Received: 17 June 2021 | Revised: 31 May 2022 | Accepted: 17 June 2022

DOI: 10.1002/ajpa.24589

RESEARCH ARTICLE

AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY
The Official Journal of the American Association of Biological Anthropologists

WILEY

Exploring the relationships between genetic, linguistic and geographic distances in Bantu-speaking populations

Miguel González-Santos[1] | Francesco Montinaro[1,2,3] | Rebecca Grollemund[4] | Davide Marnetto[2] | Masharip Atadzhanov[5] | Celia A. May[6] | Nedio Mabunda[7] | Pierre de Maret[8] | Ockie Oosthuizen[9] | Erica Oosthuizen[9] | Cristian Capelli[1,10]

[1]Department of Zoology, University of Oxford, Oxford, UK

[2]Estonian Biocentre, Institute of Genomics, Tartu, Estonia

[3]Department of Biology-Genetics, University of Bari, Bari, Italy

[4]Departments of English and Anthropology, University of Missouri, Columbia, Missouri, USA

[5]School of Medicine, Department of Internal Medicine, University of Zambia, Lusaka, Zambia

[6]Department of Genetics & Genome Biology, University of Leicester, Leicester, UK

[7]Instituto Nacional de Saúde, Distrito de Marracuene, Maputo, Mozambique

[8]Faculté de Philosophie et Sciences Sociales, Université Libre de Bruxelles, Brussels, Belgium

[9]School of Medicine, University of Namibia, Windhoek, Namibia

[10]Dipartimento di Scienze Chimiche, della Vita e della Sostenibilità Ambientale, Università di Parma, Parma, Italy

Correspondence

Miguel González-Santos and Cristian Capelli, Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK.

Email: m.gonzalezsantos@gmail.com and cristian.capelli@unipr.it

Funding information

European Regional Development Fund, Grant/Award Numbers: 2014-2020.4.01.16-0024 MOBTT53, 2014-2020.4.01.16-0030; Fundação para a Ciência e a Tecnologia, Grant/Award Number: SFRH/BD/90648/2012; John Fell Fund, University of Oxford; Leverhulme Trust, Grant/Award Number: RPG-2013-298; University of Oxford Boise Fund; Wenner-Gren Foundation

Abstract

Objectives: The predominance of Bantu languages in sub-Saharan Africa has sparked a large debate over the processes through which they came to disperse over time and space—the "Bantu expansion." The overall genetic similarity shown by Bantu-speaking populations indicates that movement of people occurred too, but the extent of the correlation between genetics, linguistics and geography has been a matter of debate among scholars of different disciplines. In this work, we aim to investigate how genetic, linguistic and geographic distances relate to each other in Bantu-speaking populations.

Methods: We analyzed genome-wide SNP array data from a set of 37 Bantu and non-Bantu-speaking populations together with related linguistic and geographic data. Due to the complex demographic relationships resulting from events of admixture in the history of these populations, we develop and implement a method for controlling the signatures of admixture.

Results: Genetic distances were only minimally correlated with linguistic and geographic distances, possibly as the result of gene flow from neighboring groups into Bantu-speaking populations. When signatures of admixture are controlled for, the correlation of genetic data with linguistic and geographic distances significantly increases.

Discussion: The increase of the correlation between linguistic and genetic distances after the signatures of admixture are taken into account is in agreement with a scenario of spatial co-dispersal of languages and people. Additional specific cultural and demographic dynamics have probably further affected the relationship between language and genetics, which will be necessary to take into account when integrating multidisciplinary data to reconstruct the history of populations.

KEYWORDS

admixture, African populations, bantu expansion, SNPs

1 | INTRODUCTION

Almost a third of the people living in Africa speak a language belonging to the Bantu family, which is part of the Niger-Congo phylum and by far the largest linguistic family in the continent (de Maret, 2013; Simons & Fennig, 2018). The predominance of these languages across most of the continent has sparked a large debate over their common origin and the processes through which they dispersed over time and space. Besides a few outliers, the close relatedness of these languages suggests that the distribution observed today results from a relatively recent and rapid dispersal—the so-called "Bantu expansion" (Bostoen et al., 2015; de Maret, 2013; Johnston, 1886; Vansina, 1979). The origin of the Bantu languages in present-day Cameroon near the border with Nigeria is well supported by linguistic and archeological studies (de Maret, 2013; Grollemund et al., 2015; Lavachery, 2001). Studies also indicate that climate and environmental changes (namely the contraction of the rainforest creating the "Sangha River Interval") might have facilitated a fast migration of these communities (Bostoen et al., 2015; Grollemund et al., 2015). An association with the spread of cultural elements has been also suggested, in particular iron smelting and agriculture, even if not in the earliest stages (Mitchell, 2002). Moreover, new evidence points to a spread-over-spread model for the dispersal of Bantu languages (Seidensticker et al., 2021).

