Article-Discoveries

Along the Indian Ocean coast: genomic variation in Mozambique provides new insights into the Bantu expansion

Armando Semo[a], Magdalena Gayà-Vidal[a], Cesar Fortes-Lima[b], Bérénice Alard[a], Sandra Oliveira[c], João Almeida[a], António Prista[d], Albertino Damasceno[e], Anne-Maria Fehn[a,f], Carina Schlebusch[b,g,h], Jorge Rocha[a,i,j]

[a]CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos
Universidade do Porto
Campus de Vairão
Rua Padre Armando Quintas, nº 7
4485-661 Vairão, Portugal

[b]Department of Organismal Biology
Evolutionary Biology Centre
Uppsala University
Norbyvagen 18C
SE-752 36 Uppsala, Sweden

[c]Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
04103 Leipzig, Germany

[d]Faculdade de Educação Física e Desporto
Universidade Pedagógica de Moçambique
Avenida Eduardo Mondlane 955
257 Maputo, Mozambique

[e]Faculdade de Medicina
Universidade Eduardo Mondlane
Avenida Salvador Allende 702
257 Maputo, Mozambique

[f]Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History
Kahlaische Strasse 10
07745 Jena, Germany

[g]Palaeo-Research Institute
University of Johannesburg
P.O. Box 524
Auckland Park, 2006, South Africa

[h]SciLifeLab, Uppsala, Sweden

[i]Departamento de Biologia, Faculdade de Ciências
Universidade do Porto
Rua Campo Alegre s/n
4169-007 Porto, Portugal

[j]To whom correspondence should be addressed. Email: jrocha@cibio.up.pt

Corresponding author:
Jorge Rocha
CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto
Campus de Vairão/ Rua Padre Armando Quintas, nº 7/4485-661 Vairão/Portugal
Mail: jrocha@cibio.up.pt
Phone: +351 961031921
ORCID: 0000-0001-5460-7615

Abstract

The Bantu expansion, which started in West Central Africa around 5,000 BP, constitutes a major migratory movement involving the joint spread of peoples and languages across sub-Saharan Africa. Despite the rich linguistic and archaeological evidence available, the genetic relationships between different Bantu-speaking populations and the migratory routes they followed during various phases of the expansion remain poorly understood. Here, we analyze the genetic profiles of southwestern and southeastern Bantu-speaking peoples located at the edges of the Bantu expansion by generating genome-wide data for 200 individuals from 12 Mozambican and 3 Angolan populations using ~1.9 million autosomal single nucleotide polymorphisms. Incorporating a wide range of available genetic data, our analyses confirm previous results favoring a "late split" between West and East Bantu speakers, following a joint passage through the rainforest. In addition, we find that Bantu speakers from eastern Africa display genetic substructure, with Mozambican populations forming a gradient of relatedness along a North-South cline stretching from the coastal border between Kenya and Tanzania to South Africa. This gradient is further associated with a southward increase in genetic homogeneity, and involved minimum admixture with resident populations. Together, our results provide the first genetic evidence in support of a rapid North-South dispersal of Bantu peoples along the Indian Ocean Coast, as inferred from the distribution and antiquity of Early Iron Age assemblages associated with the Kwale archaeological tradition.

Keywords: Mozambique, Bantu expansion, population structure, migration, admixture

Introduction

It is generally believed that the dispersal of Bantu languages over a vast geographical area of sub-Saharan Africa is the result of a migratory wave that started in the Nigeria-Cameroon borderlands around 4,000-5,000 BP (Rocha and Fehn 2016; Bostoen 2018; Schlebusch and Jakobsson 2018). Although the earliest stages of the Bantu expansions were probably not associated with plant cultivation and domestication, Bantu speech communities added agriculture and iron metallurgy to their original subsistence strategies and subsequently replaced or assimilated most of the resident forager populations who lived across sub-Saharan Africa (Mitchell and Lane 2013; Bostoen et al. 2015). For this reason, the dispersal of Bantu-speaking peoples has often been considered a prime example of the role of food production in promoting demic migrations and language spread (Diamond and Bellwood 2003).

While genetic studies had a pivotal role in demonstrating that the Bantu expansions involved a movement of people (demic diffusion) rather than a mere spread of cultural traits (Tishkoff et al. 2009; de Filippo et al. 2012; Schlebusch et al. 2012; Li et al. 2014), the majority of research on the specific routes and detailed dynamics of the spread of Bantu-speakers has been conducted in the fields of linguistics and archaeology.

Linguistic studies focusing on the reconstruction of the historical relationships between modern Bantu languages have led to some rather concrete proposals about links between individual languages and language areas, including the establishment of three widely accepted geographical subgroups: North-West Bantu, East Bantu and West Bantu (Guthrie 1948; Vansina 1995; Bostoen 2018). Among them, the East Bantu languages, which currently extend from Uganda to South Africa, have been shown to form a single monophyletic clade that is believed to be a relatively late offshoot of West Bantu (Holden 2002; Currie et al. 2013; Grollemund et al. 2015). Assuming that the phylogenetic trees inferred from the comparison of lexical data can be used to trace the migratory routes of ancestral Bantu-speaking communities, the linguistic pattern favors a dispersal scenario whereby populations from the Nigeria-Cameroon homeland first migrated to the south of the rainforest and later diversified into several branches before occupying eastern and southern Africa (Currie et al. 2013; Grollemund et al. 2015).

According to archaeological evidence, the earliest Bantu speakers in East Africa appeared around 2,600 BP in the Great Lakes region, associated with pottery belonging to the so-called Urewe tradition, also characterized by a distinctive iron smelting technology and farming (Phillipson 2005; Bostoen 2018). However, the link between Urewe and pottery traditions further west is unclear, and the historical events leading to its introduction to the inter-lacustrine area are still poorly understood (Bostoen 2007). Some interpretations of the archaeological data have proposed that, in contrast with the “late split” between East and West Bantu suggested by linguistic evidence, East Bantu peoples introduced the Urewe

tradition into the Great Lakes by migrating out of the proto-Bantu heartland along the northern fringes of the rainforest after an early separation from Bantu speakers occupying the western half of Africa (Phillipson 1977; Huffman 2007). This model, however, is not supported by recent genetic studies showing that Bantu-speaking populations from eastern and southern Africa are more closely related to West Bantu speakers that migrated to the south of the rainforest than they are to West Bantu speakers that remained in the north (Busby et al. 2016; Patin et al. 2017; Schlebusch et al. 2017).

In spite of their uncertain origins, the Urewe assemblages display pottery styles similar to the younger Kwale and Matola traditions that are distributed along coastal areas ranging from southern Kenya across Mozambique to KwaZulu-Natal (Sinclair et al. 1993; Phillipson 2005; Bostoen 2007; Bostoen 2018). This archaeological continuity has been interpreted as the earliest material evidence for an extremely rapid dispersion of East Bantu speakers from the Great Lakes, starting around the second century AD and reaching South Africa in less than two centuries (Sinclair et al. 1993; Phillipson 2005; Bostoen 2018). Such a migration remains, however, to be documented by genetic data, due to insufficient sampling of the areas lying between eastern and southern Africa that roughly correspond to present-day Mozambique.

