Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups

Cesare de Filippo,*†[1] Chiara Barbieri,*†[1] Mark Whitten,[1] Sununguko Wata Mpoloka,[2] Ellen Drofn Gunnarsdóttir,[3] Koen Bostoen,[4] Terry Nyambe,[5] Klaus Beyer,[6] Henning Schreiber,[7] Peter de Knijff,[8] Donata Luiselli,[9] Mark Stoneking,[3] and Brigitte Pakendorf[1]

[1]Max Planck Research Group on Comparative Population Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Department of Biological Sciences, University of Botswana, Gaborone, Botswana

[3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Royal Museum for Central Africa, Université libre de Bruxelles, Tervuren, Belgium

[5]Livingstone Museum, Livingstone, Zambia

[6]Department of Asian and African Studies, Humboldt University, Berlin, Germany

[7]Department of African Linguistics and Ethiopian Studies, University of Hamburg, Hamburg, Germany

[8]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

[9]Department of Experimental Evolutionary Biology, University of Bologna, Bologna, Italy

†These authors contributed equally to this work.

***Corresponding author:** E-mail: cesare_filippo@eva.mpg.de; chiara_barbieri@eva.mpg.de.

Associate editor: Sarah Tishkoff

Abstract

Technological and cultural innovations as well as climate changes are thought to have influenced the diffusion of major language phyla in sub-Saharan Africa. The most widespread and the richest in diversity is the Niger-Congo phylum, thought to have originated in West Africa ~10,000 years ago (ya). The expansion of Bantu languages (a family within the Niger-Congo phylum) ~5,000 ya represents a major event in the past demography of the continent. Many previous studies on Y chromosomal variation in Africa associated the Bantu expansion with haplogroup E1b1a (and sometimes its sublineage E1b1a7). However, the distribution of these two lineages extends far beyond the area occupied nowadays by Bantu-speaking people, raising questions on the actual genetic structure behind this expansion. To address these issues, we directly genotyped 31 biallelic markers and 12 microsatellites on the Y chromosome in 1,195 individuals of African ancestry focusing on areas that were previously poorly characterized (Botswana, Burkina Faso, Democratic Republic of Congo, and Zambia). With the inclusion of published data, we analyzed 2,736 individuals from 26 groups representing all linguistic phyla and covering a large portion of sub-Saharan Africa. Within the Niger-Congo phylum, we ascertain for the first time differences in haplogroup composition between Bantu and non-Bantu groups via two markers (U174 and U175) on the background of haplogroup E1b1a (and E1b1a7), which were directly genotyped in our samples and for which genotypes were inferred from published data using linear discriminant analysis on short tandem repeat (STR) haplotypes. No reduction in STR diversity levels was found across the Bantu groups, suggesting the absence of serial founder effects. In addition, the homogeneity of haplogroup composition and pattern of haplotype sharing between Western and Eastern Bantu groups suggests that their expansion throughout sub-Saharan Africa reflects a rapid spread followed by backward and forward migrations. Overall, we found that linguistic affiliations played a notable role in shaping sub-Saharan African Y chromosomal diversity, although the impact of geography is clearly discernible.

Key words: human, language, geography, migration, Y chromosome, Bantu.

Introduction

Modern humans originated ~200,000 years ago (ya) in Africa, subsequently colonizing the rest of the globe. Genetic studies indicate that the ancestral African populations could have been structured even before ~100,000 ya when modern humans first began migrating out of Africa (Campbell and Tishkoff 2008; Wall et al. 2009). Genetic diversity values are much higher in African populations than elsewhere (Campbell and Tishkoff 2008). Africa is also linguistically very diverse: More than 2,000 languages are reported for the whole continent, comprising 30% of the world's languages (Gordon and Grimes 2005). Disregarding some isolates, African languages have been classified into four major phyla (Greenberg 1948): Afro-Asiatic, Khoisan (which, however, is no longer considered a historical unit by several specialists, see Güldemann and Vossen 2000), Niger-Congo, and Nilo-Saharan. Of these, the largest linguistic phylum is Niger-Congo (Williamson and Blench 2000), comprising ~1,400 languages and containing many related language families and several distantly or questionably related language groups (Sands 2009). For instance,

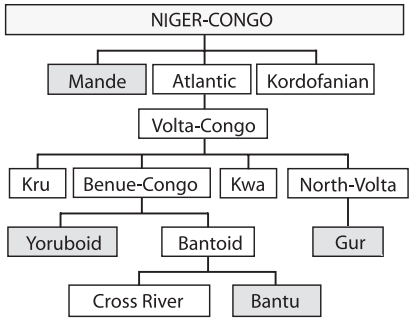

Fig. 1 Niger-Congo language tree. Schematic tree of the Niger-Congo language phylum that comprises three major branches: Mande, Kordofanian, and Atlantic-Congo (Williamson 1989). In gray boxes, linguistic families that are represented in our data set.

Mande and Kordofanian—two of the three major branches of Niger-Congo (fig. 1)—have been suggested as belonging to an earlier split, and some authors even doubt the affiliation of one or the other to the phylum (Williamson and Blench 2000; Dimmendaal 2008).

Since the migration of modern humans out of Africa, numerous population movements have played a role in shaping patterns of linguistic and genetic variation within the continent itself (Campbell and Tishkoff 2008). New forms of subsistence and technological improvements such as those derived from agriculture have driven population expansions even over long geographic distances. However, the major African linguistic phyla are assumed to have originated and spread much earlier than the advent of agriculture, which developed relatively late in sub-Saharan Africa: Cultivated plants did not appear before 4,000 ya (Neumann 2005). Indeed, it has been suggested that the expansion of Niger-Congo and Nilo-Saharan started ~12,000–10,000 ya with the improving climate at the beginning of the Holocene when speakers were still hunter gatherers (Blench 2006; Dimmendaal 2008). Nevertheless, it seems plausible that these expansions were triggered by technological innovations (e.g., bow, arrows, and domesticated dogs) and/or climatic changes (e.g., wetter conditions) in the Holocene approximately 11,000 ya (Blench 2006).

The most significant and well-known migration event in sub-Saharan Africa that has been associated—although not unanimously—with agricultural innovations, and at a later stage with iron technologies, is the expansion of the Bantu language family belonging to the Niger-Congo phylum (fig. 1). These languages are assumed to have originated in the Grassfields region between Cameroon and Nigeria not more than 5,000 ya and spread from this homeland throughout sub-Saharan Africa to Somalia in the East and as far as the Cape in the South (Nurse and Philippson 2003). The manner in which Bantu languages and speech communities spread throughout sub-Saharan Africa remains a matter of debate among specialists (Vansina 1979, 1995; Ehret 2001; Holden 2002; Eggert 2005; Holden and Gray 2006; Bostoen 2007). The general view of Diamond and Bellwood (2003) suggests that Bantu languages and agricultural techniques spread together with people throughout sub-Saharan Africa. However, this view is opposed by other investigators emphasizing the effect of cultural spread rather than movement of people (see Vansina 1995; Nichols 1997; Robertson and Bradley 2000). Several genetic studies that focused mainly on the uniparentally transmitted mitochondrial DNA (mtDNA) and Y chromosome are in favor of the first hypothesis, namely that the Bantu expansion was a joint linguistic and demographic event. As regards mtDNA, several haplogroups such as L0a, L2a, L3b, and L3e have been associated with the Bantu expansion (Salas et al. 2002), whereas for the Y chromosome, haplogroups E1b1a (defined by the single nucleotide polymorphism [SNP] M2) and B2b (defined by M150) have been connected to this event (cf. Thomas et al. 2000; Cruciani et al. 2002; Berniell-Lee et al. 2009). However, no differences have been detected in frequency and diversity levels of haplogroup E1b1a between Bantu and other Niger-Congo populations. In fact, not only does the geographic distribution of E1b1a extend far beyond the area settled by speakers of Bantu languages, but its frequency and the associated STR diversity are even higher in non–Bantu-speaking regions, such as Guinea Bissau (Rosa et al. 2007). In their extensive study of Y chromosomal variation in Africa, Wood et al. (2005) genotyped M191, which defines a sublineage of E1b1a called E1b1a7, which was also associated with the Bantu expansion (Zhivotovsky et al. 2004). They found a significant correlation between linguistic and Y chromosome variation, which is driven in large part by the correlation of Y chromosomal variation and the Bantu language family. They inferred that sex-biased migrations between expanding Bantu agriculturalists and hunter gatherers have notably affected the patterns of Y chromosomal variation in sub-Saharan Africa. However, this study was based on biallelic markers alone, and data from the entire south-central part of sub-Saharan Africa were lacking.

Although studies of autosomal polymorphisms are becoming more common as a result of technological advances (e.g., Hammer et al. 2008; Tishkoff et al. 2009; Bryc et al. 2010; Sikora et al. 2010), investigations of uniparental markers still offer valuable insights into human prehistory that cannot be obtained by autosomal markers alone. One advantage is the possibility to reconstruct phylogenies of mutations and to trace the origins of polymorphisms as well as their geographical spread, which is not possible with autosomal data due to recombination. Furthermore, uniparental markers greatly enable the detection of culturally determined sex-biased events, such as patrilocality or matrilocality or polygyny (cf. Kayser et al. 2006, 2008). Because patrilocality and/or polygyny are common social practices in sub-Saharan Africa (Pebley et al. 1988), the Y chromosome is expected to retain a clearer signal of demic migration events because the mtDNA and autosomes brought by marrying local women could with time dilute the original genetic composition.