An increasing amount of genetic data indicates shared ancestry among Bantu speakers, supporting an actual migration of people across Africa (Busby et al., 2016; de Filippo et al., 2012; Li et al., 2014; Patin et al., 2017; Tishkoff et al., 2009). The suggestion of a demic process brought additional questions, some still contentious today. Among these, some address the actual routes taken during this expansion, as these have clear implications for the genetic diversity of present-day Bantu-speaking populations (BSPs). Others question the degree of association between the cultural and genetic elements of this diffusion. However, poor resolution of the genetic markers investigated and limitations in representative sampling across the continent have restricted our ability to answer such questions. The complexity of how biological and cultural markers can be used to map population histories has been recently explored in populations from northeast Asia, highlighting how different linguistic features might be related to genetic history, different features possibly operating at different time-depths (Matsumae et al., 2021).

Before the 21st century, the reconstruction of this dispersal has been mostly based on the interpretation of linguistic, archeological, and historical data (Bastin et al., 1983; Bastin et al., 1999; de Maret, 2013; Heine, 1973; Heine et al., 1977; Henrici, 1973; Lavachery, 2001; Vansina, 1990). The two main migratory routes proposed differ primarily on when and where the BSPs crossed the Equatorial forest barrier. The "early split" model (ES), states that BSPs would have split early in their evolutionary history, with one group moving South from their homeland to most of Central and South-West Africa, and others traveling East, at the North of the rainforest, in the direction of the Great African Lakes. Only after reaching this region, would they travel South and settle across the eastern coast. On the other hand, the "late split" model (LS), proposes that the separation between Bantu groups would have only occurred after the first migration South across the rainforest. South of the forest, this common population would have then split into two groups, one traveling further South along the coastline, and the other traveling East, both in the direction of the Great Lakes and further South along the East coast of the continent (Figure 1a,b). With the appearance of phylogenetic methods applied to linguistic data in the 2000s, a flurry of phylogenetic studies emerged (Currie et al., 2013; Grollemund et al., 2015; Holden, 2002; Holden et al., 2005; Holden & Gray, 2006; Rexová et al., 2006) in order to solve the Bantu phylogeny and by extension, to explain the Bantu expansion. As studies progressed, the picture of the Bantu migrations became clearer with all these studies being in favor of the LS model. Geneticists also joined the discussion and previous attempts trying to reconcile linguistic and genetic distances have provided stronger support for the LS model (de Filippo et al., 2012). However, most of the investigated populations were from Central-West Africa and significant support was obtained only for uniparental markers, while autosomal STR data were inconclusive. More recent work focusing on genome-wide data from Mozambique and Angola similarly supported a LS model (Semo et al., 2020).

In order to investigate the patterns of human variation associated with the Bantu expansion and explore the relationships between genetic, geographic and linguistic data from BSPs, we assembled genome-wide SNP data from an array of African populations. Given the complex demographic history of these populations, we implemented a method aimed at controlling for the effects of recent admixture, which is expected to significantly shape the genetic diversity observed today (Hellenthal et al., 2014). In this work, we aimed to explore the extent of the correlation between genetic, linguistic and geographic distances in BSPs, with a particular focus on the impact of gene-flow on these relationships. Our results show that a) the interactions among sympatric communities affected BSPs genetic structure and, b) taking gene flow into consideration improves substantially the relationships between genetics and linguistics/geography. We suggest that genes and languages experienced different processes in their dispersal, complicating the attempts to reconcile them under a simple unifying model.

2 | MATERIALS AND METHODS

2.1 | Samples

DNA samples were obtained through saliva samples collected during several field trips using the Oragene® DNA collection kits (DNA Genotek, Inc., Ottawa, Ontario, Canada) and extracted according to the manufacturer's protocols. All participants were healthy adults from whom informed consent was obtained. The project and consent forms were explained to all participants by local collaborators who spoke