In this study, we fill this important gap by investigating the population history of Mozambique using ~1.9 million quality-filtered single nucleotide polymorphisms (SNPs) that were genotyped in 161 individuals from 12 populations representing all major Mozambican languages, and in 39 individuals from 3 contextual populations from Angola (fig. 1 and supplementary table 1). By making use of a maximally wide range of available genetic and linguistic data, we show that East Bantu-speaking populations display genetic substructure, and detect a strong signal for the dispersal of East Bantu peoples along a North-South cline, which possibly started in the coastal border between Kenya and Tanzania and involved minimum admixture with local foragers until the Bantu-speakers reached South Africa. Together, our results provide a strong support for reconstructions of the eastern Bantu migrations based on the distribution of Kwale archaeological sites.

RESULTS AND DISCUSSION

Genetic variation in Mozambique

To assess the genetic relationships between Angolan and Mozambican individuals, we performed principal component analysis (PCA) (Patterson et al. 2006) and unsupervised clustering analysis using ADMIXTURE (Alexander et al. 2009) (fig. 1).

The PCA patterns are closely related to geography, with the first PC (PC1) separating Mozambican and Angolan individuals, and the second PC (PC2) revealing a noticeable heterogeneity among samples from Mozambique (fig. 1B and supplementary fig. 1; Procrustes correlation: 0.89; $P<0.001$). The ADMIXTURE analysis confirmed the substantial differentiation between populations from Angola and Mozambique (at K=2), and the genetic substructure among Mozambican populations (at K=3) (fig. 1C).

Within Mozambique, the association between genetic patterns and geography is further highlighted by a strong correlation between average PC2 scores and latitude ($r=-0.97$, $P<10^{-6}$), showing that genetic variation is structured along a North-South cline corresponding to the orientation of the country's major axis (fig. 2A). The highest genetic divergence was found between Yao and Mwani speakers in the north, and Tswa-Ronga (Tswa, Changana, Ronga) and Inhambane (Bitonga and Chopi) speakers in the south, while Makhuwa, Sena, Nyanja and Shona (Manyika and Ndau) speakers occupy intermediate genetic and geographic positions (figs. 1B and 2A). Qualitatively, this trend is consistent with the geographic distribution of subclusters of Mozambican languages in the Bantu phylogeny proposed by Grollemund et al. (Grollemund et al. 2015) (cf. their supplementary fig. 1). Our own lexicostatistical analyses (supplementary fig. 2; supplementary tables 3-5), reveal significant correlations between genetic and linguistic pairwise distances (Mantel test: $r=0.68$; $P=2.9 \times 10^{-5}$), as well as between genetic and latitudinal distances ($r=0.61$; $P=7 \times 10^{-4}$), and linguistic and latitudinal distances ($r=0.79$; $P=6.3 \times 10^{-5}$). In contrast, correlations with longitude, involving either language or genetics, were not significant, further emphasizing the importance of latitude in structuring genetic and linguistic diversity in Mozambique (supplementary table 3). We also performed partial Mantel tests to evaluate the respective effect of language and geography on genetic variation. We found that while genetic and linguistic distances remained correlated when latitude was kept constant, genetic and latitudinal distances were not significantly correlated when holding language constant (supplementary table 3). The latter result indicates that language is a more important predictor of genetic differentiation than geography, as populations speaking similar languages tend to be genetically closer than expected on the basis of their location along the latitudinal axis. Since it has been recently shown that the relationships between Bantu languages can be represented by robust phylogenetic trees reflecting the fission history of Bantu-speaking groups (Currie et al. 2013; Grollemund et al. 2015), the correlation results can be interpreted

as an indication that the spatial patterns of genetic and linguistic variation in Mozambique are the outcome of successive population splits during a North-South range expansion, rather than a consequence of geographically structured gene flow underlying isolation by distance (cf. Smouse et al. 1986; Sokal et al. 1988; Smouse and Long 1992).

A stepwise reduction in levels of genetic diversity with increasing geographic distance from a reference location is generally considered to be the typical outcome of a demic migration involving serial bottlenecks (Ramachandran et al. 2005). In the global context of the Bantu expansion, a significant decrease of genetic diversity with distance to the Bantu homeland was previously reported for mitochondrial DNA and the Y-chromosome, but not for autosomes (de Filippo et al. 2012). Moreover, to our knowledge, there have been no reports for such patterns at more local scales. In order to evaluate the relationship between genetic diversity and geography, we studied the distribution of haplotype heterozygosity (HH), numbers and total lengths of runs of homozygosity (RoHs) and linkage disequilibrium (LD), as measured by the squared correlation of allele frequencies (r^2), across all sampled Mozambican populations (Supplementary Material).

We found that the number of RoHs and LD were significantly correlated with latitude, with northern populations displaying higher genetic diversity than southern populations (figs. 2B and C; supplementary figs. 3-5). We also observed a decrease of HH with absolute latitude that did not reach significance (supplementary fig. 3A; r=0.51, P=0.104). However, HH was still significantly correlated with LD (supplementary fig. 4C). Together, these results suggest that East Bantu-speaking peoples entered Mozambique from the North and underwent sequential reductions in effective population size, leading to increased genetic homogeneity and differentiation as they moved southwards.

To further assess the relationship between population structure and geography in Mozambique, we used the Estimated Effective Migration Surfaces (EEMS) method, which identifies local zones with increased or decreased migration rates, relative to the global migration across the whole country (Petkova et al. 2015) (fig. 3A). We detected two zones of low migration between northern and central Mozambique (fig. 3A): one associated with Yao speakers, located in the northwestern highlands of the Nyasa Province between lake Nyasa/ Malawi and the Lugenda River (figs. 3B and C); the other, located in the Northeast, to the north of the Ligonha River, around Makhuwa-speaking areas (figs. 3A and B). An additional low-migration zone was found around the Save River, between southern and south-central Mozambique (figs. 3A and B). Interestingly, the EEMS analysis also shows that the Zambezi River in central Mozambique is not an obstacle but rather a corridor for migration (figs. 3A and B). This is in line with archaeological findings supporting the importance of the Zambezi Basin in long-distance trading networks between the Indian Ocean Coast and the southern African hinterland from the mid-1st millennium onwards (Chirikure 2014, Nikis and Smith 2017). Overall, the geographic patterns revealed

by the EEMS method are consistent with the PC cline in showing that the highest genetic differentiation between the northernmost and southernmost populations is reinforced by intervening low migration zones, while the relative genetic proximity between central Mozambican groups was enhanced by increased migration around the Zambezi Basin (figs. 3A and B).

Genetic relationships with other African populations

To place the genetic variation of Mozambican and Angolan samples into the wider context of the Bantu expansion, we combined our dataset with available genome-wide comparative data from other African populations (fig. 4A and supplementary table 6).

Genetic clustering analysis shows that three partially overlapping components can be roughly associated with major geographic areas and linguistic subdivisions of the Niger-Congo phylum, of which the Bantu languages form part (fig. 4D and supplementary fig. 6): non-Bantu Niger-Congo in West Africa, to the north of the rainforest (beige); West Bantu, including Angolans, along the Atlantic coast (green); and East Bantu, including Mozambicans, in East Africa and along the Indian Ocean Coast (blue). A pairwise F_{st} analysis measuring the genetic divergence among Niger-Congo speaking populations further shows that the highest levels of differentiation (F_{st}=0.01) are found between non-Bantu Niger-Congo groups and East Bantu-speaking peoples (supplementary fig. 7).

