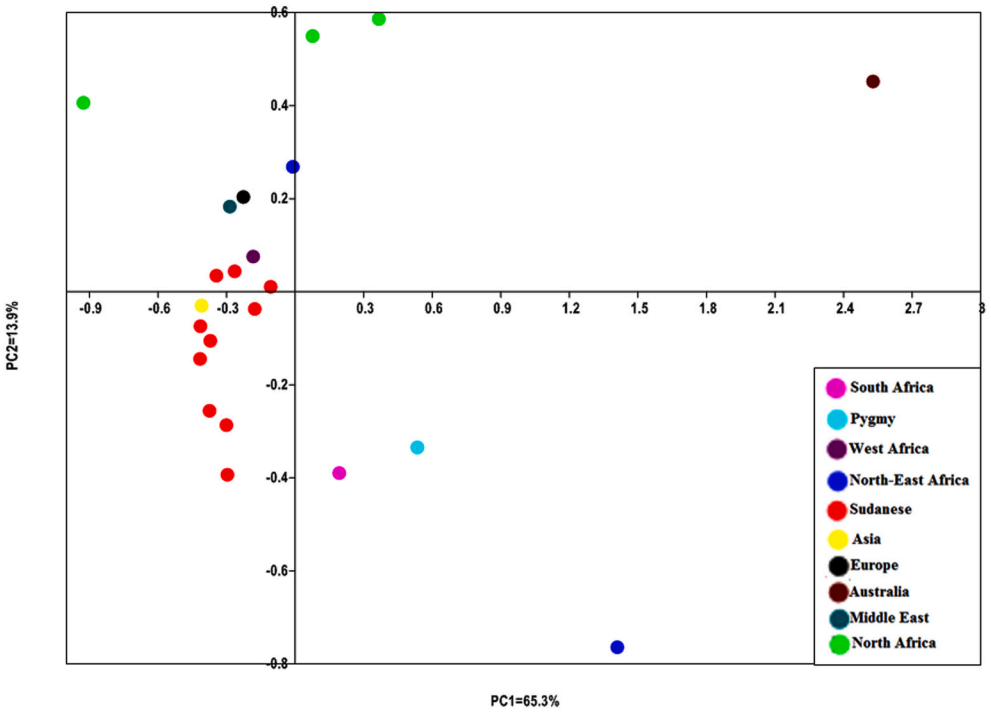

Fig. 3. PCA plot, the first component of the PCA which represent 65.3% of the variation differentiates between Sudanese and South Africans, Pygmy, Australia, possibly denoting a deep genealogical event that occurred early in human history. The second component which explains 13.9% of variations, on the other hand, differentiated the Sudanese among themselves and from Northeast Africans, Europeans, Middle Eastern and West Africans.

Table 2
Analysis of Molecular Variance (AMOVA) for Sudanese populations, within and among groups based on mtDNA HVR in relation to linguistic and geographic differences.

Groups	No. of groups	Within populations		Among populations within groups		Among groups	
		Variance (%)	F_{ST}	Variance (%)	F_{SC}	Variance (%)	F_{CT}
Linguistic groups	4	95.73	0.04273	2.02	0.02063	2.26	0.02257
Geographic groups	4	96.33	0.03672	3.02	0.03039	0.65	0.00653

P values: Vc and FST: *P*-value = 0.00000+ Vb and FSC:P-value = 0.01369 + −0.00441; Va and FCT: P-value = 0.00978.
Linguistic groups include: Afro-Asiatic, Nilo-Saharan, Niger-Kordofanian.
Geographic groups include: Northern Sudan, Eastern Sudan, Western Sudan and Southern Sudan.

Table 3
Measures of Genetic Diversity population size and neutrality among 10 Sudanese populations from the major linguistic groups estimated from mtDNA HVR sequence data.