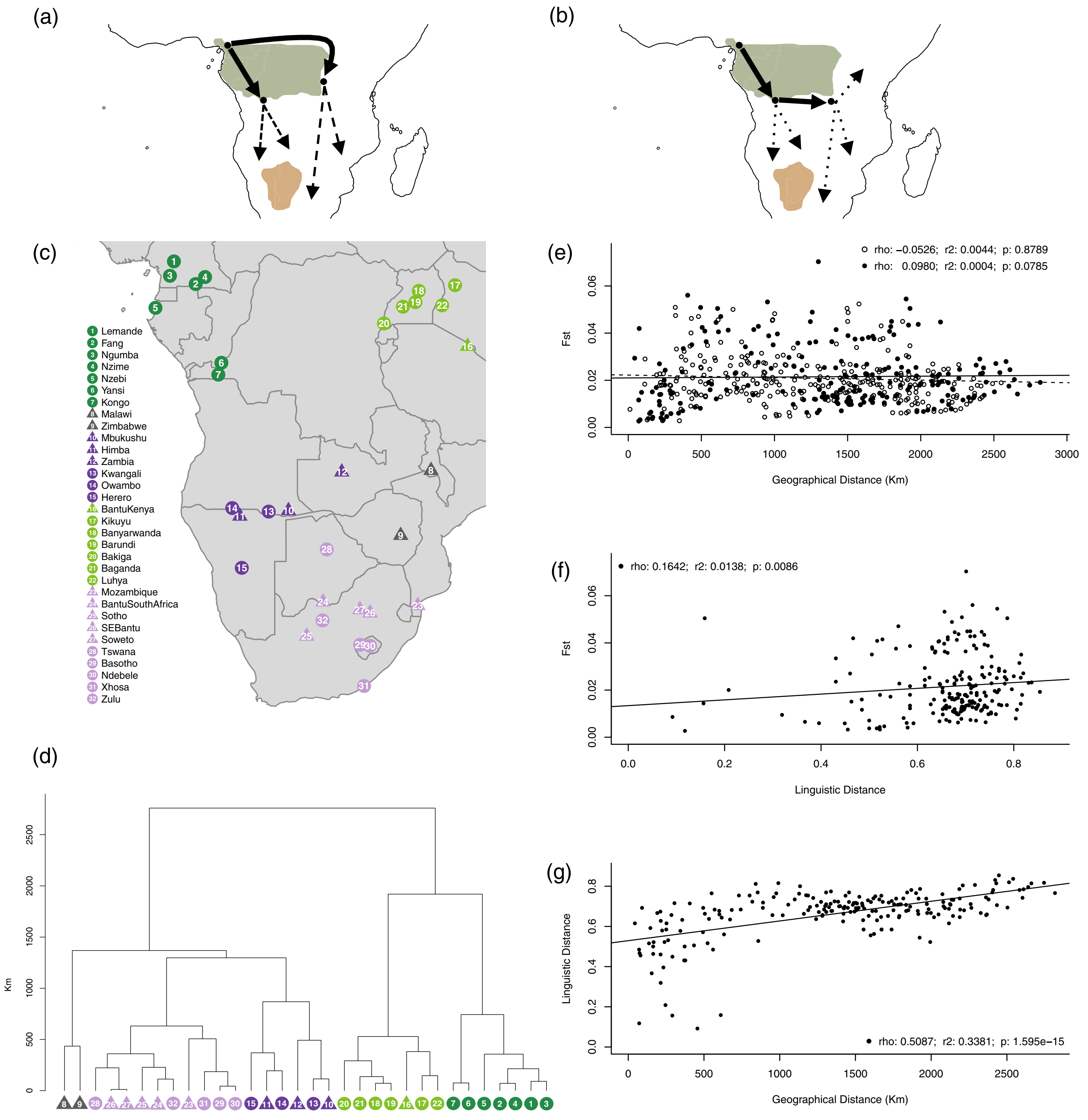

FIGURE 1 (a and b) Schematic representation of the dispersal of the Bantu languages according to the (a) early or (b) late split model (see main text). (c) Location of the Bantu-speaking populations included in the study. Circles represent samples for which both genetic and linguistic data are available (included both in the "complete" and "overlap dataset"), while triangles represent samples for which only genetic data is available (only included in the "complete dataset"). Colors refer to the five geographical clusters identified in panel d. (d) Hierarchical clustering of the great-circle distances, identifying five geographical clusters: Dark green, *cluster A*; light green, *cluster B*; dark purple, *cluster C*; gray, *cluster D*; and light purple, *cluster E*. (e) Correlation between geographical (great-circle distances) and genetic (F_{ST}) distances; open circles correspond to comparisons only observed in the "complete dataset" and full circles correspond to comparisons observed in both datasets. (f) Correlation between linguistic and genetic distances. (g) Correlation between geographical and linguistic distances.

local languages. After the explanation about the study and participation, space was given to the participant to ask for clarifications before they signed the consent form. In addition, a copy of the consent form including information about the study and contacts for future questions was provided. The ethnic and linguistic background of the donors, as well as from their parents and grandparents, was surveyed through a questionnaire. The data from Basotho (Marks et al., 2012; Marks et al., 2015), Owambo, Mbukushu, and Kwangali (González-Santos et al., 2015; Montinaro et al., 2017), and Mozambique and Zimbabwe (Ongaro et al., 2019) were previously published elsewhere.

A total of 41 novel samples from sub-Saharan Africa populations were genotyped with the Human Omni5-Quad BeadChip (Illumina, San Diego, CA, USA) in this study (Table S1). The Ndebele and Himba samples were collected in Lesotho (2009; OxTREC 28-08) and Namibia (2010; OxTREC 49-09 and OxTREC 42-11), respectively (projects reviewed by the Oxford Tropical Research Ethics Committee—OxTREC). The Yansi and Kongo samples were collected from the Democratic Republic of Congo individuals living in Belgium (2015; 20/1/2015-ULB20115) by PdM. Additional DNA samples from Zambia were provided by MA as part of a project approved by local IRB (The University of Zambia Biomedical Research Ethics Committee [UNZABREC], IRB 001131).

2.2 | Genetic dataset

The software PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007) was used to merge our novel data with available genome-wide SNP data from the literature genotyped on different Illumina and Affymetrix platforms (Figure 1c; Table S1).