Other major genetic components revealed by clustering analysis are associated with Kx'a, Tuu and Khoe-Kwadi-speaking peoples from southern Africa, also known as Khoisan (brown), Rainforest Hunter-Gatherers (RHG) (violet and light green), non-Bantu Eastern Africans (black) and Europeans (pink). As found in previous works (Pickrell et al. 2012; Schlebusch et al. 2012; Patin et al. 2017), several Bantu-speaking populations have varying proportions of these genetic components, which were likely acquired through admixture with local residents: 11% (range: 4-21%) of RHG-related component in West Bantu speakers; 16% (range: 9-38%) of non-Bantu eastern African-related component in East Bantu speakers from Kenya and Tanzania; and 17% (range: 16-18%) of Khoisan-related component in southeastern Bantu speakers from South Africa.

To mitigate the effect of admixture with resident populations, we carried out a PC analysis of all Bantu-speaking groups, together with one representative group of non-Bantu Eastern Africans (Amhara) and one representative group of southern African Khoisan (Ju|'hoansi), which are the two most important sources for external admixture with Bantu-speaking populations from the East and South, respectively. As expected, the first two principal axes are driven by genetic differentiation between the Amhara (PC1) and the Ju|'hoansi (PC2), relative to Bantu-speaking groups (supplementary fig. 8A). Moreover, some Bantu peoples from eastern (e.g., Kikuyu and Luhya) and southern Africa (e.g., Sotho and Zulu) stand

out from a tight cluster encompassing all Bantu speakers by extending toward the Amhara and Ju|'hoansi, respectively, indicating admixture of local components into the genomes of Bantu-speaking populations. When considering PCs that explain less variance, a close link between the internal differentiation of Bantu-speaking groups and geography becomes apparent (fig. 4B and supplementary fig. 8E; Procrustes correlation: 0.76; P< 0.001). PC3 represents an east-west axis displaying a noticeable gap between West and East Bantu speakers, and PC4 highlights the differentiation of Mozambican and South African groups from eastern African populations located to their north. As shown in fig. 4C, the heterogeneity of East Bantu populations is further emphasized when West Bantu speakers are removed (Procrustes correlation: 0.44; P< 0.001; supplementary fig. 8F). While PC4 is correlated with longitude ($r=-0.73$; $P<10^{-4}$), PC3 is highly correlated with latitude ($r=0.95$; $P<10^{-13}$), showing that the gradient of genetic differentiation previously observed within Mozambique extends from eastern to southern Africa (fig. 2A and supplementary fig. 9). Heuristically, the genetic differentiation among East Bantu speakers can be described by defining four groups that are broadly associated with different geographic regions in eastern and southeastern Africa, and partially correspond to various linguistic zones of Guthrie's Bantu classification (Guthrie 1948; Maho 2003) (fig. 4C, supplementary figs. 9 and 10): 1) the first group includes peoples from the western fringe of eastern Africa (Kikuyu, Luhya, Baganda, Barundi and Kinyarwanda), who live around Lake Nyanza/Victoria and mostly speak languages belonging to Bantu zone J (Lakes Bantu) (Bastin et al. 1999); 2) the second group includes populations from coastal Kenya (Chonyi, Giriama, Kambe and Kauma), who belong to the Mijikenda ethnic group and speak languages from zone E; 3) the third group is genetically intermediate between groups 1 and 2, and includes the Mzigua, Wabondei and Wasambaa from Tanzania, who speak languages from zone G; 4) the fourth group, formed by Mozambicans and South Africans**,** is an heterogeneous set of populations covering linguistic zones N, P and S, who bridge the area between eastern and southern Africa and are genetically closer to groups from Tanzania than to other East Africans.

These findings have important implications for integrating archaeological, linguistic and genetic data in the reconstruction of the Bantu migrations in the easternmost regions of Africa. Although many crucial areas like Democratic Republic of Congo, Zambia and Zimbabwe still need to be included in genome-wide analyses, the available data suggests that the occupation of eastern Africa by Bantu-speaking populations was associated with genetic structuring in the relatively small area between the Great Lakes and the Indian Ocean Coast, with Tanzanian groups being to the ancestors of south-eastern Bantu-speaking populations. This scenario agrees with the migratory path inferred from the continuity between Early Iron Age (EIA) archaeological sites from the Kwale ceramic tradition, which extend from coastal Kenya and Tanzania to South Africa across a Mozambican corridor (Sinclair et al. 1993; Phillipson 2005; Bostoen 2018).

To further investigate the origins of the migratory streams linking different Bantu-speaking groups and to better characterize the admixture dynamics between Bantu speakers and resident populations, we applied the haplotype-based approaches implemented in CHROMOPAINTER and GLOBETROTTER (Lawson et al. 2012; Hellenthal et al. 2014). We found that the haplotype copy profiles of Angolans differ significantly from Mozambicans+South Africans (figs. 5A and B): while the former derive most of their haplotypes from West Bantu-speaking populations located to their North, the latter trace most of their ancestry to Bantu-speaking groups from East Africa, in close agreement with the PCA results (fig. 4). More specifically, we found that the best donor population proxy (Mzigua) for Bantu speakers from Mozambique and South Africa is located in Tanzania (range: 72-93%), whilst Angolans derive most of their ancestry from Bantu-speaking groups in Gabon and Cameroon (range: 77-83%) (fig 5C; supplementary table 7).

Estimated Khoisan ancestry in the South African Sotho (24%) and Zulu (24%) is much higher than in their close Mozambican neighbors Ronga (5%) and Changana (4%), or in any other Mozambican group (range: 1-5%) (figs. 5B and C; supplementary table 7). This pattern suggests that Bantu speakers scarcely admixed with local foragers, in agreement with recent findings about Bantu speakers from Malawi, who displayed no Khoisan ancestry, despite the confirmed presence of a Khoisan-related genetic component in ancient samples from the region (Skoglund et al. 2017). It therefore seems that the processes governing earlier admixture events between Bantu-speakers and local hunter-gather groups in modern-day Mozambique and Malawi were very different from what has been reported for South Africa and Botswana (Pickrell et al. 2012; Schlebusch et al. 2012; González-Santos et al. 2015). As previously suggested on the basis of genetic variation in uniparental markers and archaeological modeling, the differences in admixture dynamics leading to increased Bantu/Khoisan admixture beyond the southern border of Mozambique could have been caused by a slowdown of the Bantu expansion due to adverse ecoclimatic conditions (Marks et al. 2015). In addition, the better conditions found in Mozambique and Malawi may have favored the rapid population growth of Bantu-speaking migrants, resulting in a demographic imbalance between residents and incomers and leading to low levels of Khoisan admixture, even in the event of total assimilation.

To evaluate the effect of Khoisan ancestry on the pattern of southward increase of genetic homogeneity detected in Mozambique (fig. 2; supplementary fig. 3), we reassessed the correlations between genetic diversity and latitude after masking Khoisan segments in Mozambican groups (Supplementary Material). Although the masking procedure led to a decrease in power due reduction of the number of available SNPs (950,000 vs 500,000), we still found a strong signal of southward increase in the number of RoHs, after removal of Khoisan ancestry (supplementary figure S12B). These results favor the hypothesis that the decreasing levels of genetic diversity in Mozambique are associated with a range

expansion with serial founder effects, confirming that the effect remains after masking admixed fragments.