Linguistic groups	Populations	N	HP	θ_S(SD)	D (*P*)
Afro-Asiatic	Arakien	12	11	12.60 (5.64)	– 0.90(0.19)
	Gaalien	8	7	13.97 (7.02)	– 1.08 (0.16)
	Meseria	7	7	18.79 (9.71)	– 0.53 (0.33)
	Beja	42	29	16.69 (5.35)	– 1.21 (0.10)
	Hausa	21	21	18.14 (6.73)	– 1.11 (0.13)
Nilo-Saharan	Masalit	20	20	18.38 (6.97)	– 1.18 (0.11)
	Nilotics	29	29	19.58 (6.75)	– 1.13 (0.12)
	Nubians	9	9	15.08 (7.21)	0.13 (0.59)
Niger-Congo	Fulani	12	12	12.46 (5.58)	– 0.14 (0.48)
NS + NC	Nuba	4	4	17.54 (11.62)	0.02 (0.66)

HP, number of observed haplotypes; N, sample size; SD, standard deviation; D, Tajima's D; P, P value for D. NS, Nilo-Saharan.NC, Niger-Congo.
HP, number of observed haplotypes; N, sample size; SD, standard deviation; D, Tajima's D; *P*, *P* value for D.

the Sudanese, Eritreans and other east Africans have been shown to be the source of demographic expansion that took people within Africa and beyond (Elhassan et al., 2014).

The difference in evolutionary rates of genes and sequences within the mitochondrial genome has been documented in other systems (Barker et al., 2012). We affirm such differences by analysis of the hypervariable region in addition to various coding genes. The HVRI was found to be highly correlated with mitogenome phylogeny, hence mitogenome cladistic based analysis will most likely reflect on average the evolutionary time scale of the HVRI, and HVRI based analysis may depict a reliable representation of the mitogenome phylogeny.

Furthermore, we suspect that several of the inconsistencies and confusions within the human evolutionary tree emanates from the dearth of sampling from this critical geographic area in human history. A settlement of several dangling key questions pertaining to the human origin and evolution is incumbent upon the analysis of genetic data including whole mitochondrial sequences from populations of this region.

Generally speaking, Sudanese populations show unimodal distributions of pairwise differences in their mismatch plots. Unimodal distributions, are interpreted as signs of demographic expansions, although it may also be due to other processes like population substructure, mutation rate heterogeneity (Rodríguez-Serrano et al., 2006; Marjoram and Donnelly, 1994), selection or a high level of homoplasmy that may produce the same pattern, because both reduce the correlation between sequences(Excoffier, 1990; Lundstrom et al., 1992), or from range expansion with high levels of migration between neighboring demes

Table 4
Groups, subgroups, socio-economic background linguistic affiliation and geographic location of the Sudanese populations included in this study.

Populations		Socio-economical Activities	Geographic locations	Linguistic affiliation[a]	
Groups	Population abbreviation			Family	Level
Nilotics	Ni	Pastoralists + Agri-pastoralists	South	Nilo-Saharan	Eastern Sudanic
Nubians	Nu	Agriculturists	North	Nilo-Saharan	Eastern Sudanic
Masalit	Mt	Agriculturists	West	Nilo-Saharan	Maban
Fulani	Fu	Nomadic Pastoralists	The Sahel	Niger- Kordofanian	Atlantic
Hausa	Ha	Agriculturists	Central	Afro-Asiatic	Chadic
Nuba	Nb	Agri-pastoralists	South	Nilo-Saharan + Niger- Kordofanian	Eastern Sudanic + Kordofanian
Gaalien	Ga	Agriculturists	Central	Afro-Asiatic	Semitic
Meseria	Ms	Nomadic Pastoralists	West	Afro-Asiatic	Semitic
Arakien	Ar	Agriculturists	Central	Afro-Asiatic	Semitic
Beja	Bj	Pastoralists	East	Afro-Asiatic	Cushitic