The aim of this paper is to investigate in more detail the combined Y chromosomal variation of biallelic and

Table 1. Details of the 26 Populations Included in This Study With Approximate Geographic Coordinates.

Group	Code	Sample Size	Latitude	Longitude	Linguistic Affiliation[a]	Country[b]	References
Algeria	ALG-AA	20	32.0	3.0	Afro-Asiatic	Algeria	present study
Angola Bantu	ANG-B	230	−17.0	15.0	NC—Bantu	Angola	Coelho et al. (2009)
Botswana Bantu	BOT-B	40	−24.7	25.9	NC—Bantu	Botswana	present study
Burkina Faso Gur	BF-G	183	13.0	−1.5	NC—Gur	Burkina Faso	present study
Burkina Faso Mande	BF-M	152	12.6	-3.6	NC—Mande	Burkina Faso	present study
C.A.R. Pygmies	CAR-P	23	4.0	17.0	Various	C.A.R.	present study
Cameroon Bantu	CAM-B	28	5.0	11.0	NC—Bantu	Cameroon	Berniell-Lee et al. (2009)
Cameroon Pygmies	CAM-P	27	5.0	13.4	NC—various	Cameroon	Berniell-Lee et al. (2009)
D.R.C. Bantu	DRC-B	58	−5.0	18.8	NC—Bantu	D.R.C.	present study
D.R.C. Pygmies	DRC-P	11	1.0	29.0	Nilo-Saharan	D.R.C.	present study
Ethiopia	ETH-AA	98	9.0	38.7	Afro-Asiatic	Ethiopia	present study
Gabon Bantu	GAB-B	795	−0.7	12.0	NC—Bantu	Gabon	Berniell-Lee et al. (2009)
Gabon Pygmies	GAB-P	33	0.5	13.6	NC—Ubangi	Gabon	Berniell-Lee et al. (2009)
Kenya Bantu	KEN-B	10	−3.0	37.0	NC—Bantu	Kenya	present study
Kenya Nilo-Saharan	KEN-NS	79	0.5	36.0	Nilo-Saharan	Kenya	present study
Namibia	NAM-K	6	−21.0	20.0	Khoisan	Namibia	present study
Nigeria	NIG-Y	12	8.0	5.0	NC—Yoruboid	Nigeria	present study
Senegal	SEN-M	15	14.0	−14.0	NC—Mande	Senegal	present study
South Africa Bantu	SA-B	8	−29.0	26.0	NC—Bantu	South Africa	present study
Tanzania Afro-Asiatic	TZ-AA	25	−2.8	36.0	Afro-Asiatic	Tanzania	Tishkoff et al. (2007)
Tanzania Bantu	TZ-B	64	−4.0	33.0	NC—Bantu	Tanzania	Tishkoff et al. (2007)
Tanzania Khoisan	TZ-K	121	−3.1	34.4	Khoisan	Tanzania	Tishkoff et al. (2007)
Tanzania Nilotic	TZ-NS	31	−2.1	35.4	Nilo-Saharan	Tanzania	Tishkoff et al. (2007)
Uganda	UGA-NS	118	2.7	34.3	Nilo-Saharan	Uganda	Gomes et al. (2010)
Zambia East Bantu	ZAE-B	69	−15.5	23.0	NC—Bantu	Zambia	de Filippo et al. (2010)
Zambia West Bantu	ZAW-B	480	−12.0	31.0	NC—Bantu	Zambia	present study

NOTE.—[a]NC refers to Niger-Congo linguistic phyla.
[b] C.A.R. stands for Central African Republic and D.R.C. for Democratic Republic of Congo

microsatellite markers in sub-Saharan Africa to gain insights into (pre)historic population movements, in particular those associated with the spread of the Niger-Congo language phylum. In order to obtain a more fine-grained coverage of the Y chromosomal diversity in the continent, we analyze over 1,100 samples from several populations belonging to the major linguistic phyla in West, Central, and East Africa and combine these with published data. We analyze the distribution of subclades of the widespread E1b1a lineage to obtain a more detailed view of the genetic variation present in the Niger-Congo phylum and to investigate the potential genetic effects of the Bantu migration. Furthermore, we investigate the two main hypotheses about the spread of Bantu languages over sub-Saharan Africa: a mere cultural diffusion (so-called "language shift"; Nichols 1997 and Sikora et al. 2010) or an actual movement of people via a demic diffusion (Diamond and Bellwood 2003).

Materials and Methods

Samples

A total of 1,090 saliva samples or buccal swabs were collected from healthy male volunteers after obtaining informed consent. About 480 samples from Bantu speakers from the Western Province of Zambia were collected in 2007 by C.d.F., E.D.G., T.N., K.Bo., B.P., and M.S.; 58 samples from Bantu speakers from the Democratic Republic of Congo (D.R.C.) were collected by C.d.F., K.Bo., and Joseph Koni Muluwa in 2008; 335 samples from Burkina Faso (speaking either Niger-Congo Mande or Gur languages) were collected by M.W., H.S., and K.Be. in 2008; 40 samples from Bantu speakers from Botswana were collected by S.W.M. in 2010; 98 samples from Ethiopians speaking Afro-Asiatic languages and 79 samples of Nilo-Saharan speakers from Kenya were collected by collaborators of D.L. in 2003, 2007, and 2008. DNA was extracted from the saliva samples from Botswana, Burkina Faso, D.R.C., and Zambia following the method previously described by Quinque et al. (2006). DNA extraction from the buccal swab samples from Ethiopia and Kenya was performed following the procedure described in Miller et al. (1988).

In addition, 85 unrelated sub-Saharan African individuals from the Human Genome Diversity Cell Line Panel (Cann et al. 2002) as identified by Rosenberg (2006) were included in the analyses. These include the Biaka Pygmies from the Central African Republic (C.A.R.), Mbuti Pygmies from D.R.C., Bantu speakers from Kenya, Khoisan from Namibia, Niger-Congo Yoruba from Nigeria, Niger-Congo Mandenka from Senegal, and Bantu speakers from South Africa. Furthermore, to bolster the number of Afro-Asiatic groups included in this study, the Afro-Asiatic–speaking Mozabites from Algeria were also genotyped, even though they do not belong to the geographic region of sub-Saharan Africa as such.

For the purposes of this study, the data set has been divided into 26 major geographic and/or linguistic groups as summarized in table 1 (for details of the ethnolinguistic affiliation of the groups as determined by self-identification, see supplementary table 2, Supplementary Material online).

Markers

The Nilo-Saharan samples from Kenya and some of the Ethiopian samples were initially screened at the University of Bologna through restriction fragment length polymorphism analysis of the biallelic markers M42 and M60, which define the A and B lineages, respectively. The remaining 1,174 samples were genotyped for 24 SNPs (12f2, M106, M124, M145, M168, M170, M172, M174, M175, M20, M201, M207, M213, M214, M269, M45, M52, M69, M9, M91, M96, MEH2, SRY10831, and Tat) defining the major branches of the Y chromosome tree (Karafet et al. 2008). These sites were amplified in a multiplex polymerase chain reaction (PCR) and then typed by means of two SNaPshot assays consisting of 12 SNPs each following the manufacturer's specifications (Applied Biosystems, http://www3.appliedbiosystems.com). We further genotyped seven additional SNPs (M33, M35, M2, M191, M75, U174, and U175) on those individuals ascertained to be haplogroup E for a deeper characterization of this lineage (fig. 2) in an additional multiplex PCR and SNaPshot assay. Subhaplogroups of haplogroup E have been defined according to the nomenclature specified in Karafet et al. (2008): E1b1a* (xE1b1a8 and xE1b1a7), E1b1a8, E1b1a7* (xE1b1a7a), E1b1a7a, E* (xE1b1a, xE1a, xE1b, and xE2). Genotyping details are listed in supplementary table 1 (Supplementary Material online). The markers U174 and U175 were additionally typed for this study in the samples from Eastern Zambia that had previously been genotyped for the other markers (de Filippo et al. 2010). Finally, we genotyped 12 short tandem repeat (STR) loci (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) by means of the Promega Y-Powerplex kit (http://www.promega.com). When two peaks were detected in the duplicated STR locus DYS385, the smaller allele was arbitrarily assigned to DYS385a and the larger to DYS385b. Both SNP and STR genotyping were performed on the ABI 3130xl Genetic Analyzer and analyzed using the GeneMapperID v3.2 software (Applied Biosystem).

Comparative Data

In order to extend our study of Y chromosomal variation to a wider geographical coverage of sub-Saharan Africa, we included published data sets having a similar amount of SNP and STR genotype information as our data. The published data were classified on geographic and linguistic grounds as follows: Khoisan, Afro-Asiatic, Nilo-Saharan, and Bantu speakers from Tanzania (Tishkoff et al. 2007); non-Pygmy Bantu speakers and Pygmies (Bantu and non-Bantu speakers) from both Cameroon and Gabon (Berniell-Lee et al. 2009); Bantu speakers from Angola (Coelho et al. 2009); and a Nilo-Saharan group from Uganda (Gomes et al. 2010). However, these studies genotyped individuals belonging to haplogroup E only to the level of E1b1a, with the exception of Gomes et al. (2010) who additionally genotyped M191. We therefore inferred the frequency of the haplogroup E sublineages studied here—namely E1b1a8, E1b1a7a, and E1b1a7*—from the STR haplotypes using Linear Discriminant Analysis (LDA) with the R statistical software by means of the function "lda" from the package MASS (Venables and Ripley 2002). Because Tishkoff et al. (2007) and Coelho et al. (2009) subtyped only M2 and M35 on the haplogroup E samples, we also applied LDA to those individuals who were E*(xE1b1a and xE1b1b1). Of these, the individuals from Tishkoff et al. (2007) being possibly haplogroup D or E (i.e., carrying the YAP mutation) were considered as belonging to haplogroup E under the assumption that haplogroup D is virtually absent in the African continent (Jobling and Tyler-Smith 2003; Wood et al. 2005). We tested the power of LDA to reliably infer haplogroups from STR haplotype data as described in the supplementary text (Supplementary Material online) before applying it to the above-mentioned data sets. However, it should be kept in mind that the comparative data inferred by LDA may not be as reliable as our genotyped data.