A recent genome-wide study found that the best-matching source population for South African Bantu speakers is located in Angola (Kimbundu) rather than in East Africa (as represented by the Bakiga and Luhya from around the Great Lakes) (Patin et al. 2017). Here, we used a stepwise approach to rank the best proxies for the ancestry of two South African Bantu-speaking groups (Sotho+Zulu) among all populations contained in our dataset (fig. 6; Supplementary Material; supplementary table 7). We found that the Changana and Ronga from Mozambique, and a southern Khoisan descendent group (the Karretjie People of South Africa) are the best proxies for the ancestry of the South African Bantu speakers (fig. 6A). When Mozambican populations are removed from the list of sources, the next best non-Khoisan proxies are the Mzigua from Tanzania (fig. 6B). The contribution of Angola only becomes increasingly more relevant when Tanzanian (fig. 6C), Kenyan (fig. 6D) and Great Lakes (fig. 6E) populations are successively removed from the list of donors. Nevertheless, the fact that Angola still represents a better proxy for the ancestry of southeastern Bantu speakers than populations closer to the Bantu homeland provides additional evidence in favor of a "late-split" between southwestern and southeastern Bantu-speaking groups after a single passage through the rainforest, as suggested in previous studies (Busby et al. 2016; Patin et al. 2017).

In a further step, we identified and dated signals of admixture in the history of the studied populations using GLOBETROTTER. We found no evidence for admixture between any two Mozambican populations (not shown), suggesting that the intermediate position of central Mozambique in the North-South gradient of genetic relatedness (figs. 1B and 2A) is not the result of admixture between populations from northern and southern Mozambique but rather a cline of stepwise genetic differentiation. At the same time, we found that the Khoisan ancestry detected in South Africans (Sotho and Zulu) and at low frequencies in southern Mozambican populations (Ronga, Changana, Tswa, Bitonga and Chopi) (fig. 5C) resulted from admixture events occurring around 1165 BP (range: 756-1851 BP), involving the Karretjie people from South Africa as best matching Khoisan source and the Tanzanian Mzigua as best-matching Bantu-speaking population (P-values for evidence of admixture <0.05; supplementary table 8). This date is remarkably consistent with the first Iron Age arrivals to southern Mozambique associated with the Matola pottery, which stylistically resembles the Kwale ceramics from Tanzania and has been dated to the early and mid-first millennium AD (Sinclair et al. 1993).

We also found evidence (P<0.05) for admixture with Afro-Asiatic (Amhara and Oromo) and Nilotic (Kalenjin and Maasai) speakers in Bantu-speaking groups from the Great Lakes, coastal Kenya and Tanzania (supplementary table 8). The average estimated antiquity of these admixture events dates to

~760 BP (570-1047 BP) and is in close agreement with Bantu/non-Bantu eastern African admixture dates inferred by Skoglund et al. (Skoglund et al. 2017). These estimates postdate the Bantu/Khoisan admixture inferred for Mozambique and South Africa, suggesting that the bulk of admixture between Bantu and non-Bantu speakers in East Africa occurred only after Bantu speakers had already begun their migration towards the South. This is also supported by the low eastern African ancestry detected in Bantu speakers from Mozambique and South Africa.

Conclusion

Using a country-wide sample of 12 Mozambican populations, we were able to fill an important gap in the understanding of the expansion of Bantu speakers from the Great Lakes region to the eastern half of southern Africa. Our results suggest that, in spite of the present-day homogeneity of East Bantu languages, the arrival of Bantu-speaking groups in eastern Africa was associated with a period of genetic differentiation in the area between the Great Lakes and the Indian Ocean Coast, followed by a southwards dispersal out-of Tanzania, along a latitudinal axis spanning cross Mozambique into South Africa. The resulting gradient of genetic relatedness is accompanied by a gradual reduction in genetic diversity possibly indicative of serial bottlenecks, as well as by a progressive loss of the genetic similarity between East Bantu speakers and Bantu-speaking peoples remaining in West-Central Africa. This increased genetic differentiation, however, cannot be attributed to admixture with resident populations. In fact, the absence of a substantial Khoisan contribution to the genetic make-up of Mozambican Bantu speakers (1-5%) suggests that the migrants had very low levels of admixture with resident populations until they reached the southernmost areas of eastern Africa, where Sotho and Zulu display considerable admixture proportions (24%). Moreover, the dates we obtained for admixture between Bantu speakers and Khoisan groups (~1165 BP) are remarkably close to the dates for the first archaeological attestations of the presence of Bantu speakers in southeastern Africa. We therefore conclude that our results provide a genetic counterpart to the distribution of Early Iron Age assemblages associated with the Kwale ceramic tradition, which are thought to constitute the material evidence for the southward movement of Bantu speech communities along the Indian Ocean coast.

Material and Methods

Population samples. A total of 221 samples from 12 ethnolinguistic groups from Mozambique and three groups from Angola were included in the present study (fig. 1A). Sampling procedures in Mozambique and Angola were described elsewhere (Alves et al. 2011; Oliveira et al. 2018). All samples were collected with informed consent from healthy adult donors, in collaboration with the Portuguese-Angolan TwinLab established between CIBIO/InBIO and ISCED/Huíla Angola and the Pedagogic and Eduardo Mondlane Universities of Mozambique. Ethical clearances and permissions were granted by CIBIO/InBIO-University of Porto, ISCED, the Provincial Government of Namibe (Angola), and the Mozambican National Committee for Bioethics in Health (CNBS).

Genotyping and phasing. DNA samples were extracted from buccal swabs and genotyped with the Illumina Infinium Omni2-5Exome-8 v1-3_A1 BeadChip (Gunderson et al. 2005; Steemers et al. 2006), after Whole Genome Amplification (WGA). Of a total of 2,612,357 genomic variants initially typed in 221 samples from Angola and Mozambique, a final set of 200 individuals typed for 1,946,715 autosomal SNPs was retained after applying quality control filters. Haplotypes and missing genotypes were inferred using SHAPEIT2 (Delaneau et al. 2013). Geographic locations, linguistic affiliations and sample sizes for all groups are presented in supplementary table 1. Details about DNA extraction, genotyping, haplotyping and quality control filtering are provided in Supplementary Material.

Data merging. The newly generated data from Angola and Mozambique were merged with eight publicly available datasets (Li et al. 2008; Henn et al. 2011; Schlebusch et al. 2012; 1000 Genomes Project Consortium et al. 2015; Gurdasani et al. 2015; Busby et al. 2016; Montinaro et al. 2017; Patin et al. 2017), following the approach described in Supplementary Material. The final merged dataset consists of 1,466 individuals from 89 populations typed for 105,286 SNPs (supplementary table 6).

Genetic data analysis. PCA was performed with the EIGENSOFT v7.2.1 package (Patterson et al. 2006). Unsupervised clustering analysis was done with ADMIXTURE (Alexander et al. 2009) applying a cross-validation (CV) procedure. We performed 20 independent runs for each number of clusters (K) and post-processed and plotted the results with the pong software (Behr et al. 2016). For PC and ADMIXTURE analyses, SNPs in linkage disequilibrium (r^2>0.5) were removed using PLINK 1.9 (Chang et al. 2015), which reduced the newly-generated and merged datasets to 927,435 and 98,570 independent autosomal SNPs, respectively. To assess the relationship between genetic, geographic and linguistic data, we used Procrustes analysis (Wang et al. 2010), Estimated Effective Migration Surfaces (EEMS) (Petkova et al. 2015) and Mantel tests (Mantel 1967), as detailed in Supplementary Material. Levels of genetic diversity were assessed by using Haplotype Heterozygosity (HH), Runs of Homozygosity (RoH) and Linkage Disequilibrium (LD), as described in Supplementary Material. All

reported correlations were assessed using Pearson correlation coefficient (r). To infer "painting" or copying profiles and quantify the ancestry contributions of different African groups to Bantu-speaking populations of Mozambique, Angola and South Africa, we used CHROMOPAINTER v.2 (Lawson et al. 2012) in combination with the MIXTURE MODEL regression implemented in the GLOBETROTTER software (Hellenthal et al. 2014). GLOBETROTTER was also used to infer and date admixture events. Details on the application of these methods are provided in Supplementary Material.

