mtDNA Variation in East Africa Unravels the History of Afro-Asiatic Groups

Alessio Boattini,[1]* Loredana Castrì,[1]† Stefania Sarno,[1] Antonella Useli,[1,2] Manuela Cioffi,[1] Marco Sazzini,[1] Paolo Garagnani,[3] Sara De Fanti,[1] Davide Pettener,[1] and Donata Luiselli[1]

[1]*Department of Biological, Geological and Environmental Sciences, Laboratory of Molecular Anthropology, University of Bologna, 40126, Bologna, Italy*
[2]*Department of Science for Nature and Environmental Resources, University of Sassari, 07100 Sassari, Italy*
[3]*Department of Experimental Pathology, University of Bologna, 40126 Bologna, Italy*

KEY WORDS Cushitic; Semitic; Omotic; Horn of Africa; genetic structure

ABSTRACT East Africa (EA) has witnessed pivotal steps in the history of human evolution. Due to its high environmental and cultural variability, and to the long-term human presence there, the genetic structure of modern EA populations is one of the most complicated puzzles in human diversity worldwide. Similarly, the widespread Afro-Asiatic (AA) linguistic phylum reaches its highest levels of internal differentiation in EA. To disentangle this complex ethno-linguistic pattern, we studied mtDNA variability in 1,671 individuals (452 of which were newly typed) from 30 EA populations and compared our data with those from 40 populations (2970 individuals) from Central and Northern Africa and the Levant, affiliated to the AA phylum. The genetic structure of the studied populations—explored using spatial Principal Component Analysis and Model-based clustering—turned out to be composed of four clusters, each with different geographic distribution and/or linguistic affiliation, and signaling different population events in the history of the region. One cluster is widespread in Ethiopia, where it is associated with different AA-speaking populations, and shows shared ancestry with Semitic-speaking groups from Yemen and Egypt and AA-Chadic-speaking groups from Central Africa. Two clusters included populations from Southern Ethiopia, Kenya and Tanzania. Despite high and recent gene-flow (Bantu, Nilo-Saharan pastoralists), one of them is associated with a more ancient AA-Cushitic stratum. Most North-African and Levantine populations (AA-Berber, AA-Semitic) were grouped in a fourth and more differentiated cluster. We therefore conclude that EA genetic variability, although heavily influenced by migration processes, conserves traces of more ancient strata. Am J Phys Anthropol 000:000–000, 2013.

Populations from East Africa (EA) are one of the most compelling puzzles in human diversity worldwide, both from a genetic and a linguistic perspective. In light of the long-term hominid occupation attested by the local fossil record, the Horn of Africa is very likely to have had a major role in the emergence of anatomically modern humans. Furthermore it is one of the most probable gateways for Eurasian colonization by *Homo sapiens*. However, EA complexity could also be the result of more recent events. Ethiopia, for instance, has been involved in a broad net of people movements that extends from the Levant and Arabian Peninsula—via the Bab-el-Mandeb strait—to Central Africa and the Chad Basin—via the Sahel and Southern Sahara (Kivisild et al., 2004; Cerný et al., 2007; 2009; Tishkoff et al., 2009; Cruciani et al., 2010; Musilova et al., 2011; Pagani et al., 2012). Southwards, the area approximately comprised between Kenya, Tanzania, Uganda and Sudan was affected by Bantu expansions and gene flow from Nilo-Saharan-speaking pastoralists starting from ~3,000 years before present (Castrì et al., 2008, 2009; Tishkoff et al., 2009; de Filippo et al., 2011; Gomes et al., 2010). Northwards, the Nile basin has been a privileged way of access to North-Eastern Africa for Neolithic technological innovations (i.e., pastoralism and agriculture) (Newman, 1995).

Language diversity in EA fits well with its complicated genetic history. In Fleming words, "Ethiopia by itself has more languages than all of Europe, even counting all the so-called dialects of the Romance family" (Fleming, 2006). All African linguistic phyla are found in EA: Afro-Asiatic (AA), Nilo-Saharan, Niger-Congo and Khoisan (however, the genealogical unit of Khoisan is no longer generally accepted). Among them, AA is the most differentiated, being represented by three (Omotic, Cushitic, Semitic) of its six major clades (the others being Chadic, Berber and Egyptian). Omotic and Cushitic are considered the deepest clades of AA, and both are found almost exclusively in the Horn of Africa, along with the linguistic relict Ongota that is traditionally assigned to the Cushitic family but whose classification is still widely debated (Fleming, 2006). These observations are in agreement

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

Grant sponsor: University of Bologna RFO grants 2010 and 2011 to DL.

*Correspondence to: Alessio Boattini, Department of Biological, Geological and Environmental Sciences, Via Selmi 3, 40126, Bologna, Italy.
E-mail: alessio.boattini2@unibo.it

†Diseased

Received 11 July 2012; accepted 19 November 2012

DOI 10.1002/ajpa.22212
Published online in Wiley Online Library (wileyonlinelibrary.com).

TABLE 1. List of the sampled ethnic groups from Ethiopia and Kenya with their geographic location, linguistic affiliation, approximate census size and subsistence patterns

Population	Code	Sample Size	Approximate Census size	Country	Language affiliation	Long.	Lat.	Subsistence Patterns
Dawro-Konta	37	137	259,633	Ethiopia	AA-Omotic	37.16676	7.08454	Agro-pastoralists
Ongota	38	19	89	Ethiopia	AA-Cushitic	36.98167	4.83278	Hunter-gatherers, small-scale agriculture
Hamer	39	11	42,466	Ethiopia	AA-Omotic	36.48333	4.96667	Agro-pastoralists
Rendille	40	17	34,700	Kenya	AA-Cushitic	37.46613	2.80588	Seminomadic pastoralists
Elmolo	41	52	700	Kenya	AA-Cushitic	36.71906	2.74827	Fishing
Luo	42	49	4,270,000	Kenya	Nilo-Saharan	34.4751	-0.53832	Agro-pastoralists, fishing
Maasai	43	81	590,000	Kenya	Nilo-Saharan	36.85089	0.70036	Seminomadic pastoralists
Samburu	44	35	174,000	Kenya	Nilo-Saharan	37.06512	1.25783	Seminomadic pastoralists
Turkana	45	51	451,000	Kenya	Nilo-Saharan	35.11231	4.46691	Seminomadic pastoralists

Census size information was extracted from the ethnologue web site (www.ethnologue.com).

with a North-Eastern African origin of the AA languages, most probably in pre-Neolithic times (Ehret, 1979, 1995; Kitchen et al., 2009). The main contender to the African hypothesis is the farming-language dispersal theory (Diamond and Bellwood, 2003), according to which AA languages originated in the Levant and spread in Africa after the Neolithic revolution, along with agriculture and cattle rising. However, this view might contrast with evidence for a local development of farming packages in East Africa such as the Ethiopian ensete, tef and coffee (Phillipson, 1998). Despite the fact that relationships between linguistics and genetics are often elusive, the fact that both languages and genes are influenced by the same kind of evolutionary factors, together with the correlations between genetic and linguistic variation that were observed in several studies, make the use of linguistic affiliations a reasonable way of grouping populations for genetic studies (Scheinfeldt et al., 2010).