Data Analyses

Standard measures of genetic diversity, pairwise genetic distances between groups expressed as R_{ST} and proportion of haplotypes not shared were calculated in *R*. Correspondence analysis (CA) of haplogroup frequencies in all populations was performed using the function "ca" from the R package ca (Nenadic and Greenacre 2007). Analysis of molecular variance (AMOVA) and pairwise F_{ST} between groups were carried out with Arlequin software v3.1 (Excoffier et al. 2005) based on haplogroup frequencies. A matrix of geographic great circle distances between all groups (with the exclusion of populations with less than ten individuals) was generated. We performed a Mantel test (*Z* value) to investigate whether the geographic distances are correlated with genetic distances. Individuals who had STR missing values were excluded from some analyses.

Patterns of haplotype sharing among groups were explored as follows. STR haplotypes that were shared among at least three groups were ranked based on their frequency in the entire combined data set. We explored the distribution of shared haplotypes among groups that were merged (here called metagroups) according to their geographic location as well as their linguistic affiliation (and ethnicity in the case of the Pygmies, who are known to have acquired their language from their agriculturalist neighbors). With regard to linguistic affiliation, individuals from Western Zambia who speak a language belonging to the Eastern Bantu branch (Fortune 1970; Bostoen 2009) were classified with the Bantu speech communities from Eastern Zambia. To test if the observed patterns simply reflect sample size differences among the various metagroups, we randomly assigned the shared haplotypes to groups and subsequently merged the groups into the various metagroups. We repeated this process 1,000 times and recorded the number of haplotypes shared between each pairwise comparison of metagroups to estimate the significance level.

The average squared distance (ASD) statistic (Goldstein and Pollock 1997) was calculated to estimate the time since

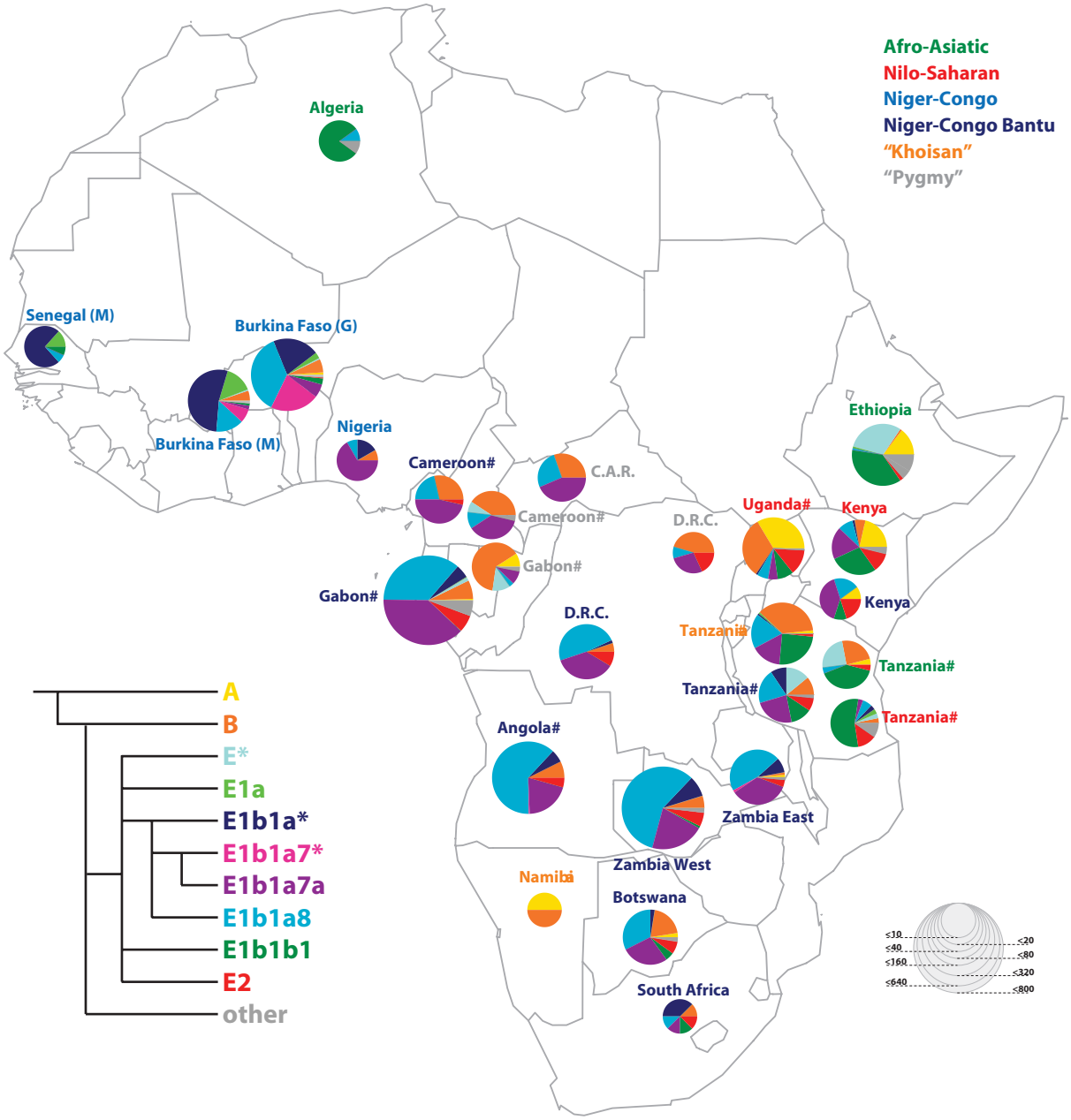

FIG. 2 Haplogroup composition of the combined data set. The size of the pie charts is proportional to the sample size as shown in the bottom right. Groups marked with # indicate that the subhaplogroup composition of E1b1a was inferred by LDA. Only the major African haplogroups (A, B, and subhaplogroups of E) are displayed; the remaining haplogroups are lumped under the label "other." Population labels are color coded according to linguistic phyla as indicated in the upper right, with Pygmy groups (gray) indicated separately from other groups.

the most recent common ancestor (tMRCA) for ten microsatellites (excluding DYS385a/b). Under the Stepwise Mutation Model, the tMRCA is expected to be ASD/2μ, where μ is the mutation rate per generation per locus, averaged across loci. Therefore, to calculate the tMRCA and associated confidence intervals (CI), the mutation rates reported in the Y-STR haplotype reference database (http://www.yhrd.org) were used, and a generation time of 25 years was considered.

Because Pygmy groups are commonly believed to have shifted from their original language to that of their agricultural neighbors, which makes their current linguistic affiliation misleading, they were considered as a separate ethnic unit, regardless of the language they speak, and excluded from the AMOVA analysis.

Results

Y Chromosome Haplogroups in Sub-Saharan Africa

Figure 2 shows the haplogroup composition for 2,736 samples belonging to 26 groups (see references in table 1). STR haplotypes and SNP haplogroups genotyped here as well as those inferred by LDA (with associated relative posterior probabilities) are reported in supplementary table 2 (Supplementary Material online) and the phylogenetic relationships of the SNPs typed are in supplementary figure

3 (Supplementary Material online). Overall, the haplogroup composition in all the groups reflects what has been previously observed in the African continent, with A (mainly present in Khoisan speakers and Eastern groups), B (mainly found in hunter-gatherer Pygmies and Khoisan as well as their neighbors), and E (in almost all groups) representing the majority (87%) of the haplogroups.

Haplogroup E1b1a (including all its sublineages typed in this study) is present in all groups (excluding the Namibian Khoisan) and was found at a frequency of ~68.5% in the entire data set. This is in agreement with previous studies of African Y chromosomal variation (Wood et al. 2005; Tishkoff et al. 2007; Berniell-Lee et al. 2009). With respect to the sublineages of E1b1a typed here, the most frequent haplogroup in the combined data set was E1b1a8 (~35%), which was found in all groups except in the Namibian Khoisan (which are, however, represented by only six individuals). All Bantu-speaking groups showed relatively high frequencies of this haplogroup, ranging from 18% to 62%, with the exception of the South African Bantu where the frequency was only 12.5%; however, this is due to the small sample size and not significantly different from the other groups (95% CI of sampling error = 3–53%). The second most common haplogroup, E1b1a7a, is present in African populations with an average frequency of 23% and shows moderately high frequencies in all Bantu and Pygmy groups. The highest frequencies are found in Nigeria (67%) and Bantu speakers from Cameroon (46%), which are both regions that are close to the putative homeland of the Bantu languages.

Another common haplogroup within haplogroup E is E1b1a* (xE1b1a8 and xE1b1a7) with an average frequency of 8.9%, which is a characteristic of all West African groups included here, with the highest frequencies in Mande speakers from Senegal (75%) and Burkina Faso (53%). Haplogroup B is also widespread, being found on average in 10.3% of the African groups included here.