The data went through a quality control (QC) process before any analysis was performed. Each individual dataset was first processed using PLINK to remove markers and then individuals with a missing call rate higher than 10%. All variant positions were also lifted to build 37 of the Human Genetic map using data provided by either Illumina or Affymetrix, accordingly.

After this first QC step, all populations were merged in a single data file. Merging genotype data produced with both Illumina and Affymetrix platforms has been previously employed with success, with no evidence of errors or biases (Henn et al., 2012; Montinaro et al., 2017; Reich et al., 2009). Then, an additional QC step was performed by removing variants and afterwards individuals with a missing call rate higher than 2%. In order to overcome the effects of markers in strong linkage disequilibrium, all markers with a correlation (r^2) greater than 0.4 were also removed, using a sliding window of 200 SNPs, shifted at 25 SNPs intervals (Behar et al., 2010). A final set of 10,809 SNPs was retained and used in all the analyses. We evaluated the performance of a reduced SNP dataset in recovering population relationships by estimating the degree of correlation between F_{ST} distances (Wright, 1949) calculated using 10,809 and 107,738 SNPs, by including a same subset of populations/individuals typed only with Illumina platforms (Table S1).

After the datasets were assembled the software KING (Manichaikul et al., 2010) was used to infer kinship between samples. All pairs of individuals with a kinship rate higher than 0.0884 (up to second-degree relationship) had one of the individuals randomly removed. The inclusion of reference populations for potential sources of admixture in some of the analyses highlighted two individuals (both from Mozambique) with an Eurasian genetic profile. These samples probably represented very recent non-African influence and were then removed. A total of 1212 individuals from 32 BSPs and 250 samples from five additional reference populations were available for analyses (Table S1).

2.3 | Linguistic dataset

The linguistic data on Bantu languages were selected from Grollemund et al. (2015). To obtain this dataset 100 basic vocabulary words were considered and cognate sets for each of these words were identified and coded as discrete multistate characters (Table S2). The wordlist was based on a Swadesh list replacing some words which are not relevant for African languages with words that are more stable and informative for these languages. In order to produce the cognate sets that are used to calculate linguistic distances, we compared words. If two words, with a similar meaning, present a similar phonetic form, we consider them as cognates, indicating that they might be related. The distance matrix between languages was estimated based on the Hamming distance, counting the number of differences between pairs of sequences/characters. A neighbor-joining algorithm with sequential agglomeration (data were combined into progressively larger overlapping clusters) was used to construct and subset a phylogenetic network based on the linguistic distance matrix of the populations (Table S3; Saitou & Nei, 1987). Lemande was here used as an outgroup since it belongs to the group of languages that have been shown to be the first to diverge within the Bantu languages tree (Grollemund et al., 2015). We also note here that Mbuti and other rainforest hunter-gatherers speak Bantu languages but were not included within the Bantu speakers dataset as representing known examples of linguistic transitions (Patin et al., 2009). The tree was constructed using the Splitstree software using the Neighbor-Joining algorithm (Saitou & Nei, 1987; Figure S1).

As the data to build the linguistic distance matrix was not always available—either because (a) the exact Bantu language spoken was not known, or, even if known, (b) linguistic data for the construction of the distance matrix was not available—a subset of the "complete dataset" (which comprises all the populations for which genetic data was available) was assembled to include only those populations that could be included in the linguistic distance matrix—the "overlap dataset" (Table S1). We also investigated a larger linguistic dataset comprising 416 Bantu-speaking populations and 3876 cognate sets coded as binary characters (Grollemund et al., 2015). The results of the correlation analysis with geography were compared across these datasets to evaluate if major differences emerged when subsets were analyzed.

2.4 | Geographical dataset

The geographical coordinates of the populations were registered during fieldwork or retrieved from the original studies. For populations not sampled in the country of origin or with no geographical information available, we used the coordinates of the capital city of the country of origin (Figure 1c; Table S1). The geographical distance between pairs of populations was calculated through two different approaches: the great-circle line distances and model-based distances.

In the first stage, the great-circle line distances between each pair of populations was calculated. However, it is known that geographical barriers played a major factor in the dispersal of human populations and thus the great-circle distances are not a true representation of the real distances that populations had to travel (Liu et al., 2006; Prugnolle et al., 2005). Model-based distances were thus calculated based on how clusters of neighboring populations were connected to each other through various waypoints (see section 3). The clusters of populations—solely based on geography and not taking into consideration other factors—were identified through a hierarchical clustering of great-circle distances between pairs of populations. We found that a cut-off of the population tree at 1000 km generated a relatively small number of clusters (5) consistent with their geographical location (Figure 1c,d). Given the geographical location of the clusters of populations we identified one or more waypoints—*mA*, *mB*, *mC1*, *mC2*, *mD1*, *mD2*, and *mE*—to represent the borders of each cluster. These waypoints (with the exception of *cluster E* due to the distribution of populations in this cluster) corresponded to the combination of either the northernmost or the southernmost population coordinates with the easternmost or the westernmost coordinates in that cluster. For waypoint *mE*—as its location would not be an accurate representation of the cluster's borders with the previous method—the coordinates were obtained as the midpoint between the northernmost and easternmost population. Additionally, some of the midpoints between these waypoints were also identified—*mAB* and *mABD*. The different paths aimed to represent different points of split between East and West Bantu languages more than providing a direct test for the hypotheses of dispersal of Bantu languages and therefore our results should be interpreted with due caution in this regard.