[a] According to Ethnologue. www.ethnologue.com

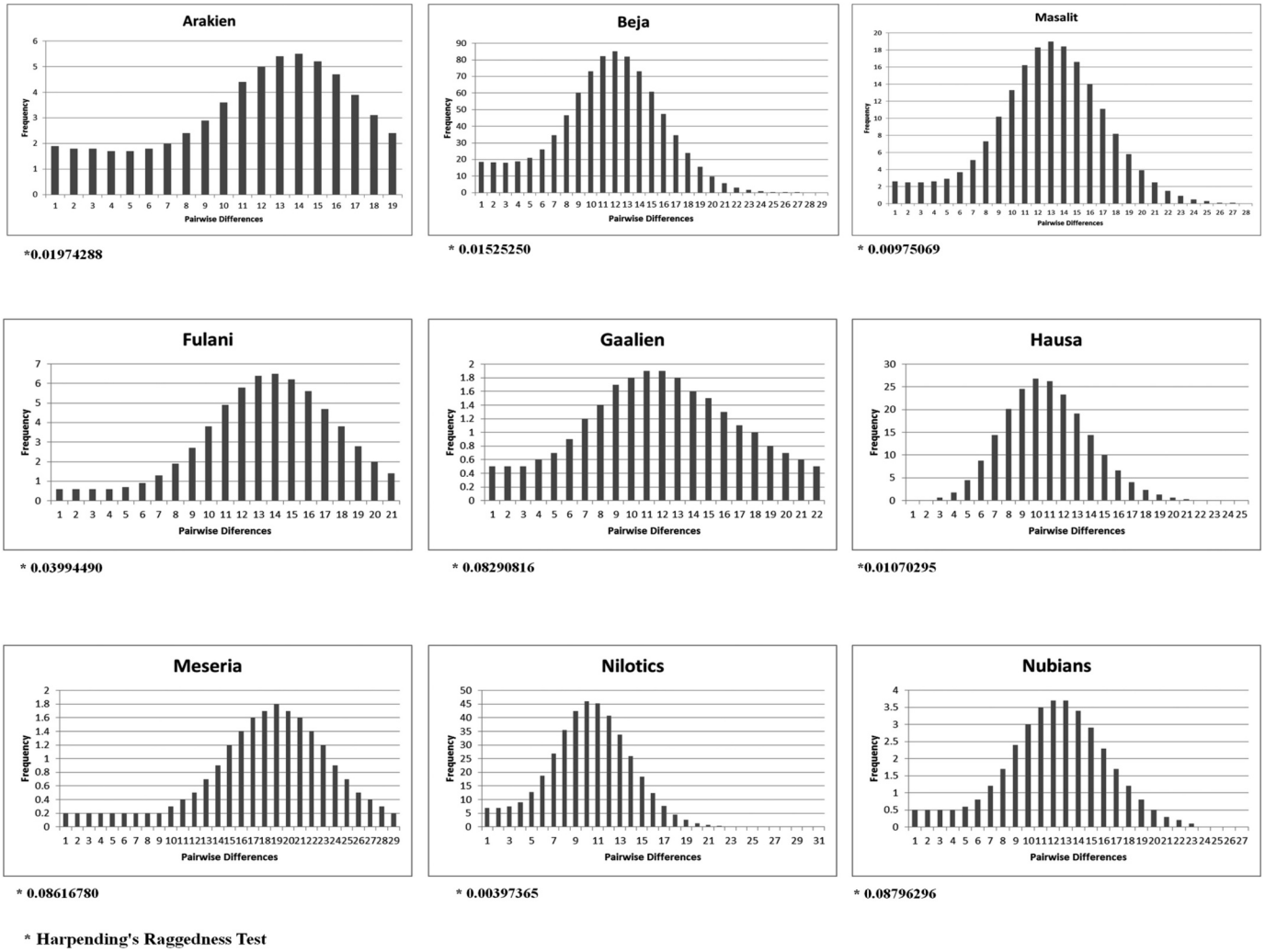

Fig. 4. Pairwise mismatch distributions of nine Sudanese populations with all populations showing normal distribution.

(Excoffier, 2004). We tend to favor this latter scenario as it tallies with a stable population structure emanating from antique mitochondrial DNA gene pool of striking diversity coupled with a continuous admixture over the millennia.

In the PCA plots all Sudanese populations cluster together possibly as a result of an extended history of mitochondrial DNA sharing and/or a recent maternal gene flow from a predominantly Nilo-Saharan indigenous speakers towards the Afro-Asiatic gene pool. An example of such gene flow is the high frequencies of haplogroups L2a which is widely distributed in sub-Saharan Africa (Salas et al., 2004) and L3f of East African origin(Watson et al., 1997) among populations labeled as Arabs.

Some of the groups like Meseria who have been shown to display Y-chromosome patterns suggesting recent migration to Africa, have strictly "native" mitochondrial profile sharing haplotypes with groups like Masalit in Western Sudan and Dinka in the South. Similar is the haplotype sharing between Arakien and Beja, which suggest that

Arakien males migrating from Arabia might have taken the eastern Sudan route. The time scale of haplotype sharing and coalescence between Masalit and Beja (L2) might be much older: either fairly early during differentiation of human populations in this part of east Africa or subsequently during the pastoralism /agricultural motivated expansion marked particularly by Y chromosome haplogroup E (Gebremeskel and Ibrahim, 2014).