This research therefore aims to provide a better knowledge of the mitochondrial genetic structure of populations from EA by taking in consideration both geography and linguistics. More precisely, we are trying to address the following questions: first, is the mtDNA genetic structure of EA more related to geography or linguistics? Second, does the East African high genetic diversity have recent (<3,000 years before present) or ancient origins? Third, from a molecular perspective, which of the two hypotheses about the AA origin is more plausible? To answer all these questions, we have investigated mtDNA variability in individuals belonging to three AA linguistic families (Omotic, Cushitic, Semitic) and two linguistic phyla (Nilo-Saharan, Niger-Congo). We decided not to take into consideration EA Khoisan-speakers, which would behave both as linguistic and genetic outliers in our sample, potentially flattening the diversity observed within AA populations. In addition, these populations are not relevant to the specific purposes of this study. Moreover, in order to better explore the genetic relationships among AA speakers, samples from the Chad basin (Chadic family), North Africa (Berber and Semitic families) and Levant (Semitic family) are included in our analyses, reaching a total of 4,641 EA and AA samples. Materials include 452 HVS-I unpublished sequences from nine populations settled in EA, among which individuals speaking the little known – from the genetic perspective – AA-Omotic family (Dawro Konta, Hamer) and of the controversial AA Ongota language.

MATERIALS AND METHODS

Population samples and locations

Buccal swabs were collected from 167 Ethiopian and 285 Kenyan unrelated apparently healthy individuals belonging to nine ethnic groups and two linguistic phyla: Dawro-Konta (137, AA-Omotic), Hamer (11, AA-Omotic), Ongota (19, AA-Cushitic), Rendille (17, AA-Cushitic), Elmolo (52, AA-Cushitic), Luo (49, Nilo-Saharan), Maasai (81, Nilo-Saharan), Samburu (35, Nilo-Saharan) and Turkana (51, Nilo-Saharan). Their geographic location, linguistic affiliation, census size and subsistence patterns are detailed in Table 1. Census sizes greatly vary between groups, spanning from Kenyan Luo, who exceed 4,000,000, to Elmolo and Ongota, whose sizes are around 700 and only 89, respectively. Not surprisingly, languages spoken by these groups are nowadays almost extinct, Elmolo having shifted to Samburu (Nilo-Saharan) and Ongota to Tsamai (AA-Cushitic). All the considered groups are basically patrilineal and clan systems and/or age-grade institutions of governance are widespread, coupled with exogamy (marriage outside clan) and polygyny, with varying degrees of intensity. For instance, Dawro-Konta people have a very strong clan system, while this is less important among Turkana. As for subsistence patterns, seminomadic pastoralism is frequent in the Lake Turkana area (Turkana, Samburu, Rendille) as well as among Maasai from Kenya. An important exception is represented by Elmolo, who are mainly fishermen (but formerly they were pastoralists, too). Dawro-Konta and Hamer from Ethiopia are agro-pastoralists (with an emphasis on the first term for Dawro-Konta, vice versa for Hamer). Ongota are the only case of hunter-gatherers. Ethnic origin, birthplace and up-to-grandfathers maternal and paternal pedigrees of all individuals were ascertained by oral interview performed in collaboration with local consultants. The collection of biological samples was performed during several expeditions conducted from 1999 to 2010. Written ethical approval for the use of samples from Kenya in this study was provided by the ethics committee of the University of Bologna (record of 2 March 2011; hard copies are available upon request). Samples from Ethiopia were procured in 2007 (Dawro-Konta) and 2010 (Hamer, Ongota) with individual informed consent and following the ethical guidelines stipulated by the research institutions involved in this project. The confidentiality of personal information for each participant to the study was assured.

Reference data from EA include 1,219 HVS-I sequences from individuals belonging to 21 different ethnic groups. A further set of 2,970 HVS-I sequences from North Africa (NA) (1,600), Central Africa (CA) (275) and the Levant (1,095) was established in order to explore the genetic variability of AA-speakers outside EA. The complete reference dataset includes 4,641 HVS-I sequences from 79 different populations. Geographic locations and linguistic affiliations of each of the considered groups are detailed in Supplementary Table 1.

HVS-I sequencing

DNA was extracted by means of a salting out modified protocol (Miller et al., 1988). mtDNA variability was investigated with a focus on the first hypervariable segment (HVS-I), by sequencing a total of 360 base pairs (bp), encompassing nucleotide positions from 16,024 to 16,383.

Polymerase chain reaction (PCR) of the HVS-I region was performed in a T-Gradient Thermocycler (Whatman Biometra, Gottingen, Germany) using L15996 and H16401 primers and following the standard protocol (Vigilant et al., 1991). PCR products were purified by ExoSap-IT® (USB Corporation, Cleveland, OH) and sequenced on an ABI Prism 3730 Genetic Analyzer (Applied Biosystem), using a Big-Dye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), according to the manufacturer's instructions and with the aforementioned primers. To reduce ambiguities in sequence determination the forward and reverse primers were used to sequence both strands of HVS-I. Sequences were then aligned to the reference sequence (Anderson et al., 1981; Andrews et al., 1999) using the DNA Alignment Software 1.3.0.1 (http://www.fluxusengineering.com/align.htm).

To ensure data quality, all sequences were aligned and edited by two researchers independently. The final consensus sequence was then generated by comparing the two independent results. No ambiguities were found.

In order to fit our data with the most updated mtDNA phylogeny (PhyloTree build 15; van Oven, 2009), haplotype motifs were obtained comparing sequences with both the Cambridge Reference Sequence (CRS) and the new Reconstructed Sapiens Reference Sequence (RSRS, Behar et al. 2012). Both CRS- and RSRS-based haplotypes, as well as the corresponding haplogroups, are detailed in Supporting Information Table 2. For comparison purposes, we used the level of phylogenetic resolution adopted in Poloni et al. (2009).

Statistical methods

All following analyses are based on haplotypes and nucleotide differences among haplotypes are taken into account.

Nucleotide diversity and Analysis of Molecular Variance (AMOVA) were calculated using the software Arlequin 3.5 (Excoffier and Lischer, 2010). A Non-Metric Multi-Dimensional Scaling bi-dimensional plot of the examined populations was obtained calculating Nei's distance (Nei, 1972) and using the function isoMDS implemented in the R software package MASS (Cox and Cox, 2001; Venables and Ripley, 2002; R Development Core Team, 2008).

Relationships between geographic coordinates of the populations and genetic variation (HVS-I allelic frequencies) were explored by means of a spatial Principal Component Analysis (sPCA) performed using the R software package adegenet (Jombart, 2008; Jombart et al., 2008; R Development Core Team, 2008). Differently from classic PCA, where eigenvalues are calculated by maximizing variance of the data, in sPCA eigenvalues are obtained maximizing the product of variance and spatial autocorrelation (Moran's I index). In order to include spatial information in the analysis, we used a weighting procedure based on a Delaunay connection network. Eigenvalues obtained by sPCA are both positive and negative, depending from Moran's I positive or negative values. The most informative components are those identified by eigenvalues with the highest absolute values. Large positive components correspond to global structures (i.e., cline-like structures), whereas large negative components correspond to local structures (i.e., marked genetic differentiation among neighbors). Only scores from the most informative components (up to ~80% of the sum of the eigenvalues absolute values) were retained. Therefore, we excluded from downstream analyses those components that convey scarce spatial information and low variance. A Model Based Clustering algorithm, as implemented in the Mclust function included in the R software Mclust package (Fraley and Raftery, 2002; 2006), was then applied to sPCA scores. Mclust explores a set of ten different models for Expectation-Maximization (EM) – each characterized by a different parameterization of the covariance matrix – and for different number of clusters and chooses the best one according to the highest Bayesian Information Criterion (BIC). The output includes the parameters of the maximum-BIC model and the corresponding classification (i.e., affiliation of each population to one of the inferred clusters).