Patterns of Y-STRs Diversity

Y-STR diversity values within specific haplogroups can be informative for discerning origins and migrations of haplogroups: In general, the highest diversity should be found in the population where the haplogroup originated, and lower diversity (due to successive founder events) may be associated with migrations. However, because STRs have a high mutation rate, these signals might be erased over time, and it can be insightful to examine the variance in repeat units. The STR variance has been described as evolutionary more stable and is correlated with the time that has elapsed since a haplogroup-defining mutation arose, thus serving as a rough estimator of tMRCA as well (Goldstein and Pollock 1997; Bosch et al. 1999). Yet, because the results of such estimates depend to a large extent on the mutation rates used, which are very variable and subject to considerable debate (Zhivotovsky et al. 2004), age estimations should be considered with due caution.

Values of diversity for 11 STR loci for all individuals as well as those carrying the E1b1a*, E1b1a7a, and E1b1a8 clades are reported in table 2. In general, regardless of the haplogroup composition and excluding populations with sample size less than ten individuals, Niger-Congo–speaking groups have slightly higher haplotype diversity than Nilo-Saharan–speaking groups (Mann–Whitney U test: $W = 27$, P value < 0.017), but these together have higher diversity values than Afro-Asiatic, Khoisan, and Pygmy groups ($W = 123.5$, P value < 0.005).

The STR haplotype diversity associated with E1b1a8 was found to be higher ($W = 45$, P value < 0.004) in all Bantu-speaking groups (except in Cameroon with a low sample size = 6) than all the other groups after removing groups with less than five individuals. However, the STR variance showed a different pattern with the highest values in Pygmies from C.A.R. (with a sample size of six) and Burkina Mande, whereas reduced values were found in all the other groups. Moreover, STR variances did not differ significantly among groups ($W = 24$, P value $= 0.526$).

For haplogroup E1b1a7a, the STR haplotype diversity levels were high (>0.90) in all groups, with the lowest values observed in Pygmies from C.A.R., Tanzanian "Khoisan," and the two Nilo-Saharan groups. Similar to E1b1a8, the highest STR variance for E1b1a7a was found in the C.A.R. Pygmies (0.49); however, the Bantu speakers from West Zambia and the Burkina Faso Gur speakers also had high STR variances (0.47 and 0.43, respectively).

With regard to the diversity associated with haplogroup E1b1a*, Niger-Congo non-Bantu have higher haplotype diversity and STR variance than the Bantu-speaking groups. Overall, there is some support for an association of E1b1a8 with higher diversity in Bantu-speaking groups and of E1b1a* with higher diversity in Niger-Congo non–Bantu-speaking groups. However, none of these patterns reach statistical significance: for E1b1a8 $W = 54$, P value $= 0.125$ and for E1b1a7a $W = 20$, P value $= 0.057$.

The tMRCA estimates for haplogroups E1b1a7 and E1b1a8 were calculated by means of the ASD statistic for the major ethno-linguistic groups (table 3). The highest tMRCA (~4,200 ya) for E1b1a7a was ascertained in the Yoruba from Nigeria, whereas the lowest (~2,000 ya) was in the Nilo-Saharans. With regard to E1b1a8, the highest tMRCA (~ 5,000 ya) was found in Mande speakers from both Burkina Faso and Senegal, whereas the lowest (~3,400 ya) was in the Bantu. The 95% CIs all overlap; overall, all these estimates are consistent with the time of the Bantu expansion (5,000–3,000 ya) and with an origin of both haplogroups in an area between West and Central Africa a few thousand years before the beginning of the expansion as indicated by the upper limits of the CIs.

Genetic Structure Within and Between Groups in Sub-Saharan Africa

To visualize the relationships among the different groups within sub-Saharan Africa, a CA was performed on the haplogroup frequencies (fig. 3). The first two dimensions together accounted for 59.2% of the total inertia and reflect both geographic and linguistic groupings. In the first dimension, the Niger-Congo–speaking groups and

Table 2. Diversity Values Based on 11 Y-STR Loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b), Where *N* is the Sample Size, HD is the Haplotype Diversity with its Standard Deviation (SD), and STR Var is the Variance of Repeat Units Averaged Across All 11 STR loci.

	ALL			E1b1a8			E1b1a7a			E1b1a*		
Group[a]	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var
Bantu speakers												
ANG-B	230	0.992 (0.002)	1.35	143	0.982 (0.005)	0.32	46	0.987 (0.009)	0.36	13	0.962 (0.041)	0.38
BOT-B	39	0.993 (0.007)	2.57	13	1.000 (0.030)	0.40	10	0.933 (0.62)	0.19	1	—	—
CAM-B	28	0.992 (0.012)	3.63	6	0.933 (0.122)	0.31	13	0.987 (0.035)	0.27	0	—	—
DRC-B	43	0.992 (0.007)	0.97	21	0.990 (0.018)	0.39	16	0.967 (0.036)	0.26	0	—	—
GAB-B	795	0.997 (0.000)	1.67	289	0.993 (0.001)	0.39	303	0.992 (0.002)	0.39	39	0.966 (0.014)	0.40
KEN-B	10	1.000 (0.045)	1.56	2	1.000 (0.500)	0.41	4	1.000 (0.177)	0.11	0	—	—
SAB	8	1.000 (0.063)	2.72	1	—	—	1	—	—	3	1.000 (0.272)	0.33
TZ-B	64	0.999 (0.003)	2.67	13	0.987 (0.035)	0.32	15	0.990 (0.028)	0.35	6	0.933 (0.122)	5.65
ZAW-B	473	0.995 (0.001)	1.12	277	0.987 (0.002)	0.30	100	0.995 (0.002)	0.47	37	0.964 (0.018)	0.25
ZAE-B	69	0.997 (0.003)	0.83	32	0.992 (0.011)	0.44	24	0.989 (0.017)	0.30	6	1.000 (0.096)	0.35
Niger-Congo non-Bantu speakers												
BF-G	173	0.994 (0.002)	1.32	65	0.973 (0.010)	0.46	11	1.000 (0.039)	0.43	36	0.992 (0.009)	0.70
BF-M	148	0.988 (0.004)	1.28	21	0.981 (0.023)	0.74	2	1.000 (0.500)	0.50	81	0.972 (0.012)	0.57
NIG-Y	12	1.000 (0.034)	1.34	1	—	—	8	1.000 (0.063)	0.34	2	1.000 (0.500)	0.55
SEN-M	15	0.990 (0.028)	0.81	1	—	—	0	—	—	11	0.982 (0.046)	0.55
Hunter gatherers												
CAM-P	27	0.980 (0.016)	4.08	3	1.000 (0.222)	0.14	10	0.956 (0.059)	0.23	0	—	—
CAR-P	23	0.964 (0.022)	4.41	6	0.800 (0.237)	0.84	10	0.911 (0.077)	0.49	0	—	—
DRC-P	11	0.964 (0.051)	4.36	1	—	—	3	1.000 (0.272)	0.45	0	—	—
GAB-P	33	0.936 (0.026)	4.00	1	—	—	3	0.667 (0.314)	0.12	0	—	—
NAM-K	4	1.000 (0.177)	2.89	0	—	—	0	—	—	0	—	—
TZ-K	121	0.982 (0.004)	2.51	22	0.970 (0.024)	0.27	19	0.936 (0.037)	0.33	1	—	—
Nilo-Saharan												
KEN-NS	45	0.990 (0.007)	1.31	6	0.800 (0.172)	0.21	10	0.933 (0.062)	0.24	0	—	—
TZ-NS	31	0.991 (0.012)	5.79	2	1.000 (0.500)	0.27	1	—	—	1	—	—
UGA-NS	118	0.988 (0.003)	2.24	7	0.905 (0.103)	0.27	6	0.933 (0.122)	0.18	1	—	—
Afro-Asiatic												
ALG-AA	20	0.963 (0.033)	0.93	2	1.000 (0.500)	0.05	0	—	—	0	—	—
ETH-AA	64	0.980 (0.007)	1.02	0	—	—	0	—	—	0	—	—
TZ-AA	25	0.963 (0.021)	6.53	1	—	—	0	—	—	0	—	—

NOTE.—[a] The group codes correspond to those reported in table 1.

Pygmies (except those from Gabon) all have values less than 0.5, and all other groups have values greater than 0.5. The Afro-Asiatic groups cluster together and the Nilo-Saharan groups from Kenya, Uganda, and Tanzania are also located close to each other along the first dimension. The eastern Bantu speakers from Tanzania (and to a minor extent from Kenya) are closer to the other East African populations than are the other Bantu–speaking groups as a result of their modest frequencies of haplogroups A and E*, respectively. Dimension 2 largely divides the Niger-Congo populations into Bantu and non-Bantu, with the Western samples (Senegal and Burkina Faso) with highest values, driven by haplogroups E1a, E1b1a7*, and E1b1a*.

Table 3. Estimates of tMRCA (in years ago) of the Two Major Haplogroups (E1b1a7a and E1b1a8) Using ASD Statistic With 10 STRs (excluding DYS385a/b) and a Generation Time of 25 Years.