Nevertheless, for completeness, we included an additional set of waypoints—*tA* and *tB*—to design in a simplified way the path possibly followed under the assumption of an early split of the Bantu languages. These waypoints were identified by considering the latitude of the northernmost and the longitude of the easternmost (*A*) or westernmost (*B*) population in the two clusters, respectively, as done for the identification of other waypoints. For each pathway, the geographical distances for populations within each cluster were calculated as great-circle distances. The different pathways differ in the way populations in different clusters are connected—the distance between them being the sum of the distance between each population and the corresponding waypoint for its cluster, and the distance between waypoints (directly or through other waypoints; see Section 3). All distances were calculated with the function *rdist.earth* of the package *fields* (Nychka et al., 2015).

Due to the different patterns of variation within and between clusters for linguistic and genetic distances, we introduced a series of modifications to all the pathways as a way to assess how these modifications altered the correlation coefficient. Furthermore, some of the modifications focused specifically on *clusters A* and *E* since they showed, respectively, the most divergent patterns for linguistic (highest within clusters heterogeneity) and genetic (highest between clusters differentiation) distances.

2.5 | Data analysis

2.5.1 | Genetic distance corrections

The software ADMIXTURE v1.23 (Alexander et al., 2009) was used to explore the genetic variation among populations in the study. This method allows for a model-based estimation of cluster allocation by implementing a maximum likelihood algorithm assigning individuals to a predefined number of clusters (*K*). The cross-validation (CV) procedure implemented in the software—which assesses the consistency between different runs of subsets of the data at any given value of *K*—was used as an indication of the most supported value of clusters, assuming that a well-supported division should have a relative lower CV error (Alexander et al., 2009; Alexander & Lange, 2011). Each value of *K* was run for several iterations until the log-likelihood between iterations increased by less than 10^{-4} (Alexander et al., 2009). We refer to these different clusters as "components," as they are often combined in different proportions to compose the profile of populations.

Given that ADMIXTURE is not a formal test for gene-flow between populations, we performed three-population admixture tests (*f3* statistics) that are based on the concept that shared genetic drift between populations implies a shared evolutionary history (Reich et al., 2009). Briefly, in a *f3* statistic with the form *f3(X;PopA,PopB)* a significantly negative value of the statistic (*Z-score* <-3) highlights a complex phylogeny for the target population ("X"), as the result of a certain amount of ancestry from populations related to *PopA* and *PopB*. All the *f3* tests were performed for windows of 100 markers using the *threepop* companion software in the TreeMix suite (Pickrell & Pritchard, 2012).

To mitigate the effects of recent admixture in BSPs, we developed a method to correct the allele frequencies used to calculate the genetic distance between pairs of populations. Haplotype-based approaches could not be implemented due to the low SNP density resulting from the merging of datasets genotyped on different platforms (Lawson et al., 2012). We based this correction on the two elements of the results of the ADMIXTURE analyses, (for $K = 7$, see section 3): (i) the fractions of each component for each individual in the dataset (Q-file); and (ii) the allele frequencies of the inferred components (P-file). From the Q-file we isolated Niger-Congo-specific components present in each BSP (see section 3) and normalized these frequencies so that the total sum of the components retained was equal to 1. By doing so, we regenerated BSPs as only composed of Niger-Congo components, removing most of non-Niger-Congo influences. The rationale of this approach is to try to remove the impact of admixture on the extant population and ideally reconstruct the allelic profile of the ancestral "un-admixed" population. In order to obtain the putative allelic frequencies in these original populations we multiplied, for each allele, the fraction of each ancestral component by the frequency of the allele in the corresponding ancestral population (from the P-file).

2.5.2 | Simulations

We validated our correction approach by implementing a series of simulations using the program *admix-simu* (https://github.com/williamslab/admix-simu). We considered five source populations (CEU, Somali, Mbuti, Ju/'hoansi, and Yoruba) and simulated admixture between Yoruba and one of the other sources occurring 40 generations ago with mixing proportions $\alpha \in (0.1, 0.2, 0.5, 0.8, 0.9)$. The correction method described above was then applied to the simulated datasets. For each pair of sources (Yoruba and each one of the other four populations) and the corresponding simulated populations, an ADMIXTURE analysis was performed and the results for $K = 3$ were used for the correction (in order to allow the identification of more than one potential Niger-Congo component in the simulated samples). For consistency, the reconstruction of the allelic frequencies was applied to all populations, simulated and sources, with the latter being reconstructed using all three of the identified components. A comparison between the F_{ST} between the simulated populations and their sources was performed both before and after the correction to visualize the changes in population affinity generated by our approach.