An independent evaluation of membership probabilities for each population to the Mclust-inferred clusters was obtained by means of Discriminant Analysis of Principal Components (DAPC). It is important to note that a certain population – attributed by the above sPCA-Mclust procedure to a certain cluster – not necessarily has its highest DAPC-based membership probability for the same cluster. Actually, DAPC membership probabilities are here used as indicators of how clear-cut genetic clusters are. Accordingly, low values can be interpreted as evidence of admixture with populations from other clusters. The DAPC method (Jombart et al., 2010) aims to describe the diversity between pre-defined groups of observations. Analyses were performed using the R software adegenet package (Jombart 2008; R Development Core Team 2008). Despite being designed to investigate individual genetic data, the method can be easily adapted to population data (and in general to all kinds of tabular data). Briefly, the DAPC procedure consists of two steps. First, original data (e.g., allele frequencies) are centered and submitted to a PCA. Second, the retained PCs are passed to a Linear Discriminant Analysis. As a result, discriminant functions are constructed as linear combinations of the original variables which have the largest between-group variance and the smallest within-group variance. Membership probabilities are based on the retained discriminant functions. Concerning the first step, it is important to observe that retaining too many PCs with respect to the number of populations can lead to over-fitting the discriminant functions, meaning that membership probabilities may become drastically inflated for the best-fitting cluster, resulting in apparent perfect discrimination. The optimal number of retained PCs is evaluated calculating the

α-score, which is the difference between the proportion of successful reassignment of the analysis (observed discrimination) and values obtained using random groups (random discrimination). The procedure relies on repeating DAPC with different randomized groups (the default setting is ten) for different numbers of retained PCs. The 'best' number of retained PCs is the one that optimizes the mean α-score (i.e., the closest to one). The same problem would hold also for the second step, e.g., the number of retained discriminant functions. In our case, given that the number of inspected clusters is low (three), all the discriminant functions were retained.

Surface plots of nucleotide diversity in Africa and the Levant and macro-haplogroup frequencies in EA were obtained with the software Surfer 8 (Golden Software, Golden, CO).

The BayeSSC software (Excoffier et al., 2000; Anderson et al., 2005) was used to perform coalescent simulations of multiple sets of HVS-I mtDNA sequences assuming different demographic scenarios. Simulation sets were used to test the following hypotheses: (1) whether the high genetic EA nucleotide diversity is mainly the result of recent (from ~3,000 YBP) or more ancient events; (2) whether mtDNA results are consistent with an EA origin of AA or with a Levantine one. In both cases, we considered 25 years generations, HVS-I substitution rate of 1.64723×10^{-7} mutations per nucleotide per year (Soares et al., 2009), a Kimura 2-Parameter model with Gamma correction of 0.4 and a transition/transversion bias of 0.91 (Poloni et al., 2009). For hypothesis (1), we proceeded as follows. We simulated four populations corresponding to the four clusters (A, B1, B2, C; see Results) identified using the above described method (sPCA, Mclust); sample sizes are equal to the number of individuals affiliated to each cluster. As a basic demographic model, we assumed three African Sub-Saharan populations (A, B1, B2) splitting from each other ~5,000 generations ago (~125,000 YBP; Garrigan et al., 2007). Experiments with higher values did not yield significant changes in results (not shown). For these clusters we assumed a constant population size. A fourth population (C)—simulating a Levantine/North African cluster—was assumed to split from A 2,400 generations ago (~60,000 YBP) according to a bottleneck scenario (followed by re-expansion) compatible with the parameter space estimated by Gravel et al. (2011). Effective population sizes were introduced in the model as prior uniform distributions varying between 1,500 and 6,500. Within this model, we tested four scenarios with different degrees of population mobility: (a) no migrations (only population splits); (b) instant gene flow (33%) at 120 generations ago (~3,000 YBP) following the direction C → A → B1 → B2; (c) continuous gene flow from 120 generations ago to the present (migration matrix based on mean DAPC membership probabilities per cluster); (d) sum of (b) and (c). For each of the four tested scenarios, we performed 2,000,000 preliminary simulations in BayeSSC. Simulated nucleotide diversity values were further processed for calculating most likely estimates (MLE) of model parameters (population effective sizes) using Approximate Bayesian Computation and retaining the best 5% simulations (Beaumont, 2008). A second set of 100,000 simulations per scenario was run based on MLE. Finally, we calculated AIC (Akaike Information Criterion) for each scenario by comparing simulated and observed nucleotide diversity values.

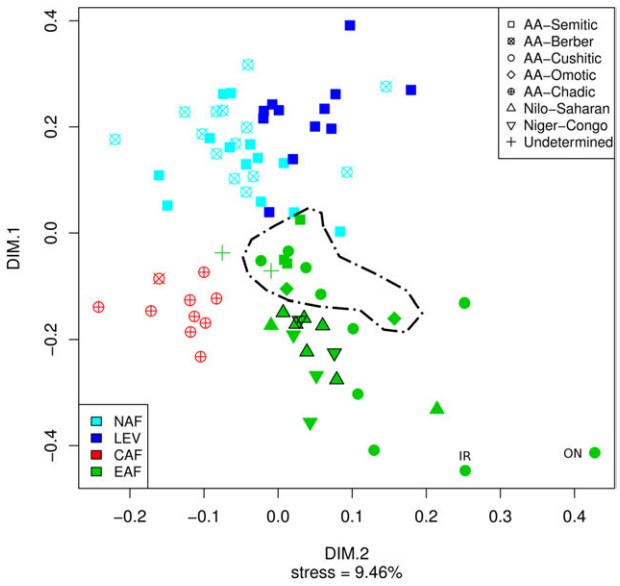

Fig. 1. Non-Metric MDS representation of the 79 examined populations. The plot is rotated right by 90° to better fit the representation with geographic coordinates. Ethiopian samples are enclosed by the dashed-dotted line. Nilo-Saharan and Niger-Congo groups from Kenya are represented by black bordered symbols. Labels in the plot indicate the position of outlier populations (ON: Ongota, IR: Iraqw). Stress value is lower than the cutoff threshold according to Sturrock and Rocha, 2000 (38.8% for two dimensions and 79 objects).

Simulations for hypothesis 2 assume substantial identity between genes and language evolution processes. In other words, these experiments may be read as a mean to understand if mtDNA variability is able to distinguish between two different scenarios (EA or Levantine origin of AA) in a fully idealized case. We simulated 200 HVS-I sequences equally shared between two populations—representing EA and the Levant—evolving independently from a common proto-AA-speaking ancestor. Based on results from precedent simulations, effective population sizes were set to 2,000 and 1,000, respectively, while we assumed a constant population size. We tested two scenarios: (a) EA origin of AA with split between the two populations at 480 generations ago (~12,000 YBP), (b) Levantine origin of AA with split between the two populations at 200 generations ago (~5,000 YBP). 100,000 simulations per scenario were run and empiric distributions for standard genetic parameters (haplotype diversity, nucleotide diversity, Tajima's D) and Nei's distance were compared.