	E1b1a7a			E1b1a8		
Groups	*N*[a]	Mean	95% CI	*N*[a]	Mean	95% CI
NC—Bantu	532	3,238	2,022–6,792	798	3,396	1,933–8,951
NC—Gur	11	2,583	1,806–3,917	65	3,458	2,444–5,543
NC—Mande	2	—	—	22	4,987	3,164–10,281
NC—Yoruba	8	4,249	2,498–10,181	1	—	—
Pygmies	26	3,707	2,629–5,468	11	3,889	2,298–10,205
Khoisan	19	2,396	1,608–3,831	22	3,484	1,771–11,263
Nilo-Saharan	17	2,049	1,326–3,595	15	4,066	2,068–12,288

NOTE.—[a]Number of STR-haplotypes used.

To test whether the genetic structure was in better accordance with linguistic or geographic groupings, AMOVA analyses were performed (table 4). As mentioned in the Methods section, the four Pygmy populations were excluded from these analyses because of their assumed recent language shift. Both linguistic affiliation and geographic location are in good agreement with the Y chromosomal variation because the variance between groups is always higher than that between populations within a group. The variance among all the populations included in the study accounts for 15.4% of the total. When these are grouped according to their classification in one of the four major linguistic phyla, the between-group variability reaches 14.8%, whereas the variance within the linguistically defined groups is 8.7%. Grouping populations by geography into North, West, East, Central, and South Africa decreased the between-group variability to 9.96% and the variance within groups to 6.75%. When only

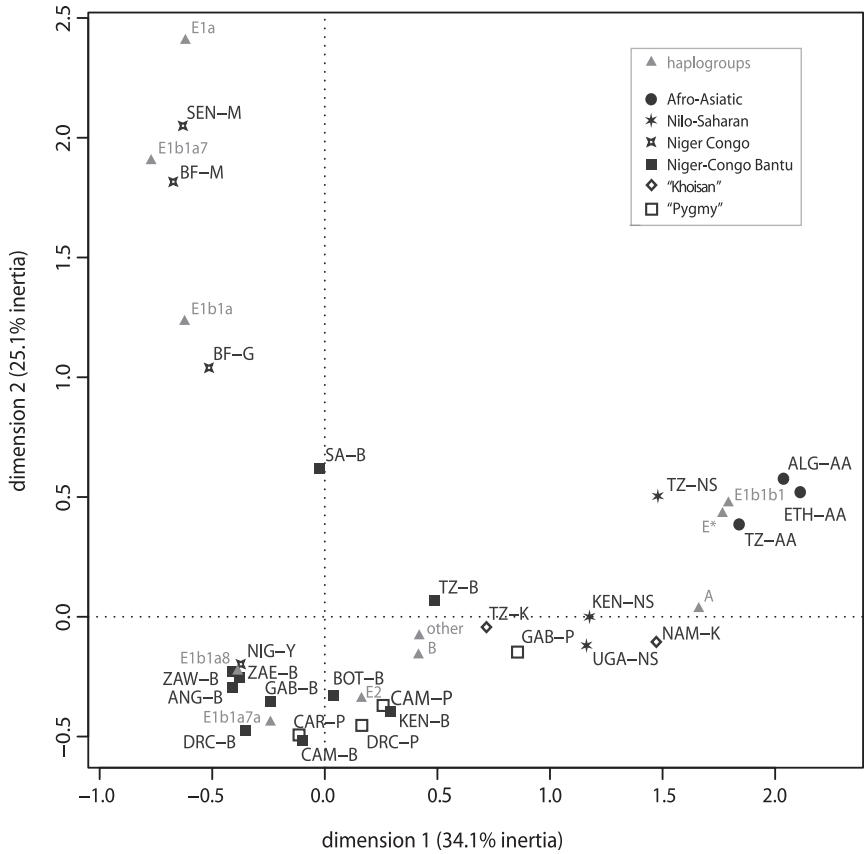

Fig. 3 Correspondence analysis performed on haplogroup frequencies. The population labels correspond to those reported in table 1.

Bantu-speaking populations were compared, the proportion of variance explained by differences between populations is much lower but still significant (4.7%, *P* value = 0).

We performed another AMOVA to quantify the differences between Niger-Congo, non-Bantu, and Bantu populations (see fig. 1). This highlighted a large amount of variation (11.6%, *P* value < 0.018) due to differences among groups and only 5.31% within groups. When performing this AMOVA with the lower haplogroup resolution used in previous studies (e.g., Wood et al. 2005)—that is, only E1b1a*(xE1b1a7) and E1b1a7 without their subhaplogroups E1b1a8 and E1b1a7a—the proportion of variation observed between Bantu and non-Bantu became nonsignificant (0.28%, *P* value = 0.35). This is a strong indication that the more fine-grained haplogroup genotyping used here adds considerably to our power to detect genetic substructure in Africa.

Mantel tests of correlation between geographic and genetic distances further confirmed that geography has had an important influence on Y chromosomal diversity

Table 4. AMOVA Based on Haplogroup Frequencies.

Number of Groups	Grouping[a]	Total Number of Populations	Proportion of variation (%)		
			Among Groups	Among Populations Within Group	Within Populations
1	All populations	22	—	15.39**	84.61**
1	Bantu	10	—	4.69**	95.31**
5	Geography[b]	22	9.96**	6.75**	83.29**
4	Language[c]	22	14.08**	8.68**	77.24**
2	Niger-Congo[d]	14	11.58*	5.31**	83.10**
2	Niger-Congo (low)[e]	14	0.28	5.67**	94.06**

Note.—All values are significant with *P* value < 0.05* and *P* value < 0.01**, except for that in boldface.

[a] Pygmy groups were excluded because they are known to have undergone language shift.

[b] Geographic subdivision as follows: North (Algeria), West (Senegal, Burkina Faso, and Nigeria), Central (Cameroon, D.R.C., and Gabon), East (Ethiopia, Kenya, Tanzania, and Uganda), and South (Angola, Zambia, Botswana, Namibia, and South Africa).

[c] Linguistic grouping with the four major African phyla: Afro-Asiatic, "Khoisan," Niger-Congo, and Nilo-Saharan.

[d] Niger-Congo Bantu vs. non-Bantu.

[e] Niger-Congo Bantu vs. non-Bantu with a lower haplogroup resolution: E1b1a*(xE1b1a7) and E1b1a7. See main text for details.

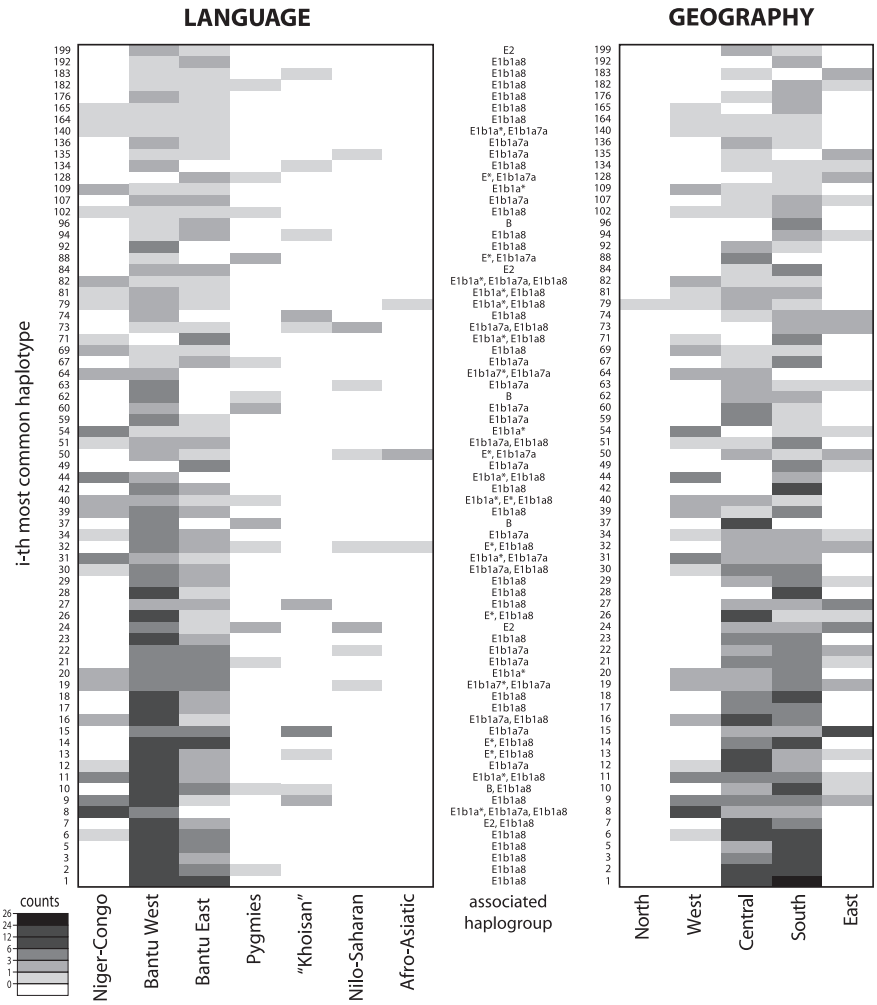

Fig. 4 Patterns of haplotype sharing. Heat plots showing the count of the most common haplotypes from 11 STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b) shared among at least three individual groups. Individual groups are combined into metagroups according to their linguistic affiliation (left) and geographic location (right); the same heat plot, but for single groups, is reported in supplementary figure 5 (Supplementary Material online).

in Africa. Indeed, both pairwise F_{ST} and R_{ST} matrices were correlated with the matrix of great circle geographic distances: $Z = 0.47$ (one-tail P value < 0.001) and 0.26 (one-tail p value < 0.015), respectively. When only Niger-Congo groups were considered, F_{ST} values were correlated with geography ($Z = 0.50$, one-tail P value < 0.001), but R_{ST} values were not ($Z = -0.02$, one-tail P value $= 0.51$). In contrast, the correlation of R_{ST} and geographic distances was still present when all the other groups (excluding Niger-Congo) were considered. In addition, pairwise R_{ST} values between groups were calculated for haplogroups E1b1a7a, E1b1a8, and E1b1a* and compared with the geographic distances between them. Only R_{ST} values associated with haplogroups E1b1a8 and E1b1a* exhibited a correlation with geographic distances, with $Z = 0.36$ (one-tail P value < 0.03) and 0.67 (one-tail P value $= 0.034$), respectively. However, because the dimension of the matrices might have an effect on the significance of the Mantel test, we controlled for the number of groups by redoing the test using only those groups that have both E1b1a7a and E1b1a8. In this test, no correlation was observed between geographic distances and pairwise R_{ST} for either haplogroups E1b1a7a or E1b1a8.