Linguistic data analysis. We collected published lexical data from 24 languages from Mozambique (10), Angola (3), eastern (9) and southern Africa (2) (supplementary figs. 2 and 10), based on the wordlist published by Grollemund et al. (Grollemund et al. 2015) consisting of 100 meanings (supplementary table 4). Using reconstructions provided in the online database Bantu lexical reconstructions 3 (Bastin et al. 2002) in combination with standard methodology from historical-comparative linguistics, we identified 636 cognate sets, and all languages were coded for presence (1) or absence (0) of a particular lexical root. Based on our coded dataset (supplementary table 5), we used the software SplitsTree v4.14.2 (Huson and Bryant 2006) to generate a matrix of pairwise linguistic distances (1-the percentage of cognate sharing) and computed Neighbor-Joining networks with 10,000 Bootstrap replicates (supplementary figs. 2A and 10A). We further applied to our coded dataset a Bayesian phylogenetic approach as implemented in the BEAST2 software (Bouckaert et al. 2014), using the Continuous Time Markov Chain (CTMC) model (Greenhill and Gray 2009) included in the Babel package (Bouckaert 2016). We assumed 10,000,000 generations and sampled every 1,000th generation. The first 1,000 generations were discarded as burn-in. The resulting consensus tree was converted in a radial tree using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) (supplementary figs. 2B and 10B).

Data Availability

The newly generated data will be made available for academic research use through the ArrayExpress database (accession number TBD).

Acknowledgments

We are grateful to all subjects who participated in this research. Financial support for this work was provided by Foundation for Science and Technology (FCT, Portugal) under the project PTDC/BIA-GEN/29273/2017 and by projects Variabilidade Biológica Humana em Moçambique and STEPS at the Pedagogic University and Eduardo Mondlane University of Mozambique. Genotyping was performed by the SNP&SEQ Technology Platform in Uppsala (www.genotyping.se). The facility is part of the National Genomics Infrastructure supported by the Swedish Research Council for Infrastructures and Science for Life Laboratory, Sweden. The SNP&SEQ Technology Platform is also supported by the Knut and Alice Wallenberg Foundation. The computations were performed at the Swedish National Infrastructure for Computing (SNIC-UPPMAX). AS was supported by the FCT grant SFRH/BD/114424/2016, MGV by POCI-01-0145-FEDER-006821 funded through the Operational Programme for Competitiveness Factors (COMPETE, EU) and UID/BIA/50027/2013 from FCT, AMF by CEECIND/02765/2017 from FCT, BA by a post-doctoral fellowship of the Fyssen foundation, and CS and CFL by the European Research Council (ERC - no. 759933). We would like to thank Ezekia Mtetwa, Per Sjödin and Jeffrey Wills for useful discussions, and two anonymous reviewers for valuable suggestions.

Author contributions

JR, CS, AP, and AD designed research; AP and AD performed research; AS, JR, MGV, CFL, BA, SO, JA, and AMF analyzed data; JR, CS, AS, CFL, and AMF wrote the paper.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19:1655–1664.

Alves I, Coelho M, Gignoux C, Damasceno A, Prista A, Rocha J. 2011. Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations. *Hum Biol*. 83:13–38.

Bastin Y, Coupez A, Mann M. 1999. Continuity and divergence in the Bantu languages: perspectives from a lexicostatic study. Tervuren: Musée Royal de l'Afrique Centrale.

Bastin Y, Coupez A, Mumba E, Schadeberg TC. 2002. Bantu lexical reconstructions 3 / Reconstructions lexicales bantoues 3. Available from: http://linguistics.africamuseum.be/BLR3.html.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. Pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 32:2817–2823.

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

Bostoen K. 2018. The Bantu expansion. In: Oxford Research Encyclopedia of African History. Oxford: Oxford University Press. Available from: http://africanhistory.oxfordre.com/view/10.1093/acrefore/9780190277734.001.0001/acrefore-9780190277734-e-191.

Bostoen K, Clist B, Doumenge C, Grollemund R, Hombert J-M, Muluwa JK, Maley J. 2015. Middle to Late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of western Central Africa. *Curr Anthropol*. 56:354–384.

Bouckaert R. 2016. Babel. BEAST analysis backing effective linguistics. Available from: https://github.com/rbouckaert/Babel.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 10: e 1003537.

Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM, et al. 2016. Admixture into and within sub-Saharan Africa. Elife. 5:e15266.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 4:7.

Chirikure S. 2014. Land and sea links: 1500 years of connectivity between southern Africa and the Indian Ocean rim regions, AD 700 to 1700. *Afr Archaeol Rev*. 31: 705-724.

Currie TE, Meade A, Guillon M, Mace R. 2013. Cultural phylogeography of the Bantu languages of sub-Saharan Africa. *Proc R Soc B*. 280:20130695.

Delaneau O, Zagury J-F, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 10:5–6.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science. 300:597–603.

de Filippo C, Bostoen K, Stoneking M, Pakendorf B. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc R Soc B*. 279:3256–3263.

González-Santos MG, Montinaro F, Oosthuizen O, Oosthuizen E, Busby GBJ, Anagnostou P, Destro-Bisol G, Pascali V, Capelli C. 2015. Genome-wide SNP analysis of southern African populations provides new insights into the dispersal of Bantu-speaking groups. *Genome Biol Evol*. 7:2560–2568.

Greenhill SJ, Gray RD. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In: Adelaar KA, Pawley A, editors. Austronesian historical linguistics and culture history: a Festschrift for Robert Blust. Canberra: Pacific Linguistics. p. 375–397.

Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proc Natl Acad Sci USA*. 112:13296–13301.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*. 37:549–554.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature. 517:327–332.

Guthrie M. 1948. The classification of the Bantu languages. London: Oxford University Press for the International African Institute.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science. 343:747–751.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA*. 108:5154-5162.

Hijmans RJ, van Etten J. 2011. raster: geographic analysis and modeling with raster data. R Packag. version 2.5-2.

Holden CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc R Soc B*. 269:793–799.

Huffman TN. 2007. Handbook to the Iron Age: the archaeology of pre-colonial farming societies in southern Africa. Scottsville: University of KwaZulu-Natal Press.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 23:254–267.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet*. 8:e1002453.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 319:1100–1104.

Li S, Schlebusch C, Jakobsson M. 2014. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc R Soc B*. 281:20141448.

Maho J. 2003. A classification of the Bantu languages: an update of Guthrie's referential system. In: Nurse D, Philippson G, editors. The Bantu Languages. London: Routledge. p. 639–651.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res*. 27:209–220.

Marks SJ, Montinaro F, Levy H, Brisighelli F, Ferri G, Bertoncini S, Batini C, Busby GBJ, Arthur C, Mitchell P, et al. 2015. Static and moving frontiers: the genetic landscape of southern African Bantu-speaking populations. *Mol Biol Evol*. 32:29–43.