RESULTS

Nucleotide diversity in the whole dataset (Supporting Information Table 1) varies between 0.0115 ± 0.0064 (Libyan Tuareg, 22) and 0.0305 ± 0.0156 (Datoga, 55), the mean being 0.0206. Values higher than the third quartile of the empiric distribution (0.0244) are found almost exclusively in EA, the only exception being a Chadic-speaking population from CA (Hide, 73). The highest values are observed in Kenya and Tanzania, with a decreasing gradient moving towards NA and the Levant (Supporting Information Fig. 1).

As a first overview of mtDNA genetic landscape in the considered populations, we performed a MDS analysis (Fig. 1). Results show that Levantine populations are

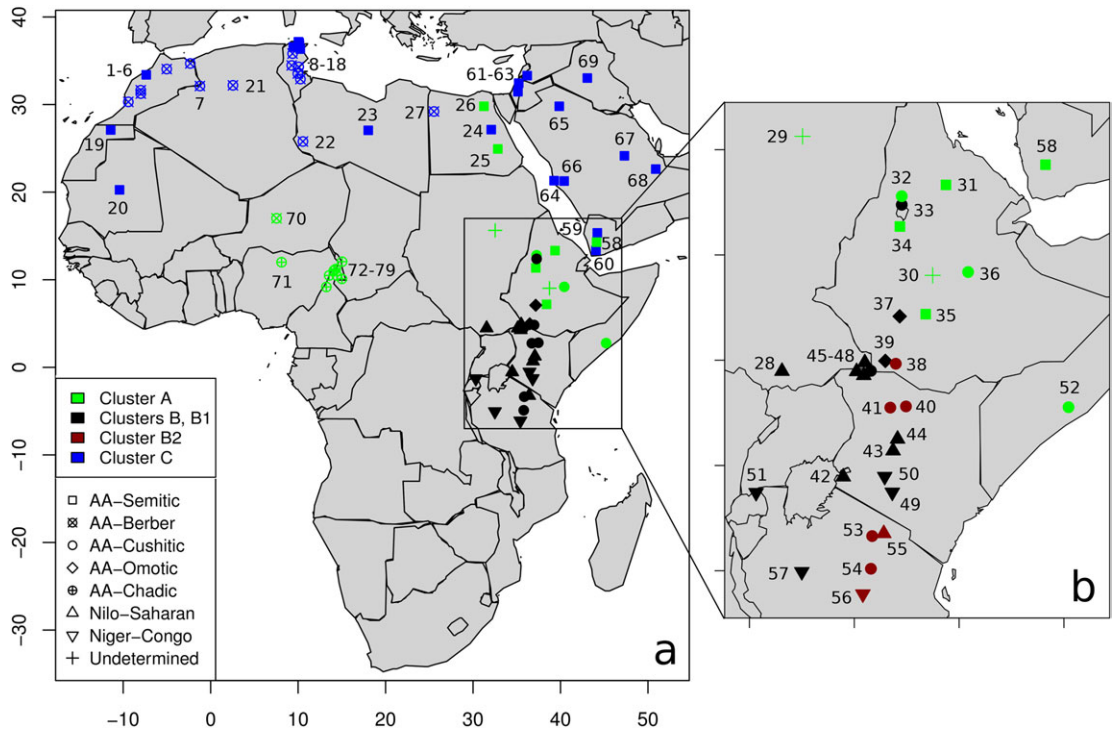

Fig. 2. Geographic distribution of the sPCA-Mclust-based clusters calculated on the whole dataset (a) and on East Africa only (b).

separated from EA ones along the first dimension, while the second dimension highlights a West-East gradient. Berber and Semitic groups from NA are undistinguishable from the mtDNA point of view, being interspersed between each other. Chadic-speaking populations from CA form a tight and fairly homogeneous group. On the contrary, EA populations show a remarkable degree of variability. Among them, Cushitic speakers are particularly heterogeneous, sometimes behaving as outgroups (Ongota, 38; Iraqw, 53). Groups from Ethiopia fall in the center of the plot (dashed-dotted line), behaving like a bridge between the Levant, NA and EA. Nilo-Saharan and Niger-Congo samples from Kenya form a homogeneous cluster.

sPCA and Mclust analyses were performed at two different geographic scales: (a) on the whole African and Levantine dataset (including all AA-speaking populations); (b) on populations from EA.

In the first case, after having performed sPCA on allelic frequencies data, we retained the first three global sPCs, explaining respectively 73.30%, 10.17%, and 4.37% of the sum of the absolute values of eigenvalues, for a total of 87.83%. sPC scores were fed to Mclust, that settled (as a best model) to an ellipsoidal model with equal shape (VEV) and three clusters (Supporting Information Fig. 2a). Figure 2a shows the geographic distribution of the three inferred clusters (A, B, C).

Cluster A (green) is frequent in EA, mostly in Ethiopia, and in CA. Interestingly, one Yemenite and two Egyptian populations are included in this cluster. Cluster B (black) occupies a relatively tighter area, spanning from Tanzania to Southern Ethiopia. From a linguistic point of view, cluster B includes non-AA speakers (namely Nilo-Saharan and Niger-Congo), as well as AA-Cushitic and Omotic speakers. Cluster C (blue) collects populations from NA and the Levant. From a linguistic point of view, these populations are affiliated to AA-Semitic and AA-Berber families.

The corresponding membership probabilities were calculated using DAPC and retaining eight principal components. Results are detailed in Supplementary Table 1. Cluster C shows the highest mean membership probability (0.897 ± 0.003); individual lowest values are found in a group from Tunisia (14), one from Yemen (59) and two from Saudi Arabia (64, 65). All of them show relevant contributions from cluster A, which can be interpreted as traces of probably recent events of admixture. Mean membership probability in cluster A is 0.810 ± 0.007. It includes two Egyptian populations (25 and 26) that show evidence of admixture with cluster C. Cluster B is characterized by the lowest mean membership probability (0.773 ± 0.015); accordingly, great part of its populations show evidence of admixture with other clusters, most notably with A (in particular 28, 33, 39, 40, 44, 57).

Focusing on EA, and including one population from Yemen (58, cluster A), we retained four sPCs encompassing for 59.32%, 8.23%, 5.40% and 4.82%, respectively, of the sum of the absolute values of eigenvalues, for a total of 77.77%. In that case, Mclust found that the best model was a diagonal, equal shape (VEI) with three clusters (Supporting Information Fig. 2b). As for the geographic location of the clusters, (Fig. 2b) we observe that the 'green' cluster coincides exactly with the EA distribution of cluster A, hence we maintain the same label. More interestingly, cluster B from previous analysis here diverges in two groups, which we call clusters B1 and B2. Cluster B2 (dark red) shows a discontinuous geographic pattern, being divided into a northern portion – located on the border between Ethiopia and Kenya –

TABLE 2. AMOVA results according to geographic-, linguistic-, and mclust-based groupings; (a) whole dataset; (b) East Africa only