2.5.3 | Correlation tests

In order to generate matrices of genetic distances, the pairwise F_{ST} between pairs of populations was calculated through a custom-made script, using the classic Wright's measure of F_{ST} (Wright, 1949). As the index value is calculated for each individual marker, in order to combine the estimates across multiple SNPs and estimate the genome-wide F_{ST} value we used Weir and Cockerham's approach (Bhatia et al., 2013; Weir & Cockerham, 1984). The *pvclust* R package was used to build a hierarchical tree based on the pairwise F_{ST} values using the complete linkage method, and to assess the significance of its topology based on 10,000 bootstrap replications (R Core Team, 2016; Suzuki & Shimodaira, 2006).

In order to test for correlations between the genetic, geographical, and linguistic distances between pairs of BSPs, we performed a Mantel test—using the non-parametric Spearman's rank correlation method—using the *vegan* R package on all our distance matrices (Oksanen et al., 2017). A linear regression analysis was fitted to the data and the coefficient of determination—r^2, explaining the proportion of variation in one of the variables that is explained by the other variable analyzed—was calculated. This coefficient was used to evaluate the fit of the various modifications tested on the data and for comparisons across different scenarios. We additionally explored the correlation of the three variables (genetics, linguistics and geography) by the way of a Procrustes analysis using the *vegan* R package (Oksanen et al., 2017). Briefly, we started by standardizing the genetic and linguistic data for all variables and calculating a Principal Component Analysis of each set of data. A Procrustes test was then performed across the PCA results and its significance was tested based upon 10,000 permutations (Peres-Neto & Jackson, 2001).

3 | RESULTS

3.1 | Correlations between geographic, genetic and linguistic distances

We aimed to explore the relationships among BSPs in sub-Saharan Africa by combining and comparing their spatial, genetic, and linguistic variation (Figure 1; Tables S1–S3; Figure S1). In order to do so, we started by testing the correlation between different pairwise distances in our dataset—genetic (F_{ST}), geographical (great-circle distances), and linguistic distances (based on the number of differences between pairs of words; Grollemund et al., 2015). As the linguistic distance matrix represented only a subset of the total samples in the dataset, comparisons between genetics and geography were done both for the "complete dataset" and the "overlap dataset" (Table S1). The geographical and genetic distances showed no significant correlation, for either dataset ($p > 0.05$; Figure 1e). On the contrary, linguistic distances showed a significant ($p < 0.05$) positive correlation with both geographical and genetic distances (Figure 1f,g). However, even if statistically significant, the amount of variation explained is small, less than 2% when linguistics is correlated with genetics, but above 30% with geography.

The restricted SNP dataset generated F_{ST} estimates highly correlated with those calculated when a number of SNPs 10 times larger was considered ($r^2 = 0.999$; $p < 0.001$; Figure S2). The amount of variation explained by geography for the more comprehensive linguistic dataset—composed of 416 Bantu-speaking populations and 3876 cognate sets—was virtually identical to the one used in this study (34% and 33%, respectively; Figure S3). Due to the larger number of information analyzed in the more comprehensive linguistic dataset the data in this dataset had to be coded as binary characters instead, leading to a decrease in the overall distances obtained. However, the encoding of the data does not affect the topology of a tree for a given set of populations—as evidenced by the very strong correlation ($r^2 = 0.93$) observed between the distances for the two linguistic datasets (Figure S4). These results validate the use of a smaller number of genetic markers and populations in our analyses.

3.2 | Signatures of admixture in BSPs

In order to better understand the associations between genetics and the other two variables, we investigated the genetic make-up of our dataset. We ran ADMIXTURE for a range of a possible number of K ancestral populations (from 2 to 10; Figures S5 and S6) in the "complete dataset" and included five additional populations as possible sources of gene-flow [(i) Europeans (CEU); (ii) East African Cushitic