Mitchell P, Lane P eds. 2013. The Oxford Handbook of African Archaeology. Oxford: Oxford University Press.

Montinaro F, Busby GBJ, Gonzalez-Santos M, Oosthuitzen O, Oosthuitzen E, Anagnostou P, Destro-Bisol G, Pascali VL, Capelli C. 2017. Complex ancient genetic structure and cultural transitions in southern African populations. Genetics. 205:303–316.

Nikis N, Smith, AL. 2017. Copper, trade and polities: exchange networks in southern Central Africa in the 2nd Millennium CE. *J South Afr Stud*. 43: 895-911.

Oliveira S, Fehn AM, Aço T, Lages F, Gayà-Vidal M, Pakendorf B, Stoneking M, Rocha J. 2018. Matriclans shape populations: insights from the Angolan Namib Desert into the maternal genetic history of southern Africa. *Am J Phys Anthropol*. 165:518–535.

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science. 356:543–546.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190. Petkova D, Novembre J, Stephens M. 2015. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet*. 48:94–100.

Phillipson DW. 1977. The later prehistory of eastern and southern Africa. London: Heinemann.

Phillipson DW. 2005. African Archaeology (3rd edition). Cambridge: Cambridge University Press.

Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, Kure B, Mpoloka SW, Nakagawa H, Naumann C, et al. 2012. The genetic prehistory of southern Africa. *Nat Commun*. 3:1143.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA*. 102:15942–15947.

Rocha J, Fehn A-M. 2016. Genetics and demographic history of the Bantu. In: eLS. John Wiley & Sons, Ltd: Chichester.

Schlebusch CM, Jakobsson M. 2018. Tales of human migration, admixture, and selection in Africa. *Annu Rev Genomics Hum Genet*. 19:405–428.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H, et al. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 358:652–655.

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MGB, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 338:374–379.

Sinclair Paul, Morais J, Adamowicz L, Duarte R. 1993. A perspective on archaeological research in Mozambique. In: Shaw T, Sinclair P, Andah B, Okpoko A, editors. The archaeology of Africa: food, metals and towns. London: Routledge. p. 409–431.

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al. 2017. Reconstructing prehistoric African population structure. Cell. 171:59-71.

Smouse P, Long JC. 1992. Matrix correlation analysis in anthropology and genetics. *Yearb Phys Anthropol*. 35:187-213.

Smouse P, Long JC, Sokal R. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool*. 35:627-632.

Sokal R. R. 1988. Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci USA*. 85: 1722-1726.

Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. 2006. Whole-genome genotyping with the single-base extension assay. *Nat Methods*. 3:31–33.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. Science. 324:1035–1044.

Vansina J. 1995. New linguistic evidence and 'The Bantu Expansion.' *J Afr Hist*. 36:173–195.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using procrustes analysis. *Stat Appl Genet Mol Biol*. 9: Article 13.

FIGURES

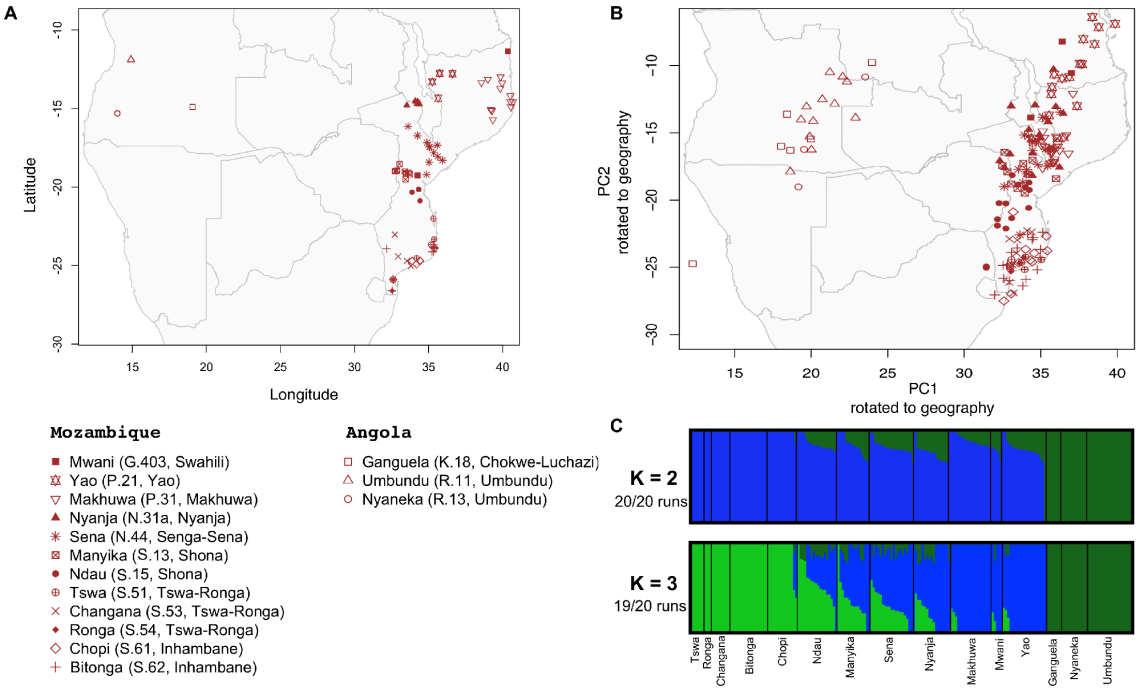

FIG. 1. Genetic structure in Angolan and Mozambican populations. (A) Geographic locations of sampled individuals. The geographic subgroups of Bantu languages ("Guthrie zones") following Maho (Maho 2003) are given in parentheses in the legend. (B) Principal components 1 and 2 of Angolan and Mozambican individuals rotated to fit geography (Procrustes correlation: 0.89; P<0.001). (C) Population structure estimated with ADMIXTURE assuming 2 and 3 clusters (K). Vertical lines represent the estimated proportion of each individual's genotypes that are derived from the assumed genetic clusters (note that the order of individuals in K=2 is not the same as K=3). The lowest cross-validation error (CV) was associated with K=2 (CV values are reported in supplementary table 2).

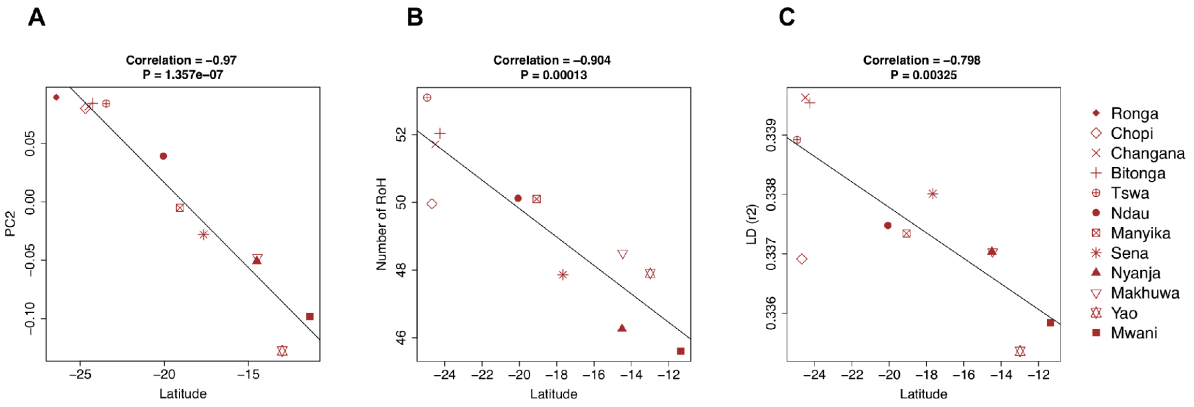

FIG. 2. Genetic variation and geography in Mozambique. The plots show the correlations between latitude and (A) average PC2 scores (supplementary fig. 1B) (B) average number of RoHs, and (C) average LD (r^2). In B and C, Tswa and Ronga were lumped and are identified by the Tswa symbol (see Supplementary Material).