The low values of variation among geographic and linguistic groups indicate that both geography and language have no significant role in shaping the mtDNA genetic structure of the Sudanese, which is expected since mtDNA variation has long predated the divergence of languages (Diamond, 2003; Barbujani, 1997). This could suggest that matrilineal and patrilineal patterns in Sudan may have had different genetic history, although this result should be taken with caution given the differences in mutation rates between the mitochondria and Y-chromosome.

In conclusion, our data shows remarkable feature of genetic diversity among the Sudanese population consistent with the antiquity and in-situ evolution of the mitochondrial DNA gene pool, and relating some of the populations to groups displaying the most ancestral lineages of maternal lines. The mitochondrial HVRI remains surrogate of whole mitochondrial genome information and may be utilized for rapid screening of populations of interest. Based on HVRI sequence, and in addition to affirming north-eastern Africa as a prime scene in early evolution of modern humans, variation in this gene pool seems to have been influenced to lesser extent by migrations than by in-situ evolution.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mgene.2020.100837.

Funding

This research did not receive a specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

H.Y.H., H.M., and H·S performed lab work and generated genotype data. M.E.K performed NGS analysis and extracted mitogenomes. M.M.O., E.I.G., H.Y.H., M.A.E. M.E.K., and T.G., analyzed data. M.E.I., M.M.O., H.Y.H., E.I.G., H.M., and H·S analyzed and interpreted results and wrote the paper.

Credit statement

Authors Role

1) Maha M. Osman analysis Writing - original draft Writing, Methodology
2) Hisham Y. Hassan Data curation, analysis, original draft writing
3) Mohammed A. Elnour analysis
4) Heeran Makkan Data curation
5) Eyoab Iyasu Gebremeskel, analysis
6) Thoyba Gais: analysis
7) Mahmoud E. Koko: analysis, Data curation;
8) Himla Soodyall: Conceptualization Data curation review & editing Funding acquisition
9) Muntaser E. Ibrahim: Conceptualization of the study, Data curation, review & editing, Funding acquisition,

Declaration of Competing Interest

The authors state no conflict of interest.

Acknowledgement

The authors wish to acknowledge the assistance of Carina Schlebusch from HDDRL, Division of Human Genetics, School of Pathology, and University of the Witwatersrand. The authors are grateful to the individuals who donated their samples.

References

Abu-Amero, K.K., et al., 2008. Mitochondrial DNA structure in the Arabian peninsula. BMC Evolutionary Biology 8 (1), 45. https://doi.org/10.1186/1471-2148-8-45.

Bandelt, H.J., Forster, P., Röhl, A., 1999. 'Median-joining networks for inferring intraspecific phylogenies', *Molecular Biology and Evolution*. Society for Molecular Biology and Evolution 16 (1), 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036.

Bandelt, H.J., et al., 2001. 'Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade', *Annals of Human Genetics*. Blackwell Science Ltd 65 (6), 549–563. https://doi.org/10.1046/j.1469-1809.2001.6560549.x.

Barbujani, G., 1997. DNA variation and language affinities. *American journal of human genetics*. Elsevier 61 (5), 1011–1014. https://doi.org/10.1086/301620.

Barker, F.K., et al., 2012. Contrasting evolutionary dynamics and information content of the avian mitochondrial control region and ND2 gene. PLoS ONE. Public library of science 7 (10). https://doi.org/10.1371/journal.pone.0046403 e46403.

Behar, D.M., et al., 2008. The Dawn of human matrilineal diversity. American Journal of Human Genetics. Elsevier 82 (5), 1130–1140. https://doi.org/10.1016/j.ajhg.2008.04.002.

Chen, Y.-S., et al., 2000. mtDNA variation in the south African kung and Khwe—and their genetic relationships to other African populations. The American Journal of Human Genetics Elsevier 66 (4), 1362–1383. https://doi.org/10.1086/302848.

Cruciani, F., et al., 2002. A Back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. The American Journal of Human Genetics Elsevier 70 (5), 1197–1214. https://doi.org/10.1086/340257.

Diamond, J., 2003. Farmers and their languages: the first expansions. Science 300 (5619), 597–603. https://doi.org/10.1126/science.1078208.

Drummond, A.J., et al., 2012. Bayesian P hylogenetics with BEAUti and the BEAST 1. 7. Mol. Biol. Evol. 29 (8), 1969–1973. https://doi.org/10.1093/molbev/mss075.

Elhassan, N., et al., 2014. The episode of genetic drift defining the migration of humans out of Africa is derived from a large east African population size. PLoS ONE 9 (5). https://doi.org/10.1371/journal.pone.0097674. Edited by D. Mishmar. e97674.