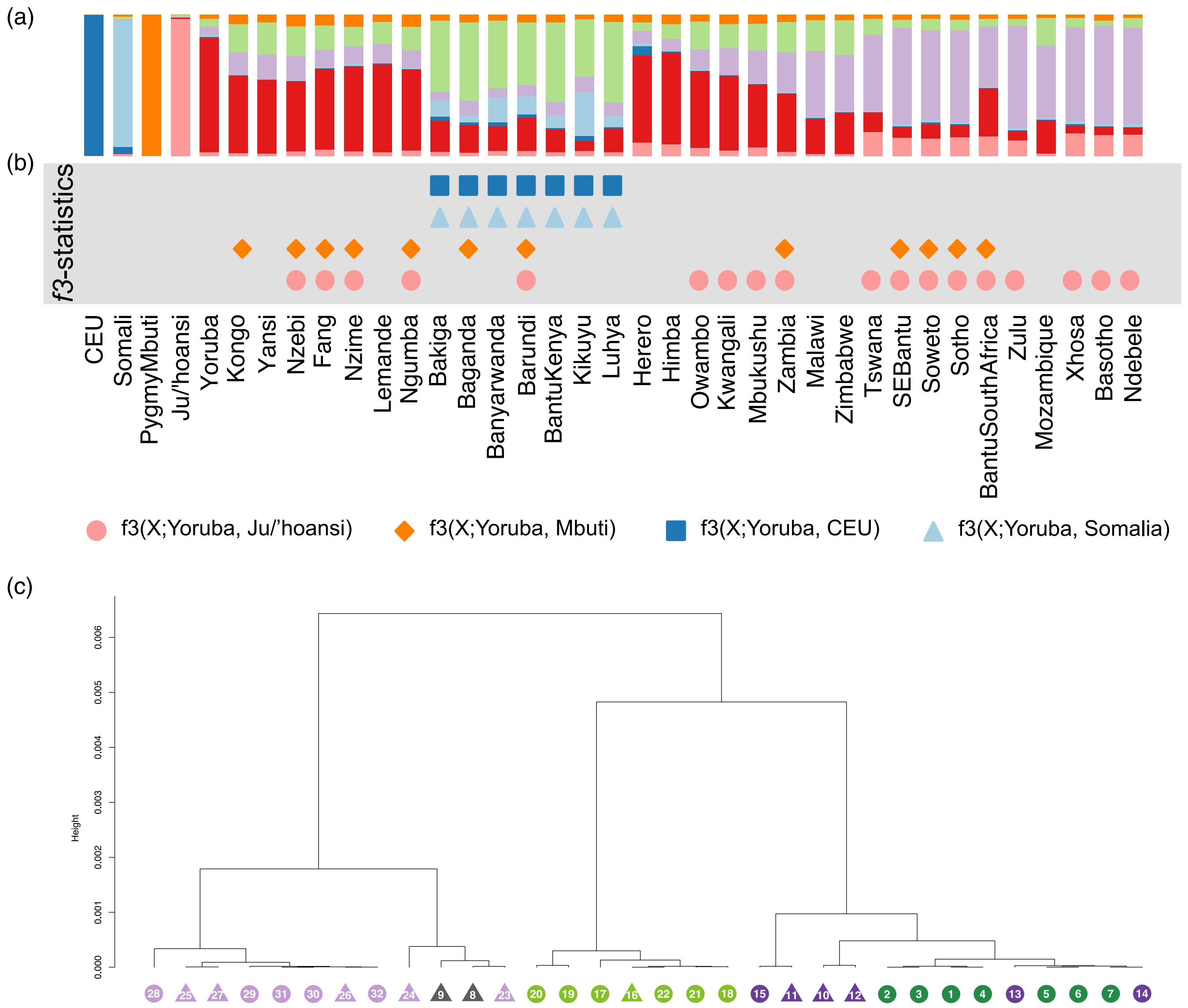

FIGURE 2 (a) ADMIXTURE plot for $K = 7$ for the "complete dataset," plus five source populations. (b) Significant *f3 tests* for different combinations of sources/test population. (c) Hierarchical population tree based on the corrected genetic distances. All the nodes showed bootstrap values above 60%. The bootstrap values are shown in Figure S9. Population labels as in Figure 1.

(Somalia); (iii) Kx'a (Ju/'hoansi); (iv) rainforest hunter-gatherers (Mbuti); and (v) West African Niger-Congo (Yoruba)]. Even though the lowest CV value was observed for $K = 6$, $K = 7$ also showed a relatively low CV error (Figure S5). The main difference observed was that for $K = 7$ the Mbuti population was represented by a single specific component, while two different components (one associated with Ju/'hoansi and another observed across multiple BSPs) were observed for $K = 6$ (Figure S5). For $K = 7$, components strongly associated with both linguistics and geography could be identified, as previously reported (Tishkoff et al., 2009; Figure 2a). One specific component was modal in each of the five populations here used as possible sources of admixture: CEU (dark blue), Ju/'hoansi (pink), Mbuti (dark orange), Somalia (light blue), and Yoruba (red). Two additional components were found in significant amounts almost exclusively in BSPs (light green and purple). BSPs seem to be mainly characterized by different amounts and combinations of these two components and the "Yoruba" component. Given the information related to population-specific components and low CV value associated with $K = 7$, we used this number of clusters in the subsequent analyses.

The three components characterizing all BSPs showed pairwise F_{ST} values below 0.025, lower than the values between any of these individual components and the remaining ones (the lowest being 0.068; Table S4). These three components are not randomly distributed across BSPs. The "Yoruba" component shows higher frequencies in populations from Central-West Africa, while the other two are more prevalent in East Africa BSPs ("Bantu East" component, light green) and in populations from the southern regions of the continent

("Bantu South" component, purple). Non-Niger-Congo components are also observed in several BSPs.