Grouping	N° of Groups	N° of Pops	Proportion of variation (%) Among Groups	Among Populations Within Group	Within Populations
(a) Whole Data Base					
All populations	1	79	-	8.80***	91.20***
Geography	4[a]	79	7.66***	3.19***	89.15***
Geography (AA populations only)	4[a]	64	6.22***	3.12***	90.66***
Language	3[b]	76	8.87***	6.17***	84.96***
Language (AA clades)	7[c]	76	7.17***	3.55***	89.28***
Language (AA clades only)	5[d]	64	5.51***	3.88***	90.61***
Mclust	3	79	9.04***	2.96***	88.01***
(b) East Africa Only					
All populations	1	31	–	4.91***	95.09***
Geography	7[e]	31	2.49***	2.90***	94.61***
Geography (AA populations only)	5[f]	16	3.41*	3.38***	93.21***
Language	3[b]	28	1.06*	4.56***	94.38***
Language (AA clades)	5[g]	28	0.92	4.44***	94.63***
Language (AA clades only)	3[h]	16	-0.25	6.18***	94.07***
Mclust	3	31	3.84***	2.43***	93.73***

*** P-value < 0.001; **P-value < 0.01; *P-value < 0.05.
[a] North Africa, East Africa, Levant, Central Africa.
[b] AA, Nilo-Saharan, Niger-Congo.
[c] AA-Semitic, AA-Berber, AA-Cushitic, AA-Omotic, AA-Chadic, Nilo-Saharan, Niger-Congo.
[d] AA-Semitic, AA-Berber, AA-Cushitic, AA-Omotic, AA-Chadic.
[e] Ethiopia, Kenya, Sudan, Somalia, Rwanda, Tanzania, Yemen.
[f] Ethiopia, Kenya, Somalia, Tanzania, Yemen.
[g] AA-Semitic, AA-Cushitic, AA-Omotic, Nilo-Saharan, Niger-Congo.
[h] AA-Semitic, AA-Cushitic, AA-Omotic.

and a southern one, located in Tanzania. Cluster B2 collects most of the AA-Cushitic speaking populations from these areas (including Ongota). Cluster B1, on the contrary, includes AA-Omotic and non-AA (Nilo-Saharan and Niger-Congo) speaking populations.

Concerning DAPC-based membership probabilities (eight PCs retained), cluster B1 shows the highest mean value (0.949 ± 0.005) and conversely the lowest traces of admixture with other clusters. The opposite is true for cluster B2, revealing low mean membership probability (0.700 ± 0.052) and strong evidences of introgressions from cluster B1 (40, 53, 56). Cluster A is characterized by a mean membership probability equal to 0.863 ± 0.024, showing some relevant contributions from cluster B1 (35).

To further test the reliability of the above clusters, we performed AMOVA at both the full dataset level and the EA level (Table 2). Results were compared with geography- and language-based groupings. In both cases, the highest proportion of variance among groups (9.04%, $p < 0.001$ and 3.84%, $p < 0.001$, respectively) is reached with the Mclust inferred groups. At the full dataset level, significant results were obtained also with geography- and language-based groupings, albeit with lower values. At the EA level, language-based groupings did not yield significant (or only marginally significant, but not after Bonferroni correction) results, while a better score is obtained with geography-based groupings.

Coalescent simulations under different scenarios were performed to test whether the observed patterns of nucleotide diversity within the identified clusters were consistent with recent (~ 3,000 YBP) gene flow. Among the considered scenarios (Table 3, Supporting Information Fig. 3), the least-fitting one (AIC = 1.17) is characterized by absence of migration between populations. Notably, the lowest AIC figure (0.214) was obtained for the scenario involving the highest degree of population mobility, i.e., instant gene flow at 120 generations ago followed by continuous migrations (with rates based on DAPC membership probabilities).

TABLE 3. Akaike Information Criterion (AIC) values calculated comparing observed and simulated nucleotide diversity values assuming four different demographic scenarios

	Scenario			
	Splits	Gen. Flow.	Mig. Mat.	AIC
a	Y	N	N	1.178
b	Y	Y	N	1.062
c	Y	N	Y	0.617
d	Y	Y	Y	0.214

Each scenario is characterized by the presence/absence (Y/N) of the following demographic events: population splits (Splits), instant gene flow (33%) at 120 generations ago (Gen. Flow), continuous gene flow from 120 generations ago to the present (Mig. Mat.). For details see the Methods.

Assuming identity between processes leading to language and genetic variability, a second set of simulations was performed to test whether mtDNA variability may help to discern between an EA and a Levantine origin of AA. Empiric distributions for standard genetic parameters (haplotype diversity, nucleotide diversity, Tajima's D) and Nei's distance (Supporting Information Fig. 4) for the two scenarios show that their confidence intervals are largely overlapping.

Table 4 details distributions of the considered haplogroups in the tree EA clusters, while frequencies for each single EA population are included in Supplementary Table 3. Cluster A is characterized by high frequencies of L2a, M and R0a haplogroups, with lower frequencies of the L0 and L3 lineages, except for L3f.

TABLE 4. Frequencies of the considered mtDNA haplogroups in East African clusters A, B1, and B2

	A	B1	B2
	N (%)	*N* (%)	*N* (%)
H	15 (2.1)	0 (0)	1 (0.4)
HV	20 (2.8)	7 (0.8)	6 (2.7)
I	5 (0.7)	5 (0.6)	16 (7.2)
J	15 (2.1)	2 (0.2)	2 (0.9)
K	20 (2.8)	9 (1.1)	0 (0)
L0	0 (0)	5 (0.6)	0 (0)
L0a	40 (5.6)	108 (12.8)	66 (29.6)
L0b	0 (0)	8 (0.9)	1 (0.4)
L0d	1 (0.1)	2 (0.2)	1 (0.4)
L0f	5 (0.7)	39 (4.6)	32 (14.3)
L0g	0 (0)	5 (0.6)	0 (0)
L0k	2 (0.3)	0 (0)	0 (0)
L1	0 (0)	2 (0.2)	0 (0)
L1b	6 (0.8)	7 (0.8)	0 (0)
L1c	2 (0.3)	11 (1.3)	1 (0.4)
L2	0 (0)	4 (0.5)	0 (0)
L2a	103 (14.4)	78 (9.2)	8 (3.6)
L2b	12 (1.7)	11 (1.3)	1 (0.4)
L2c	2 (0.3)	1 (0.1)	0 (0)
L2d	3 (0.4)	6 (0.7)	0 (0)
L2e	1 (0.1)	0 (0)	0 (0)
L3	2 (0.3)	7 (0.8)	1 (0.4)
L3a	6 (0.8)	17 (2)	23 (10.3)
L3b	6 (0.8)	38 (4.5)	2 (0.9)
L3c	0 (0)	2 (0.2)	2 (0.9)
L3d	20 (2.8)	18 (2.1)	1 (0.4)
L3e	15 (2.1)	36 (4.3)	1 (0.4)
L3f	39 (5.5)	35 (4.1)	2 (0.9)
L3h	19 (2.7)	53 (6.3)	18 (8.1)
L3i	26 (3.6)	36 (4.3)	1 (0.4)
L3x	22 (3.1)	41 (4.8)	2 (0.9)
L4	2 (0.3)	3 (0.4)	0 (0)
L4a	20 (2.8)	11 (1.3)	0 (0)
L4b2	14 (2)	87 (10.3)	22 (9.9)
L5	1 (0.1)	17 (2)	0 (0)
L5a	7 (1)	23 (2.7)	0 (0)
L5b	3 (0.4)	2 (0.2)	0 (0)
L5c	4 (0.6)	29 (3.4)	0 (0)
L6	25 (3.5)	12 (1.4)	4 (1.8)
M	85 (11.9)	41 (4.8)	2 (0.9)
N	34 (4.8)	8 (0.9)	5 (2.2)
R	3 (0.4)	0 (0)	0 (0)
R0a	52 (7.3)	9 (1.1)	1 (0.4)
T	16 (2.2)	2 (0.2)	1 (0.4)
U	28 (3.9)	8 (0.9)	0 (0)
Others	12 (1.7)	1 (0.1)	0 (0)

Cluster B2 shows the highest frequencies of L0a (reaching 29.6%), L0f, L3a, L3h and I, while it has the lowest frequencies of L2 and L5 lineages, as well as M and R0a. Cluster B1 occupies an intermediate position between A and B2, while showing the highest frequencies of L4 lineages. Contour maps of the most frequent macro-haplogroups (L0, L2, L3, L4, M) are reported in Supporting Information Figure 5.