Excoffier, L., 1990. Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. 30 (2), 125–139. https://doi.org/10.1007/BF02099939.

Excoffier, L., 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Molecular Ecology. Blackwell Science Ltd 13 (4), 853–864. https://doi.org/10.1046/j.1365-294X.2003.02004.x.

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

Fox, C.L., 1997. mtDNA analysis in ancient Nubians supports the existence of gene flow between sub-Sahara and North Africa in the Nile valley. Annals of Human Biology Taylor & Francis 24 (3), 217–227. https://doi.org/10.1080/03014469700004952.

Gandini, F., et al., 2016. 'Mapping human dispersals into the horn of Africa from Arabian ice age refugia using mitogenomes', *Scientific Reports*. Nat. Publ. Group 6 (April), 1–13. https://doi.org/10.1038/srep25472.

Garrison, E. and Marth, G. (2012) 'Haplotype-based variant detection from short-read sequencing', *arXiv preprint arXiv*, pp. 1–8. Doi: arXiv:1207.3907 [q-bio.GN].

Gasca-Pineda, J., et al., 2013. Effective population size, genetic variation, and their relevance for conservation: the Bighorn sheep in Tiburon Island and comparisons with managed artiodactyls. PLoS ONE. Public library of science 8 (10). https://doi.org/10.1371/journal.pone.0078120 e78120.

Gebremeskel, E.I., Ibrahim, M.E., 2014. Y-chromosome E haplogroups: their distribution and implication to the origin of Afro-Asiatic languages and pastoralism. European journal of human genetics : EJHG 22 (12), 1387–1392. https://doi.org/10.1038/ejhg.2014.41.

Gonder, M.K., et al., 2007. Whole-mtDNA genome sequence analysis of ancient african lineages. Mol. Biol. Evol. 24 (3), 757–768. https://doi.org/10.1093/molbev/msl209.

González-martín, A. *et al.* (2015) 'Demographic history of indigenous populations in Mesoamerica based on mtDNA sequence data demographic history of indigenous populations in Mesoamerica based on mtDNA sequence data', *PLOS ONE*. Public library of science, 10(AUGUST), p. e0131791. Doi: 10.13140/RG.2.1.3914.9281.

Gordon, R.G., 2005. Ethnologue: Languages of the World, Fifteenth edition. SIL International, Dallas, Texas.

Guo, Y., et al., 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. Bioinformatics 29 (9), 1210–1211. https://doi.org/10.1093/bioinformatics/btt118.

Hall, T. (1999) 'BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT', *Nucleic Acids Symposium Series*, pp. 95–98. Doi: citeulike-article-id:691774.

Hassan, H.Y., et al., 2008. Y-chromosome variation among sudanese: restricted gene flow, concordance with language, geography, and history. Am. J. Phys. Anthropol. 137 (3), 316–323. https://doi.org/10.1002/ajpa.20876.

Hazkani-Covo, E., Graur, D., 2007. A comparative analysis of numt evolution in human and chimpanzee. *Molecular Biology and Evolution* United States 24 (1), 13–18. https://doi.org/10.1093/molbev/msl149.

Huoponen, K., et al., 1996. European mtDNAs from an analysis of three European populations. Genetics Society of America 144 (4), 1835–1850.

Ingman, M., Gyllensten, U., 2003. Mitochondrial genome variation and evolutionary history of Australian and new Guinean aborigines. Genome Research 1600–1606. https://doi.org/10.1101/gr.686603.the.

Ingman, M., et al., 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*. Macmillian Magazines Ltd. 2173 (1994), 708–713.

Kivisild, T., et al., 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *American journal of human genetics*. Elsevier 75 (5), 752–770. https://doi.org/10.1086/425161.

Knight, A., et al., 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Current Biology* Elsevier 13 (6), 464–473. https://doi.org/10.1016/S0960-9822(03)00130-1.

Kot, M., Daniel, W.A., 2008. The relative contribution of human cytochrome P450 isoforms to the four caffeine oxidation pathways: an in vitro comparative study with cDNA-expressed P450s including CYP2C isoforms. Biochem. Pharmacol. 76 (4), 543–551. https://doi.org/10.1016/j.bcp.2008.05.025.

Krings, M., et al., 1999. mtDNA analysis of Nile River valley populations: a genetic corridor or a barrier to migration? Am. J. Hum. Genet. 64 (4), 1166–1176. https://doi.org/10.1086/302314.