Distribution of Shared Haplotypes

Contrary to the geographical and linguistic structure apparent in the haplogroup data, a network based on 11 STR loci showed no structure at all; rather, haplotypes from East African and Central African Bantu groups are found clustered together. The extensive reticulation made it difficult to observe any patterns of overall haplotype sharing (supplementary fig. 4, Supplementary Material online). Therefore, in order to elucidate the relationships among groups from different geographic areas that may be due to common origin and/or recent migration, the combined data set was screened for widespread and shared haplotypes. Figure 4 shows the distribution of shared haplotypes among groups that were merged (here called metagroups as described in the Material and Methods), whereas the haplotype-sharing patterns for individual populations are shown in supplementary figure 5 (Supplementary Material online). The total number of haplotypes shared by at least

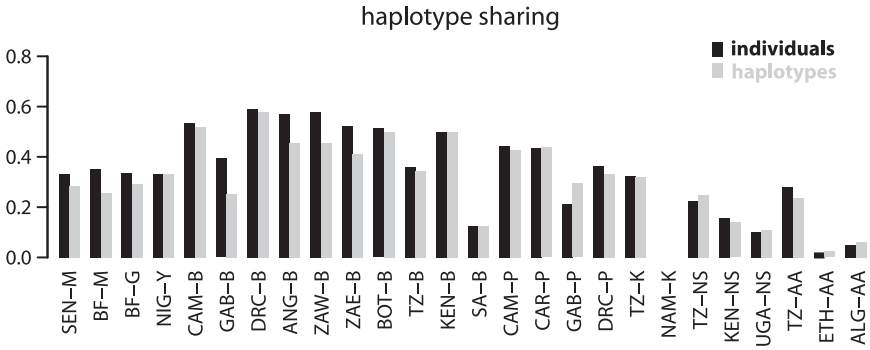

Fig. 5 Proportion of shared haplotypes. Histogram of the proportion of shared haplotypes between one group and all other groups based on 11 STRs. Black bars represent the proportion of all individuals sharing their haplotype (with any of the other groups) over the total number of individuals in a group; gray bars represent the proportion of unique shared haplotypes over the total number of haplotypes detected in a group.

three groups was 73, which is significantly less than expected if individuals are assigned to groups at random (mean = 166, range = 152–183; P value < 0.001 based on 1,000 permutations). This analysis indicates that there is a significant effect of population structure on the shared haplotypes and also indicates that the observed pattern was not caused by differences in group sample sizes. None of the 73 shared haplotypes was shared across all the metagroups. Also, no haplotype was found in all the groups within each metagroup (supplementary fig. 5, Supplementary Material online).

When grouped according to linguistic/ethnic affiliation, the West Bantu metagroup, which includes samples from Cameroon, Gabon, D.R.C., Angola, and Western Zambia and corresponds to the majority of the data set, shares 69 of 73 haplotypes with at least one of the other metagroups. Nilo-Saharan and Afro-Asiatic groups shared a low proportion of haplotypes with all other groups, ranging from 1 to 8 and from 0 to 3, respectively.

When grouped according to geography, the Southern and Central African metagroups share the most haplotypes (55), with fewer haplotypes shared between Central and Western Africa (23), Central and Eastern Africa (21), or Western and Southern Africa (26). The presence of significant structure detectable in this analysis in the STR data (which are subject to different patterns of mutation and variation as compared with the more stable haplogroup data) contrasts with the lack of structure in the network but is in good accordance with the results seen in the CA and AMOVA. This provides further indication that the inferred haplogroup frequencies are fairly accurate because the STR data were all genotyped.

To what extent do these haplotype-sharing patterns (fig. 4) simply reflect sample size differences among the various metagroups? The results of our permutation test (described in the Material and Methods and shown in supplementary table 3, Supplementary Material online) indicate that for the linguistic metagroups, the Western and Eastern Bantu do share more haplotypes than expected by chance, whereas the Niger-Congo (non-Bantu) shares significantly fewer haplotypes than expected by chance with the Pygmy, Nilo-Saharan, and Afro-Asiatic metagroups. Similarly, for the geographic metagroups, there is significantly more sharing between Central and Southern Africa and significantly less sharing between Eastern Africa and all other groups (except Southern Africa). Overall, this test demonstrates that the haplotype-sharing patterns in figure 4 do indicate population relationships and not just overall sample size differences between metagroups. In particular, there is more haplotype sharing than expected by chance involving groups toward the center of Africa (i.e., Western and Eastern Bantu and Central and Southern Africa). Moreover, the Bantu from D.R.C.—who are located in the center of the geographic area studied herein (fig. 2) and who are on average closest geographically (≈2,022 km) to all other African populations—shows the highest proportion of shared haplotypes with other groups (fig. 5).

Discussion

Haplogroup Variation Within Niger-Congo Speech Communities and Sub-Saharan Africa

The Niger-Congo phylum is one of the major language groups in the world and is the largest in the African continent in terms of number of languages, number of speakers, and geographical area it covers. To a certain extent, the linguistic branching pattern displayed in figure 1 is paralleled by Y chromosomal markers characteristic of the different subgroups of the Niger-Congo phylum included here: Mande, Gur, and Bantu. Indeed, haplogroups E1b1a* and its derivative E1b1a8 are characteristic of the Mande, which belong to the earliest split of the linguistic tree. The derived haplogroup E1b1a7* is characteristic of Gur speakers, and the most derived haplogroup analyzed here, E1b1a7a, is characteristic of Bantu-speaking groups, who represent one of the most derived branches of the Niger-Congo linguistic tree.

Although previous genetic studies on Y chromosome variation have linked haplogroup E1b1a and its sublineage E1b1a7 (when genotyped) specifically to the Bantu

expansion (Thomas et al. 2000; Cruciani et al. 2002; Zhivotovsky et al. 2004; Wood et al. 2005; Berniell-Lee et al. 2009), our results demonstrate that this association extends to all of Niger-Congo, not just Bantu. Indeed, E1b1a does not differ in frequency between Niger-Congo non-Bantu and Bantu, and this is also true if E1b1a7 is taken into account. In fact, an AMOVA with the haplogroup resolution used previously (Wood et al. 2005), that is, only E1b1a*(xE1b1a7) and E1b1a7—for Bantu versus Niger-Congo non-Bantu results in nonsignificant variation (0.28%, P value = 0.35) between these two groups. Therefore, to increase resolution, we for the first time analyzed two additional markers (U174 and U175) in a large number of African populations, resulting in a total of four E1b1a sublineages. Notably, the AMOVA carried out with this increased haplogroup resolution now finds significant variation between Bantu and Niger-Congo non-Bantu (11.58%, P value < 0.018). In addition, with these new markers, we were able to detect the presence of substructure even within the Niger-Congo non−Bantu-speaking groups as described below.

Niger-Congo non−Bantu-speaking groups in West Africa are distinct from Bantu speakers and groups belonging to the other African phyla as shown in the CA plot (fig. 3). This distinct position is mainly driven by haplogroup E1b1a* (almost absent in all non−Niger-Congo groups), which has high frequencies in Mande speakers and exhibits a clinal reduction from western toward eastern and southern Africa. A strong positive correlation was ascertained between the haplotype diversity levels and STR variance associated with E1b1a*. These results suggest that this haplogroup was present for a longer time in Western Africa—which is the presumed place of origin of the defining M2 mutation (Rosa et al. 2007)—and that two of the derived mutations considered here (e.g., M191 and U174) did not occur in the ancestors of the Mande; the low frequencies of E1b1a7a found in these groups could be due to later admixture. On the other hand, only Gur speakers are characterized by the presence of haplogroup E1b1a7*, which was previously associated with the Bantu expansion with a probable origin in western Central Africa (Underhill et al. 2000; Cruciani et al. 2002; Zhivotovsky et al. 2004; Wood et al. 2005) and that here we found practically restricted to Burkina Faso. Instead, a new sublineage of E1b1a7, namely E1b1a7a, which may also have originated in western Central Africa, is associated with the Bantu expansion. Indeed, we found that this marker has its highest frequencies in Nigerian Yoruba (where this haplogroup also appears to be oldest, with an estimated tMRCA of ~4,200 ya, cf. table 3) and Cameroonian Bantu speakers, both of whom are located close to the homeland of the Bantu languages. Furthermore, for other studies reporting high frequencies of M191 in Bantu-speaking groups, we suggest that those individuals are likely to harbor the derived mutation U174 (see, e.g., Appendix A in Wood et al. 2005). This is confirmed by the results of the LDA for the Ugandan data set where all individuals who had been genotyped as E1b1a7 were inferred to belong to E1b1a7a.