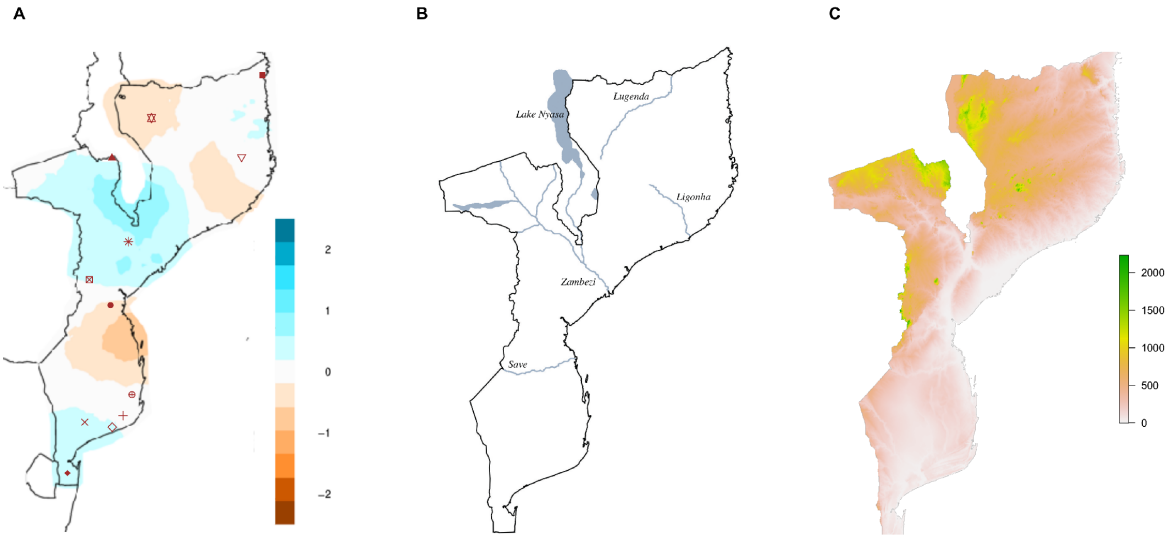

FIG. 3. Estimated Effective Migration Surface (EEMS) analysis. See fig. 1 for legend of population symbols. (A) EEMS estimated with 12 Mozambican populations. (B-C) Major rivers (B) and mountains (C) associated with barriers and corridors of migration. The effective migration rates are presented in a log10 scale: white indicates the mean expected rate in the dataset; blue and brown indicate migration rates that are X-fold higher or lower than average, respectively. The orographic map (C) was generated with the raster package (Hijmans and van Etten 2011). Altitude is given in meters.

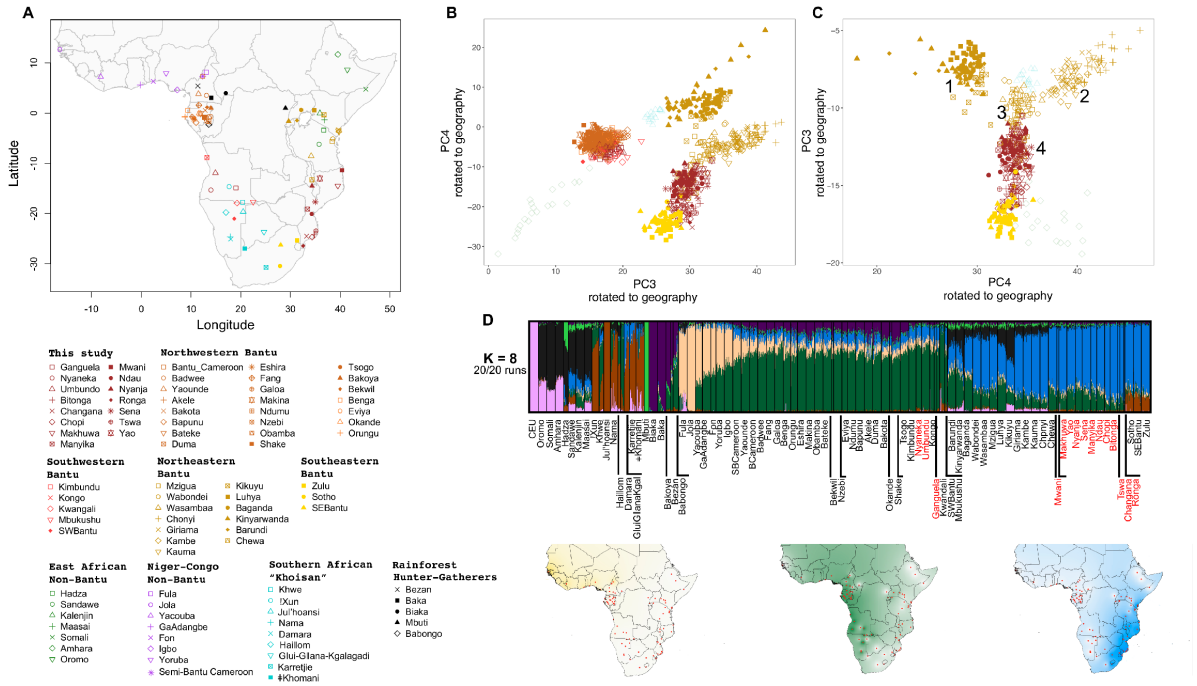

FIG. 4. Genetic structure in African populations. (A) Geographic locations of sampled populations. (B-C) PC plots rotated to geography using Procrustes analysis. (B) All Bantu-speaking populations (Procrustes correlation: 0.76; P< 0.001). (C) Only East Bantu-speaking populations (Procrustes correlation: 0.44; P<0.001). The numbers in (C) refer to groups of populations that are discussed in the text. Additional PCA and ADMIXTURE plots are shown in supplementary figs. 6 and 8. (D) Population structure estimated with ADMIXTURE assuming 8 clusters (K=8), with Mozambican and Angolan groups from this study labeled in red. Vertical lines represent the estimated proportions of each individual's genotypes that are derived from the assumed genetic clusters (CV values are reported in supplementary table 2). The maps, obtained by interpolation, display the mean proportions of major ADMIXTURE components (K=8) from Niger-Congo-speaking populations. The colors in the maps match the colors in the ADMIXTURE plot.