DISCUSSION

In their mtDNA-based survey of East African variability, Poloni and colleagues found "no strong association between linguistically-defined and genetically differentiated groups". Furthermore, they observed that EA "combines a high level of within population-diversity with strong genetic structure among populations". They argue that such results "may be explained [as a consequence of] periodical episodes of admixture in these populations, separated by periods of isolation and genetic drift" (Poloni et al., 2009). Our results largely agree with these observations and, in addition, we were able to uncover traces of an underlying and as yet uncovered genetic structure.

Anyway, a possible drawback of our procedure relies on the fact that sPCA may minimize the role of drift and isolation on single populations, while DAPC maximizes between-group variability, hence underestimating the component of variance generated by gene flow. Another possible source of distortion, however independent from the statistical methods used, could be due to discrepancies in the sampling criteria used in reference studies. Although we cannot exclude some minor effects, we do not observe any detectable relationship between different data sources (i.e., reference studies) and our results.

Our analyses indicate that the structure of EA mtDNA diversity is characterized by three population clusters: A, B1 and B2 (Fig. 2, Supporting Information Table 1). Such structure appears to be related both with geography and linguistic affiliation. On the contrary, to the best of our knowledge there is no evidence of relationships with other socio-cultural variables such as mating behavior (patrilocality is widespread in EA, as well as clan exogamy), social structure (clan-based structures are present in almost all of our samples) and subsistence patterns. The same can be said for demographic dimensions, given that each cluster does include populations with widely differing census sizes (Table 1). Cluster A is centered in Ethiopia and highlights long-range connections of Ethiopian Semitic- and Cushitic-speaking groups with Chadic ones from Central Africa, and Semitic ones from Egypt and the Arabic peninsula. This finding is highly consistent with the role of Ethiopia as a primary hub for recent human migrations already detected in other studies. In fact, movements between Ethiopia and the Arabian peninsula via the Bab-el-Mandeb strait were revealed by mtDNA analyses (Kivisild et al., 2004; Musilova et al., 2011), confirming ancient links between the two coasts of the Red Sea (at least since 8,000 – 9,000 YBP). A reconstruction of the phylogeny of the Semitic linguistic family suggested a single, presumably Levantine origin for Semitic languages in the Horn of Africa (Ethiosemitic), dating their diversification at approximately 2,850 YBP (Kitchen et al., 2009). Evidences of introgressions from the Levant to Ethiopia in the same time frame were indeed revealed by a very recent whole genome study (Pagani et al., 2012). Further population movements along the Nile Valley are suggested by the affiliation to cluster A of two Egyptian populations (25, Gurna and 26, Upper Egypt). They could be related to the spread of Neolithic technologies – according to Newman (1995), the first evidences of pastoralism and agriculture in highland Ethiopia date to ~5,000 YBP – or as the remnants of an ancient AA unity (Egyptian is an extinct branch of AA) extending from EA to Egypt (Stevanovitch et al., 2004). Furthermore, mtDNA (Cerný et al., 2009) revealed traces of ancient movements between EA and Central Africa. (Based on Y-chromosome, Cruciani et al., 2010, instead proposed a different route linking Central Africa with North-Eastern Africa). These last migrations were suggested to be responsible for the introduction of Chadic languages (along with their speakers) in the Chad basin area. The high frequencies of haplogroups M and R0a and, to a lesser extent, of T and U (Table 4) – all of them related with the Levant and Asia (Rosa and Brehem, 2011) – fit well with the high mobility patterns detected for this area.

Contrarily to cluster A, clusters B1 and B2 are restricted to EA only, at least in our panel of populations. This means that groups belonging to these clusters do not have 'relatives' among AA-speaking populations outside EA, but maybe they could have them in non-AA groups that were not included in our study. Cluster B2 shows an interesting association with southern Cushitic groups, including the Ongota (38), who are problematic from a linguistic point of view. In fact, the Ongota language, despite being traditionally assigned to the Cushitic family, is suspected to be the remnant of an independent clade of AA (Fleming, 2006), while other scholars (Savà and Tosco, 2003) propose that it may be considered a Cushitic language retaining a Nilo-Saharan substratum. Notably, this last hypothesis implies mixed ancestry for the Ongota, helping to explain their outlying position in genetic space (Fig. 1). In addition, B2 encompasses populations as Rendille (40) and Elmolo (41), that, despite sharing Cushitic languages and the same geographic area (Marsabit district, North-Eastern Kenya), are at present characterized by different subsistence strategies (pastoralism and fishing, respectively, but Elmolo were formerly herders, too). Elmolo's affiliation to cluster B2 is of particular interest, given their current reduced census size (~700) and the fact that their language is almost extinct, being largely substituted by Samburu (44, cluster B1). A second group of B2 populations is located far more south, in Tanzania (53, Iraqw; 54, Burunge; 55, Datog; 56, Turu). Indeed, the association between B2 and AA-Cushitic seems particularly relevant given the discontinuous geographic distribution of the cluster and the fact that Cushitic is considered one of the deepest and most ancient clades of AA (Ehret, 1995); cluster B2, also, shows the highest frequencies of L0 lineages (in particular L0a and L0f), the deepest clade of the human mtDNA phylogeny (Rosa and Brehem, 2011).

Cluster B1 is widespread from the Ethiopian-Kenyan border to Tanzania (Fig. 2), almost encapsulating cluster B2, that, in turn, shows strong signals of admixture with B1. To add other elements to the picture, various studies demonstrated that EA was involved in Bantu expansions (~3,000 YBP, Scheinfeldt et al., 2012), acting both as a point of arrival of migrations from the West and as a starting point for further movements directed to the south, while contributing largely to the diffusion of Niger-Congo languages (Castrì et al., 2008, 2009; de Filippo et al., 2011). Furthermore, a recent Y-chromosome study (Gomes et al., 2010) showed that Kenya and Tanzania (along with Uganda and Sudan) were also affected by a dispersal of pastoralist people speaking Nilo-Saharan languages around 3,000 YBP. Looking at the big picture, it seems reasonable to hypothesize that cluster B1 may be, at least in part, the result of migrations from West of non-AA groups. These migratory processes may have caused both disruption of the geographic distribution of Cushitic cluster B2, as well as admixture, and, at a certain extent, language shifts. For instance, Ehret (1998) argues for a number of now extinct South Cushitic languages in Tanzania on the basis of loan word evidence in local Bantu languages. Further investigations are needed to clarify this important issue.