Bantu and non−Bantu-speaking groups can be distinguished by a second haplogroup, namely E1b1a8. However, we could not associate it unambiguously with the Bantu populations because the highest tMRCA estimate (~5,000 ya, table 3) was found in the Mande-speaking group and it also is found at high frequency in the Burkina Faso Gur speakers and in other western Central African populations (cf. table 1 in Veeramah et al. 2010). Nevertheless, we believe that further subtyping of markers on the background of U175 might reveal new insights concerning its association with Bantu-speaking groups (as we found with U174). Likewise, the discovery of further subclades within E1b1a7 and E1b1a8 might add more structure to the data and erase this apparent homogeneity of the Bantu groups.

The presence of both E1b1a7a and E1b1a8 in all Pygmy groups—directly genotyped in the C.A.R. and D.R.C. Pygmies and inferred from STR data for the Cameroon and Gabon Pygmies—may be the result of sex-biased migrations between agriculturalist and hunter-gatherer societies, where paternal lineages move from the former into the latter (Destro-Bisol et al. 2004; Tishkoff et al. 2007; Quintana-Murci et al. 2008). However, judging from the networks for both haplogroups (supplementary fig. 4, Supplementary Material online), recent admixture with Bantu-speaking neighbors may not account for the origin of all of these haplotypes. Although some haplotypes are shared with, or differ by only a few mutational steps from, Bantu speakers and hence may indeed reflect recent admixture, other haplotypes found at the periphery of the network are unique to Pygmies. The Pygmy groups tend to exhibit high levels of STR variance along with low levels of haplotype diversity, indicating the presence of a few very divergent (and therefore probably old) lineages. The older age of E1b1a8 in Pygmies than in Bantu, in contrast to the similar age of E1b1a7a in both Pygmies and Bantu (table 3), suggests the possibility that a few individuals belonging to haplogroup E1b1a8 were present in Pygmies prior to their contact with Bantu-speaking groups; individuals belonging to E1b1a7a were introduced at an early stage of the expansion (for instance, when the Bantu agriculturalist started to explore the rain forest), with later introgression of new haplotypes of both haplogroups after contact. Furthermore, this scenario of E1b1a7a introgression may have been mirrored on the Western side of sub-Saharan Africa as indicated by the young tMRCA estimate in Gur from Burkina Faso (table 3).

Overall, the distribution of the four E1b1a sublineages reflects what has been suggested from historical linguistic studies about the prehistory of Niger-Congo languages that had "[. . .] a long standing epicenter of spread in West Africa, with spreads through the forest and well to the south" (Nichols 1997).

Eastern Africa exhibits distinct patterns of Y chromosome haplogroups compared with Western and Central Africa. Eastern African Nilo-Saharan and Afro-Asiatic groups are characterized in general by high frequencies of lineages A and B as well as E* and E1b1b1, leading to

their clustering in the CA plot (fig. 3). The inclusion of Algeria as an additional Afro-Asiatic–speaking group, even though it is located outside sub-Saharan Africa, confirms that E1b1b1 is characteristic of Afro-Asiatic–speaking populations. It has been suggested that this marker may have spread with agropastoralist migrations from their putative origin in East Africa toward Northern Africa (Cruciani et al. 2002; Arredi et al. 2004) and Southern-Central Africa (Henn et al. 2008). In this study, E1b1b1 is absent in Angola and present at only very low frequency (<1%) in our Zambian sample but is found in appreciable frequency in Botswana (5%). This raises the question whether the demic diffusion of pastoralism from Eastern to Southern Africa followed an eastern route that circumvented Angola and Zambia or whether the later arrival of Bantu-speaking groups replaced the former pastoralist populations in Angola and Zambia but not Botswana. Investigations of samples from southeastern Africa (e.g., Mozambique and Zimbabwe) are needed to disentangle these questions.

The Nilo-Saharan samples also have relatively high frequencies of haplogroup E2. Both E2 and E1b1b1 are also common in eastern Bantu speakers, and E2 is additionally found in the D.R.C. Pygmies, possibly introduced by contact with neighboring populations. Finally, another haplogroup found in relatively high frequencies in some of the East African groups (but also present in Cameroon and Gabon Pygmies) is E*. However, because this haplogroup is defined not by a shared derived allele but by the absence of derived alleles, we cannot exclude that these individuals belong to sublineages of M96 not tested here.

In general, a similar pattern of haplogroup composition is characteristic of all neighboring groups of Eastern Africa. This appears to suggest gene flow between the groups regardless of their language; however, the low number of shared haplotypes (fig. 4) in the area (especially between eastern Bantu from Kenya and Tanzania and the Nilo-Saharan and Afro-Asiatic groups) indicates little recent contact. Possibly, the similarities in haplogroup composition are an indication of more ancient contact.

Pattern of Diversity and the Bantu Expansion(s)

In contrast to the structure observable at the level of Y chromosomal haplogroups, there is a notable absence of structure at the resolution of STR markers. There is no obvious geographic patterning to the networks (supplementary fig. 4, Supplementary Material online); in particular, haplotypes are widely shared, especially between Eastern and Western Bantu-speaking groups. There are also no clear patterns of clinal reduction in haplotype diversity and STR variance for both haplogroups E1b1a7a and E1b1a8 in the Bantu speakers (contrary to other studies, e.g., Pereira et al. 2002) as would be expected with a serial founder event of male lineages expanding from their homeland throughout sub-Saharan Africa. These data might seem to contradict the most widely cited model of the Bantu expansion, which involves the joint movement of people and language together with the diffusion of agriculture (Diamond and Bellwood 2003). However, this model has been called into question not only by linguists (Nichols 1997) and historians (Vansina 1995) but also in a recent genetic study on ~2,800 autosomal SNPs (Sikora et al. 2010). Although Nichols (1997) and Sikora et al. (2010) assert that the Bantu expansion could rather have taken place by cultural diffusion alone (i.e., "language shift" where the original inhabitants of sub-Saharan Africa would have adopted a Bantu language without major immigration of Bantu peoples), Vansina (1995) calls into question the overly simplistic assumptions of either population replacement or language shift. However, although our data do not provide evidence for the serial founder effect expected by a migration of peoples over long geographical distances—with levels of diversity (e.g., haplotype diversity and STR variance; see table 3) reduced proportionally to the distance from the homeland—the overall genetic homogeneity of the Bantu-speaking groups included here and the widespread sharing of haplotypes on the background of E1b1a7a and E1b1a8 reject the hypothesis of mere cultural diffusion. Under this assumption, one would expect greater differences between geographically distant groups because they would have developed in situ for a long time. The overall genetic homogeneity of Bantu-speaking groups was also detected in a recent study of a large number of autosomal STR loci in a large number of African populations (Tishkoff et al. 2009), although the most widespread ancestry component derived from STRUCTURE analysis extended beyond Bantu-speaking groups to include all Niger-Congo groups. Another factor to be considered is the recent time of this expansion suggested to be 3,000–5,000 ya (Blench 2006), which would reduce the accumulation of variability and structure among populations. The tMRCAs estimated for the sublineages E1b1a7a and E1b1a8 are in accordance with a recent expansion. We suggest that a more plausible scenario is one in which there was continuous backward and forward migration after an initially rapid spread as indicated by the significant amount of haplotype sharing between Western and Eastern Bantu-speaking groups (fig. 4 and supplementary fig. 5 and supplementary table 3, Supplementary Material online). Thus, our Y chromosome evidence suggests recent expansion and ongoing contacts over the large geographic area occupied by Bantu speakers. This is in good accordance with linguistic evidence showing that the Bantu languages as we know them today have been shaped over the last four millennia through successive stages of "punctuation" and "equilibrium" (Dixon 1997). Punctuational bursts of change at the time of language splitting can account for only 31% of the total divergence in the basic vocabulary of Bantu languages (Atkinson et al. 2008), whereas convergence effects due to multilingualism and intensive and protracted contacts between speech communities certainly played an equally important role in shaping the current Bantu language area (Schadeberg 2003). For instance, the emergence of a relatively homogenous group of so-called "Savannah Bantu" languages, sometimes seen as a Bantu

"subclade" (e.g., Ehret 2001), is most likely the result of intensive contact between languages originally belonging to distinct Eastern and Western Bantu branches (Möhlig 1981; Nurse and Philippson 2003; Bostoen and Grégoire 2007). Phenomena such as political centralization and economic integration involving communities separated over long distances is equally reflected in the archaeological record of several regions of Central, Eastern, and Southern Africa, certainly from the last millennium onward but even earlier (Fagan 1977; Denbow 1990; Chami 1999; De Maret 2005; Phillipson 2005).

Our conclusion contradicts the conclusion of Sikora et al. (2010), who suggest language shift in southeastern Bantu from Mozambique as an explanation for their distinctiveness from three other Bantu populations in the data set (the Luhya from Kenya as well as the Kenyan and South African Bantu groups included in our study). These discrepancies may be explained by the differences in the populations included (southeastern Bantu from Mozambique being unfortunately absent in our data set) or in the type of markers used because autosomal and Y chromosomal markers underlie different demographic trajectories. In summary, our interpretation of the spread of Bantu as a major migratory phenomenon provides a better explanation for the present-day distribution of the paternal lineages in Africa than the alternative scenario of cultural diffusion of the Bantu languages but need not necessarily hold true for the maternal lineages or autosomal markers.