Lundstrom, R., Tavare, S., Ward, R.H., 1992. Estimating substitution rates from molecular data using the coalescent. Proc. Natl Acad Sci USA 89 (13), 5961–5965.

Macaulay, V., 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308 (5724), 1034–1036. https://doi.org/10.1126/science.1109792.

Macaulay, V., et al., 1999. The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am. J. Hum. Genet. 64 (1), 232–249. https://doi.org/10.1086/302204.

Marjoram, P., Donnelly, P., 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136 (2), 673–683. https://doi.org/10.1016/j.jvolgeores.2013.01.004.

Mcbrearty, S., Brooks, A.S., 2000. The revolution that wasn't: a new interpretation of the origin of modern human behavior. J. Hum. Evol. 39 (5), 453–563. https://doi.org/10.1006/jhev.2000.0435.

McDougall, I., Brown, F.H., Fleagle, J.G., 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433 (7027), 733–736. https://doi.org/10.1038/nature03258.

Mishmar, D., et al., 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. In: Human Mutation, 23(2). Wiley Subscription Services, Inc., A Wiley Company, pp. 125–133. https://doi.org/10.1002/humu.10304.

Quintana-Murci, L., et al., 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat. Genet. 23 (4), 437–441. https://doi.org/10.1038/70550.

Rito, T., et al., 2013. The first modern human dispersals across Africa. *PLoS ONE*. Public Library of Science 8 (11). https://doi.org/10.1371/journal.pone.0080031 e80031.

Rodríguez-Serrano, E., Cancino, R.A., Palma, R.E., 2006. Molecular Phylogeography of Abrothrix Olivaceus (Rodentia: Sigmodontinae) in Chile. Journal of Mammalogy, United States 87 (5), 971–980. https://doi.org/10.1644/05-MAMM-A-393R2.1.

Rogers, A.R., 1995. Genetic evidence for a Pleistocene population explosion. Evolution 49 (4), 608. https://doi.org/10.2307/2410314.

Ruiz-Pesini, E., 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303 (5655), 223–226. https://doi.org/10.1126/science.1088434.

Salas, A., et al., 2002. The making of the African mtDNA landscape. The American Journal of Human Genetics, Elsevier 71 (5), 1082–1111. https://doi.org/10.1086/344348.

Salas, A., et al., 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *American journal of human genetics*. Elsevier 74 (3), 454–465. https://doi.org/10.1086/382194.

Semino, O., et al., 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *American journal of human genetics*. Elsevier 70 (1), 265–268. https://doi.org/10.1086/338306.

Soares, P., et al., 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. American Journal of Human Genetics, Elsevier 84 (6), 740–759. https://doi.org/10.1016/j.ajhg.2009.05.001.

Tamura, K., et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28 (10), 2731–2739. https://doi.org/10.1093/molbev/msr121.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. ClustalW: improving the sensitivity of progressive multiple sequence aligment through sequence weighting, position specific gap penalties and weight matrix choice. Nucl Acids Res 22 (22), 4673–4680.

Tishkoff, S.A., et al., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271 (5254), 1380–1387. https://doi.org/10.1126/science.271.5254.1380.

Tishkoff, S.A., et al., 2009. The genetic structure and history of Africans and African Americans. Science, United States 324 (5930), 1035–1044. https://doi.org/10.1126/science.1172257.

Torroni, A., et al., 2001. Do the four clades of the mtDNA Haplogroup L2 evolve at different rates? Am. J. Hum. Genet. 69, 1348–1356. https://doi.org/10.1086/324511.

Underhill, P.A., et al., 2000. Y chromosome sequence variation and the history of human populations. Nat. Genet. 26 (3), 358–361. https://doi.org/10.1038/81685.

Underhill, P.A., et al., 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. 65 (1), 43–62. https://doi.org/10.1046/j.1469-1809.2001.6510043.x.

Vigilant, L., et al., 1989. Mitochondrial DNA sequences in single hairs from a southern African population. Proc. Natl. Acad. Sci. 86 (23), 9350–9354. https://doi.org/10.1073/pnas.86.23.9350.

Wall, J.D., 2003. Estimating ancestral population sizes and divergence times. Genetics 163 (1), 395–404.

Watson, E., et al., 1997. Mitochondrial footprints of human expansions in Africa. Am. J. Hum. Genet. 61 (3), 691–704. https://doi.org/10.1086/515503.