Omotic speakers, despite the fact that Omotic is generally considered the deepest branch in AA (Blench, 2006), do not show any particular pattern of differentiation, at least from the mtDNA perspective. In fact, they are included in cluster B1, suggesting that wide and ancient phenomena of admixture and/or language shift may have occurred in the area between southern Ethiopia and Kenya. Indeed, high mobility rates are one of the most likely keys to explain the elevated nucleotide diversity observed in EA (Supporting Information Tab. 1, Supporting Information Fig. 1), as well as its (up to now) almost undecipherable genetic structure. Our coalescent simulation experiments reveal indeed that the most likely scenario is the one that implies the most elevated degree of population mobility (Table 3). Such scenario assumes major migration events around 3,000 YBP (corresponding to Bantu and Nilo-Saharan migrations in EA) followed by continuous migratory flows (whose rates are based on DAPC membership probabilities).

Nearly all populations from Northern Africa and the Levant, corresponding to the Semitic and Berber families of AA, are affiliated to cluster C. This finding agrees with contacts and bi-directional migratory exchanges involving the wide corridor between the Maghreb and the Near East already detected for Y-chromosome, autosomal STRs and SNPs (Semino et al., 2004; Tishkoff et al., 2009, Scheinfeldt et al., 2012; Henn et al., 2012). Cluster C is also the most divergent one, showing the highest mean value of membership probability and, consequently, only very limited signals of admixture with other African clusters. These last observations are consistent with the postulated Western Eurasian origin of large part of North African mtDNA lineages (Maca-Meyer et al., 2003; Cherni et al, 2008; Ennafaa et al., 2009, Coudray et al., 2009).

Our results are largely consistent with those of Tishkoff et al. (2009), who adopted a different clustering method using multilocus autosomal data. As in our case, they observe that within EA, clustering is primarily (but not exclusively) associated with language phyla (AA, Nilo-Saharan, Niger-Congo, Khoisan). For instance, their results show that AA-Cushitic-speaking populations from Tanzania as Iraqw (53) and Gorowa cluster with Nilo-Saharan Datoga (55), who are close geographically, mirroring almost perfectly the southern portion of our B2 cluster. Interestingly, according to their results Elmolo (41, cluster B2) are related with formerly Cushitic-speaking groups as Yaaku from Kenya and Akie from Tanzania (not available for our study), strengthening our interpretation of B2 as a Cushitic-specific cluster. As for AA-Chadic speaking groups, Tishkoff et al. (2009) find some shared ancestry between these populations and AA groups from EA, but they conclude that their spread in Central Africa "was not accompanied by large amounts of AA gene flow".

Coming back to the three questions that introduced our study, we are now able to point out the following answers. Concerning the relationships between EA genetic structure and geographic/linguistic affiliations, we observe that both of them contribute to explain our results. Clusters A and B1 are both related to geography, the first being mainly located in Ethiopia and Central Africa, the second in Kenya and Tanzania. Nevertheless, cluster B2 shows an interesting association with some Eastern and all Southern Cushitic populations; cluster A itself is exclusively associated with AA language families, namely Chadic, Cushitic and Semitic. If the arrival of Semitic in EA is relatively recent (~2,850 YBP; Kitchen et al., 2009), Cushitic seems much older (at least ~7,000 YBP, according to Ehret, 1979) and the same holds for proto-Chadic expansions in Central Africa

(~7,000 YBP, according to Ehret, 2002). The fact that Cushitic groups are separated into two different clusters (A and B2) may be an indirect proof of their antiquity. On the whole, these observations indicate that languages had an important role in shaping the matrilineal genetic structure of EA.

As a second point, we asked whether the high mtDNA genetic diversity observed in EA (Supporting Information Table 1, Supporting Information Fig. 1) may be interpreted as the outcome of recent events. According to our results, the answer is yes. It has to be mentioned here that our method for detecting clusters enhances the geographic structure of mtDNA variability, by retaining only those sPCA components with the highest absolute value of eigenvalues. For instance, Ethiopia-centered cluster A can be read as the outcome of recent and repeated migration events, which, from a longitudinal point of view, extend from the Arabian Peninsula to Central Africa (being likely related to the spread of Chadic languages), and from a latitudinal point of view, reach Egypt through the Nile Valley. This is consistent with the local linguistic structure, overlapping in the very same area an ancient and autochthonous AA clade (i.e., Cushitic) with a more recent and exogenous one (i.e., Semitic). Similarly, cluster B, spreading from Southern Ethiopia to Tanzania, was affected by different migration events. In particular, cluster B1 may be interpreted as the result of recent contributions (starting from ~3,000 YBP) from Bantu (Niger-Congo) and Nilo-Saharan pastoralists. Cushitic-specific cluster B2 itself shows clear traces of the same migration events, as suggested by its low mean membership probabilities, as well as by evidences of introgression from B1 (Supporting Information Table 1). In addition, coalescent simulation experiments (Table 3, Supporting Information Fig. 3) suggest that the observed nucleotide diversity patterns can be best explained assuming high population mobility. Nevertheless, our results showed that all these migrations did not manage to completely delete more ancient genetic structures. In particular, cluster B2 seems to be the remnant of an ancient Cushitic continuity between Kenya and Tanzania. On the whole, EA has functioned both as a contact point between already differentiated populations and languages, and as an ancient center of expansion.

Concerning the third point, i.e., the place of origin of AA (EA or the Levant), our results do not allow us to make conclusive statements. Indeed, coalescent simulations of different genetic parameters (Supporting Information Fig. 4) according to the two mentioned hypotheses show that—even assuming complete correlation between languages and mtDNA variability—their confidence intervals largely overlap. Thus, we limit ourselves to the following observations. First, EA shows the highest levels of nucleotide diversity among the studied populations with a decreasing cline towards NA and the Levant (Supporting Information Fig. 1 and Supporting Information Table 1). This is true not only for the Ethiopian cluster A, but also, and especially, for groups belonging to clusters B1 and B2. Second, EA hosts the two deepest clades of AA, Omotic and Cushitic. These families are found exclusively in EA, while the presence of Semitic in this area is much more recent. Third, cluster C – collecting Berber- and Semitic-speaking populations from NA and the Levant – shows only modest signals of admixture with clusters A and B (Fig. 2, Supporting Information Table 1). None of these points, taken by itself, is conclusive, but undoubtedly the hypothesis of origin of AA in EA is the most parsimonious one, if compared to the Levant.

CONCLUSIONS

This study confirms the central role of EA and the Horn of Africa in the genetic and linguistic history of a wide area spanning from Central and Northern Africa to the Levant. Our results confirm high mtDNA diversity and strong genetic structuring in EA. We were indeed able to identify three population clusters (A, B1, B2) that are related both to geography and linguistics, and signaling different population events in the history of the region. The Horn of Africa (cluster A), in accordance with its role as a major gateway between sub-Saharan Africa and the Levant, shows widespread contacts with populations from CA (AA-Chadic speakers), the Arabian peninsula and the Nile Valley. Southwards, Kenya, and Tanzania (clusters B1 and B2), despite being both heavily involved in Bantu and Nilo-Saharan pastoralist expansions, reveal traces of a more ancient genetic stratum associated with Cushitic-speaking groups (cluster B2). Conversely, Berber- and Semitic-speaking populations of NA and the Levant show only marginal traces of admixture with sub-Saharan groups, as well as a different mtDNA genetic background, making the hypothesis of a Levantine origin of AA unlikely. In conclusion, EA genetic structure configures itself as a complicated palimpsest in which more ancient strata (AA-Cushitic-speaking groups) are largely overridden by recent different migration events. Further explorations of AA-Cushitic-speaking populations – both in terms of sampled groups and typed genetic markers – will be of great importance for the reconstruction of the genetic history of EA and AA-speakers.