Conclusions

The pattern of Y chromosomal variation in sub-Saharan Africa appears to be driven by the joint effect of both linguistic affiliation and geographical distribution, which to some extent are also correlated. These results were quantified by means of an AMOVA where the percentage of variance explained by differences between groups is larger for the grouping based on linguistic affiliation (~14%) than for that based on geographical criteria (~10%). This somewhat larger effect of language over geography was also found in other studies (Tishkoff et al. 2009 and Bryc et al. 2010). However, there is still a strong effect of geographical proximity (i.e., isolation by distance) on the patterns of Y chromosomal variation as demonstrated by the significant correlation observed between geographic and genetic distances calculated as F_{ST} or R_{ST} values (for haplogroups and STRs, respectively). When considering only Niger-Congo groups, the correlation between R_{ST} and geographic distances is no longer significant, probably because of the recent expansion of the language phylum.

The data presented here make it clear that there is considerable structure within haplogroup E1b1a in Africa. Analyzing the four sublineages E1b1a*(xE1b1a8), E1b1a8, E1b1a7 (xE1b1a7a), and E1b1a7a together with STRs allowed deeper insights into the Y chromosomal variation in this continent and one of the events that shaped it, namely the Bantu expansion. We suggest that the M2 mutation was present in the ancestors of the Niger-Congo populations at an early stage and was subsequently involved in the spread of the language phylum; furthermore, mainly the E1b1a subhaplogroups E1b1a7a and E1b1a8 are implicated in the Bantu expansion. However, some portions of Africa remain understudied; only when these lacunae have been filled will it be possible to come to more definitive insights into the prehistory of this area.

Supplementary Material

Supplementary figs S1—S5, supplementary text, and supplementary tables S1—S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to all the donors of the samples genotyped here; to Vicent Katanekwa, Dudu Musway, Joseph Koni Muluwa, Manuela Cioffi, Gianluca Frinchillucci, and Francesca Lipeti for invaluable assistance with sample collection; to Michael Cysouw, Michael Dannemann, Roger Mundry, and Marc Bauchet for assistance with the statistical analyses and R programming, as well as to Antje Müller for help with DNA extractions and genotyping. This study was supported by the Max Planck Society.

References

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet*. 75:338—345.

Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008. Languages evolve in punctuational bursts. *Science* 319:588.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol Biol Evol*. 26:1581—1589.

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

Bosch E, Calafell F, Santos FR, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J. 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet*. 65: 1623—1638.

Bostoen K. 2007. Pots, words and the Bantu problem: on lexical reconstruction and early African history. *J Afr Hist*. 48: 173—199.

Bostoen K, Grégoire C. 2007. 'La question bantoue: bilan et perspectives'. Mémoires de la Société de Linguistique de Paris (NS) 15, special issue: tradition et rupture dans les grammaires comparées de différentes familles de langues. Leuven (Belgium): Peeters p. 73—91.

Bostoen K. 2009. Shanjo and Fwe as part of Bantu Botatwe: a diachronic phonological approach. In: Ojo A, Moshi L, editors. Selected Proceedings of the 39th Annual Conference on African Linguistics. Sommerville (MA): Cascadilla Proceedings Project. p. 110—130.

Bryc K, Auton A, Nelson MR, et al. (11 co-authors). 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A*. 107:786–791.
Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet*. 9:403–433.
Cann HM, de Toma C, Cazes L, et al. (38 co-authors). 2002. A human genome diversity cell line panel. *Science* 296:261–262.
Chami FA. 1999. Roman beads from the Rufiji Delta, Tanzania: first incontrovertible archaeological link with the Periplus. *Curr Anthropol*. 40(2):237–241.
Cann HM, Sequeira F, de Toma C, Cazes L. (38 co-authors). 2002. A human genome diversity cell line panel. *Science*. 296:261–262.
Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol Biol*. 9:80.
Cruciani F, Santolamazza P, Shen P, et al. (16 co-authors). 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*. 70:1197–1214.
de Filippo C, Heyn P, Barham L, Stoneking M, Pakendorf B. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *Am J Phys Anthropol*. 141:382–394.
De Maret P. 2005. From pottery groups to ethnic groups in Central Africa. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 420–440.
Denbow JR. 1990. Congo to Kalahari: data and hypotheses about the political economy of the Western stream of the early iron age. *Afr Archaeol Rev*. 8:139–176.
Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, Spedini G, Capelli C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol*. 21:1673–1682.
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.
Dimmendaal GJ. 2008. Language ecology and linguistic diversity on the African continent. *Lang Linguist Compass*. 2:840–858.
Dixon RMW. 1997. The rise and fall of languages. Cambridge: Cambridge University Press.
Eggert M. 2005. The Bantu problem and African archaeology. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 301–326.
Ehret C. 2001. Bantu expansions: re-envisioning a central problem of early African history. *Int J Afr Hist Stud*. 34:5–27.
Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online*. 1:47–50.
Fagan BM. 1977. Early trade and raw materials in South Central Africa. In: Konczacki ZA, Konczacki JM, editors. An economic history of tropical Africa. Volume 1: the pre-colonial period. London: Frank Cass. p. 179–192.
Fortune G. 1970. The languages of the western province of Zambia. *J Lang Assoc East Afr*. 1:31–38.
Goldstein DB, Pollock DD. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered*. 88:335–342.
Gomes V, Sanchez-Diz P, Amorim A, Carracedo A, Gusmao L. 2010. Digging deeper into East African human Y chromosome lineages. *Hum Genet*. 127:603–613.
Gordon RG, Grimes BF. 2005. Ethnologue: languages of the world. Dallas (TX): SIL International. p. 1272.
Greenberg JH. 1948. The classification of African languages. *Am Anthropol*. 50:24–30.
Güldemann T, Vossen R. 2000. Khoisan. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 99–122.
Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet*. 4:e1000202.
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 U S A*. 105:10693–10698.
Holden CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc R Soc Lond B Biol Sci*. 269:793–799.
Holden CJ, Gray RD. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In: Forster P, Renfrew C, editors. Phylogenetic methods and the prehistory of languages. Cambridge: The MacDonald Institute for Archaeological Research. p. 19–31.
Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*. 4:598–612.
Kayser M, Brauer S, Cordaux R, et al. (12 co-authors). 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol*. 23:2234–2244.
Kayser M, Lao O, Saar K, Brauer S, Wang X, Nurnberg P, Trent RJ, Stoneking M. 2008. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet*. 82:194–198.
Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*. 18:830–838.
Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 16:1215.
Möhlig WJG. 1981. Stratification in the history of the Bantu languages. *Sprach Gesch Afr*. 3:251–316.
Nenadic O, Greenacre M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *J Stat Softw*. 20:1–13.
Neumann K. 2005. The romance of farming: plant cultivation and domestication in Africa. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 249–275.
Nichols J. 1997. Modeling ancient population structures and movement in linguistics. *Annu Rev Anthropol*. 26:359–384.
Nurse D, Philippson G. 2003. The Bantu languages. London and New York: Routledge. p. 708.
Pereira L, Gusmao L, Alves C, Amorim A, Prata MJ. 2002. Bantu and European Y-lineages in sub-Saharan Africa. *Ann Hum Genet*. 66:369–378.
Pebley A, Mbugua W, Goldman N. 1988. Polygyny and fertility in sub-Saharan Africa. *Fertil Determ Res Notes*. 21:6–10.
Phillipson D. 2005. African archaeology. Cambridge: Cambridge University Press.
Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. 2006. Evaluation of saliva as a source of human DNA for population and association studies. *Anal Biochem*. 353:272–277.
Quintana-Murci L, Quach H, Harmant C, et al. (23 co-authors). 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A*. 105:1596–1601.
Robertson JH, Bradley R. 2000. A new paradigm: the African early iron age without Bantu migrations. *Hist Afr*. 27:287–323.
Rosa A, Ornelas C, Jobling MA, Brehm A, Villems R. 2007. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol Biol*. 7:124.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet*. 70:841–847.

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

Sands B. 2009. Africa's linguistic diversity. *Lang Linguist Compass*. 3:559–580.

Schadeberg T. 2003. Historical linguistics. In: Nurse D, Philippson G, editors. The Bantu languages. London and New York: Routledge. p. 143–163.

Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. 2010. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet*. (online)

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB. 2000. Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba—the "Black Jews of Southern Africa". *Am J Hum Genet*. 66:674–686.

Tishkoff SA, Gonder MK, Henn BM, et al. (12 co-authors). 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol*. 24:2180–2195.

Tishkoff SA, Reed FA, Friedlaender FR, et al. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Underhill PA, Shen P, Lin AA, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet*. 26:358–361.

Vansina J. 1979. Bantu in the crystal ball .1. *Hist Afr*. 6:287–333.

Vansina J. 1995. New linguistic evidence and the Bantu expansion. *J Afr Hist*. 36:173–195.

Veeramah KR, Connell BA, Pour NA, Powell A, Plaster CA, Zeitlyn D, Mendell NR, Weale ME, Bradman N, Thomas MG. 2010. Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. *BMC Evol Biol*. 10:92.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. New York: Springer. p. 495.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*. 26:1823–1827.

Williamson K. 1989. Niger-Congo overview. In: Bendor-Samuel JT, Rhonda LH, editors. The Niger-Congo languages—a classification and description of Africa's largest language family. Lanham (MD): University Press of America. p. 3–45.

Williamson K, Blench R. 2000. Niger-Congo. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 11–42.

Wood ET, Stover DA, Ehret C, et al. (11 co-authors). 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet*. 13:867–876.

Zhivotovsky LA, Underhill PA, Cinnioglu C, et al. (17 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet*. 74:50–61.