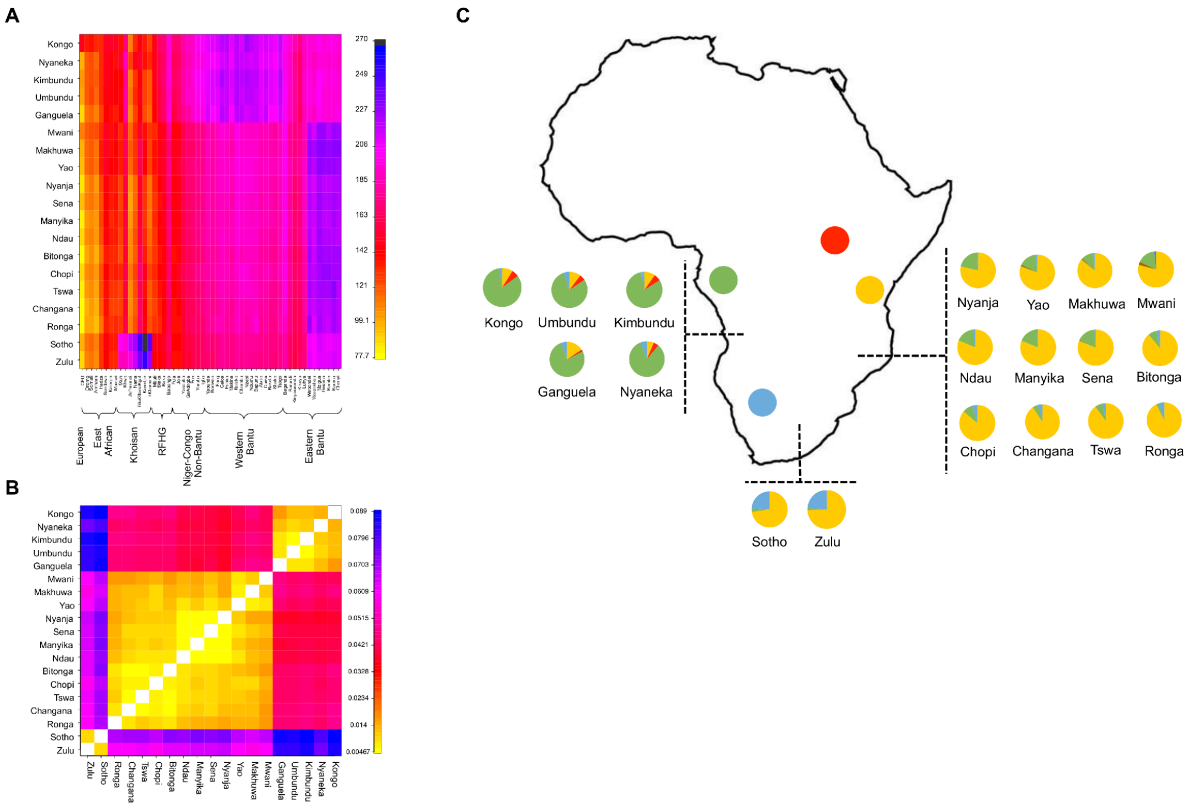

FIG. 5. Inferred ancestry of Bantu-speaking groups from Angola, Mozambique and South Africa. (A) CHROMOPAINTER coancestry matrix based on the number of haplotype segments (chunk counts) shared between representative donor groups (columns) and recipient populations (rows) from Angola, Mozambique and South Africa. The copy profile of each recipient group is an average of the copy profiles of all individuals belonging to that group. (B) Matrix of pairwise TVDxy values based on the ancestry profiles of Angolan, Mozambican and South African groups. The scales of chunk counts and TVDxy values are shown to the right of the matrices in (A) and (B), respectively. (C) Ancestry profiles of Angolan, Mozambican and South African populations (pie charts) as inferred by the MIXTURE MODEL implemented in GLOBETROTTER. The colored circles indicate the most important contributing regions where best source populations were found: West Bantu-speaking groups (green); Tanzanian East Bantu-speaking groups (yellow); Great Lakes Bantu-speaking groups (red); and Khoisan groups (blue).

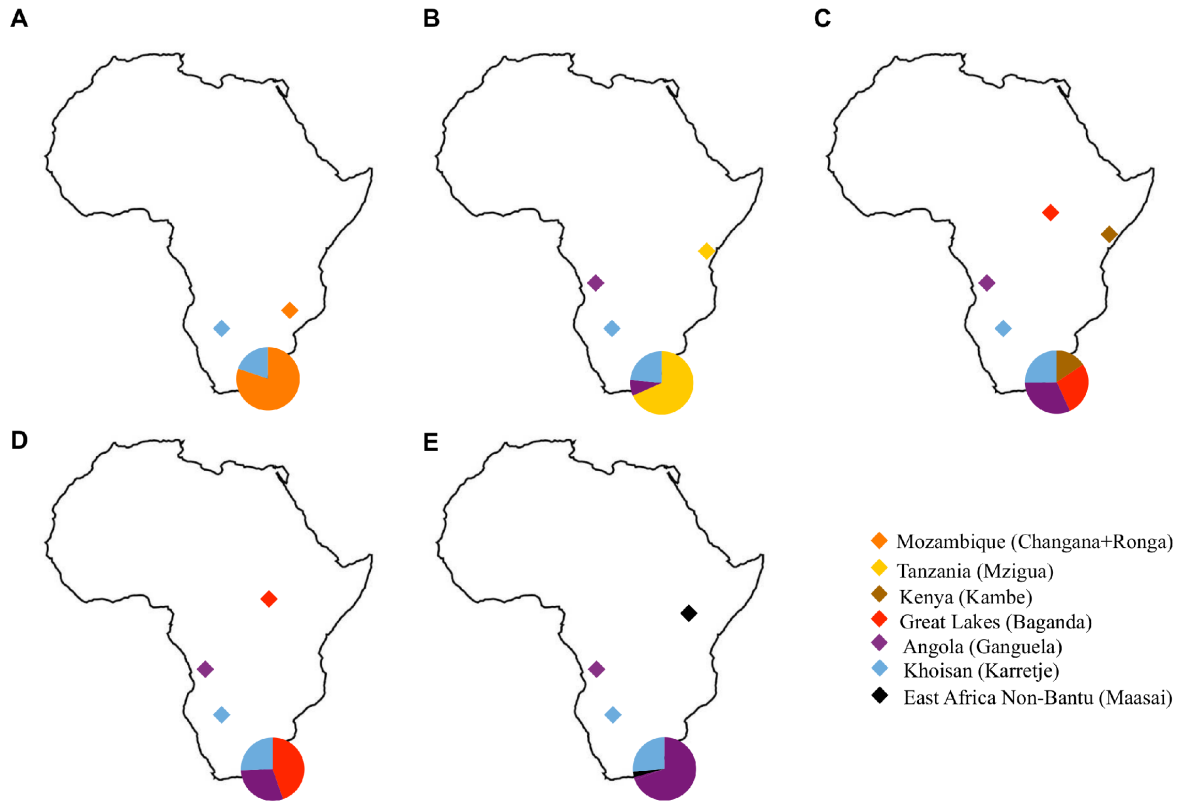

FIG. 6. Inferred average ancestry of Bantu-speaking groups from South Africa. The most important contributing regions and best source populations are provided in the legend. (A) 71 source populations from Sub-Saharan Africa. (B) As in (A), but removing Mozambique from the list of sources. (C) As in (B), but removing Tanzanian Bantu speakers from the list of sources. (D) As in (C), but removing Bantu speakers from coastal Kenya from the list of sources. (E) As in (D) but removing Bantu speakers from the Great Lakes from the list of sources. Full lists of source populations are provided in supplementary table 7.