ACKNOWLEDGMENTS

The authors wish to dedicate this article to the memory of our dear friend and colleague Loredana Castrì. We would like to acknowledge all the participants to the study as well as Francesca Lipeti (Fatima Health Center, Lengesim, Kenya), Samantha Semproli, Serena Tucci and Gianluca Frinchillucci (Perigeo onlus, www.perigeo.org) for their invaluable help in designing and performing the sampling campaigns for this research. AB would like to thank Luca Pagani for his comments that helped to improve the manuscript. This study was in part funded by University of Bologna RFO grants 2010 and 2011 to DL. The authors declare no conflict of interest with the publication of the present study.

LITERATURE CITED

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

Anderson CNK, Ramakrishnan U, Chan YL, Hadly EA. 2005. Serial SimCoal: a population genetic model for data from multiple populations and points in time. Bioinformatics 21:1733–1734.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

Beaumont MA. 2008. Joint determination of topology, divergence time and immigration in population tree. In: Matsu-

mura S, Forster P, Renfrew C, editors. Simulation, genetics and human prehistory. Cambridge: McDonald Institute for Archaeological Research, University of Cambridge. p 135–154.
Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675–684.
Blench R. 2006. Archaeology, language, and the African past. Lanham, MD: Altamira Press.
Castrì L, Garagnani P, Useli A, Pettener D, Luiselli D. 2008. Kenyan crossroads: migration and gene flow in six ethnic groups from Eastern Africa. J Anthropol Sci 86:189–192.
Castrì L, Tofanelli S, Garagnani P, Bini C, Fosella X, Pelotti S, Paoli G, Pettener D, Luiselli D. 2009. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. Am J Phys Anthropol 140:302–311.
Cerný V, Fernandes V, Costa MD, Hájek M, Mulligan CJ, 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.
Cerný V, Salas A, Hájek M, Zaloudková M, Brdicka R. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71:433–452.
Cherni L, Fernandes V, Pereira JB, Costa MD, Goios A, Frigi S, Yacoubi-Loueslati B, Amor MB, Slama A, Amorim A, El Gaaied AB, Pereira L. 2009. Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia. Am J Phys Anthropol 139:253–260.
Coudray C, Olivieri A, Achilli A, Pala M, Melhaoui M, Cherkaoui M, El-Chennawi F, Kossmann M, Torroni A, Dugoujon JM. 2009. The complex and diversified mitochondrial gene pool of Berber populations. Ann Hum Genet 73:196–214.
Cox TF, Cox MAA. 2001. Multidimensional scaling. Boca Raton: Chapman and Hall.
Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, Watson E, Colomb EB, Dugoujon JM, Moral P, Scozzari R. 2010. Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. Eur J Hum Genet 18:800–807.
de Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, Nyambe T, Beyer K, Schreiber H, de Knijff P, Luiselli D, Stoneking M, Pakendorf B. 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol 28:1255–1269.
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science 300:597–603.
Ehret C. 1979. On the antiquity of agriculture in Ethiopia. J Afr Hist 20:161–177.
Ehret C. 1995. Reconstructing proto-Afro-Asiatic. Berkeley, CA: University of California Press.
Ehret C. 1998. An African classical age: eastern and southern Africa in world history, 1000 B.C. to A.D. 400. Charlottesville: University of Virginia Press.
Ehret C. 2002. The civilizations of Africa: a history to 1800. Oxford: James Currey.
Ennafaa H, Cabrera VM, Abu-Amero KK, González AM, Amor MB, Bouhaha R, Dzimiri N, Elgaaïed AB, Larruga JM. 2009. Mitochondrial DNA haplogroup H structure in North Africa. BMC Genet 10:8.
Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. J Hered 91:506–509.
Excoffier L, Lischer HE. 2010. Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–7.
Fleming HC. 2006. Ongota: a decisive language in African prehistory. Wiesbaden: Harrassowitz.
Fraley C, Raftery AE. 2006. MCLUST Version 3 for R: normal mixture modeling and model-based clustering. Technical Report No. 504, Department of Statistics, University of Washington (revised 2009).
Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis and density estimation. J Am Statist Assoc 97:611–631.
Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, Soodyall H, Strassmann B, Destro-Bisol G, de Knijff P, Novelletto A, Friedlaender J, Hammer MF. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177:2195–2207.
Gomes V, Sánchez-Diz P, Amorim A, Carracedo A, Gusmão L. 2010. Digging deeper into East African human Y chromosome lineages. Hum Genet 127:603–13.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA,The 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. PNAS 108:11983–11988.
Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, Bustamante CD, Comas D. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397.
Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci USA 105:10693–10698.
Jombart T, Devillard S and Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94.
Jombart T, Devillard S, Dufour AB, Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity (Edinb) 101:92–103.
Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.
Kitchen A, Ehret C, Assefa S, Mulligan CJ. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc Biol Sci 276:2703–2710.
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.
Maca-Meyer N, González AM, Pestano J, Flores C, Larruga JM, Cabrera VM. 2003. Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet 4:15.
Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16:1215.
Musilová E, Fernandes V, Silva NM, Soares P, Alshamali F, Harich N, Cherni L, Gaaied AB, Al-Meeri A, Pereira L, Cerný V. 2011. Population history of the Red Sea: genetic exchanges between the Arabian Peninsula and East Africa signaled in the mitochondrial DNA HV1 haplogroup. Am J Phys Anthropol 145:592–598.
Nei M. 1972. Genetic distances between populations. Am Nat 106:283–292.
Newman JL. 1995. The peopling of Africa: a geographic interpretation. New Haven: Yale University Press.
Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Bekele E, Bradman N, Balding DJ, Tyler-Smith C. 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet 91:83–96.
Phillipson DW. 1998. Ancient Ethiopia. Aksum: its antecedents and successors. London: British Museum Press.
Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, Langaney A, Sanchez-Mazas A. 2009. Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann Hum Genet 73:582–600.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3–900051-07–0, Available at: URLhttp://www.R-project.org.

Rosa A, Brehem A. 2011. African human mtDNA phylogeography at-a-glance. J Anthropol Sci 89:25–58.

Savà G, Tosco M. 2003. The classification of Ongota. In: Bender ML, Takács G, Appleyard DL, editors. Selected comparative-historical Afrasian linguistic studies. Munich, Germany: LINCOM Europa. p 307–316.

Scheinfeldt LB, Soi S, Tishkoff SA. 2010. Colloquium paper: working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. Proc Natl Acad Sci USA 107Suppl 2:8931–8938.

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74:1023–1034.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759.

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Béraud-Colomb E. 2004. Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann Hum Genet 68:23–39.

Sturrock K, Rocha J. 2000. A multidimensional scaling stress evaluation table. Field Methods 12:49–60.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394. Available at:http://www.phylotree.org.

Venables WN, Ripley BD. 2002. Modern applied statistics with S, 4th ed. New York: Springer.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507.