In order to formally test for admixture with non-Niger-Congo groups in BSPs we performed the three-population tests with the format *f3*(*X*;*Yoruba*, "Source")—"Source" being either CEU, Somalia, Ju/'hoansi, or Mbuti (Reich et al., 2009). Significant *f3* statistics (*Z-score* <-3) were observed for several of the tested trios (Figure 2b). Signatures of admixture with Ju/'hoansi are the most widespread in the dataset (18 out of 32 BSPs). Some of these populations, mainly in central-western and southern Africa, showed signatures of gene flow also with Mbuti, which might suggest Ju/'hoansi and Mbuti might act as relatively good proxies for each other or alternatively be an example of an "outgroup case" (Patterson et al., 2012). The widespread signatures of admixture with both Ju/'hoansi and Mbuti reflect the interaction with local communities that BSPs experienced in their dispersal across sub-Saharan Africa (Montinaro et al., 2017; Patin et al., 2014). East Africa BSPs were the only populations where admixture with the neighboring Somalia and CEU was observed, with all BSPs in this region showing signatures of admixture with both sources. However, it is important to note that Somalia also showed significant signatures of admixture with CEU itself (*f3* statistic $= -0.00545$; *Z-score* $= -19.98$; data not shown), possibly associated with the out-of-Africa and subsequent back migration (Pagani et al., 2012).

3.3 | Correlation between corrected genetic distances, linguistic and geographical distances

In an attempt to correct for the impact of gene flow in BSPs we masked the effects of admixture by generating a reconstruction of the original "un-admixed" populations using the results of the ADMIXTURE analysis (see Methods). We validated this approach via simulations, generating mixed populations from two different sources, as described in Methods. The corrected genetic distances for the simulated populations were always closer to the Niger-Congo source population (Yoruba) than the other sources (Somali, Ju/'hoansi, Mbuti or CEU) and the corrected F_{ST} values estimated for Yoruba were substantially smaller than non-corrected ones (Figure S7).

We applied the ADMIXTURE-based correction to our dataset and then calculated the pairwise F_{ST} between the newly generated populations. The tree produced by these distances shows a clustering of BSPs strongly associated with their geographical distribution (Figures 1c,d and 2c; Figures S8 and S9).

We analyzed the impact of this genetic correction on the correlation with both the geographical and the linguistic distances (the original distances will be referred to from here on as "non-corrected genetic distances" and the adjusted ones as "corrected genetic distances"). For the "overlap dataset," both these correlations were now significant ($p < 0.05$) and the r^2 increased dramatically (Figure S10). In the case of linguistic distances, the percentage of variation explained by the genetic distances increased from less than 2% (Figure 1d) to 27% (Figure S10b). When analyzing geographical distances this was even more striking. Before the correction, no statistically significant correlation between the two distances was observed (Figure 1e). On the contrary, geographical and corrected genetic distances were now strongly correlated ($p < 0.05$), with more than 54% of the variation explained (Figure S10a). Similarly, when analyzing the "complete dataset," the variation of the corrected genetic distances explained by geography was larger (42%, $p < 0.05$; Figure S10*a*). From here on, we refer to the corrected F_{ST} when mentioning genetic distances unless otherwise indicated.

These results were also confirmed by the way of a Procrustes analysis, where all the tested correlations but one—between non-corrected genetic distances and linguistic distances— are significant ($p < 0.05$; Table S5). In the case of the genetics distances, the correlation in a symmetric Procrustes rotation is increased for the corrected distances using both the complete (coefficient of correlation increased from 0.4553 to 0.7517) and the overlap datasets (coefficient of correlation increased from 0.4974 to 0.9134) with the geographic distances; the correlation between genetics and linguistic distances becomes significant when corrected genetic distances are used ($p < 0.05$; Table S5).

As some of the populations we analyzed are characterized by small sample sizes, we explored the impact on the analyses when groups with less than 5 (two populations) and 10 individuals (seven populations) were removed (Tables S1 and S6). The removal of groups with small sample sizes generally improved the coefficient of correlation in the non-corrected dataset, for all the comparisons. To the contrary, the analyses without these populations did not always show an improvement in the coefficient of correlation when the corrected F_{ST} was taken into consideration. In line with the observations reported for the full dataset, when populations with less than 5/10 individuals were removed there was an increase of the correlation coefficient reported when estimates based on the uncorrected to the corrected F_{ST} values were compared (Table S6). Overall these results support the F_{ST} correction approach proposed here, with correlation patterns being consistent independently of population sizes; the full dataset was therefore used for the subsequent analyses.

3.4 | Exploring different spatial pathways for the "Bantu expansion"

Great-circle distances are not a true representation of the actual distance that needs to be traveled between any two sites, as they ignore the existence of potential barriers to the movement of people. To evaluate how differences in the way distances between populations are calculated affect the correlation with genetic and linguistic variation, we generated several alternative pathways linking populations and calculated the associated traveling distances (model-based distances; see methods; Figure 3). Given the geographical location of the clusters of populations we initially identified one or more waypoints—*mA*, *mB*, *mC1*, *mC2*, *mD1*, *mD2*, and *mE*—to represent the borders of each cluster (Figure 3a). These waypoints corresponded to the combination of the highest and/or lowest coordinate for both latitude and

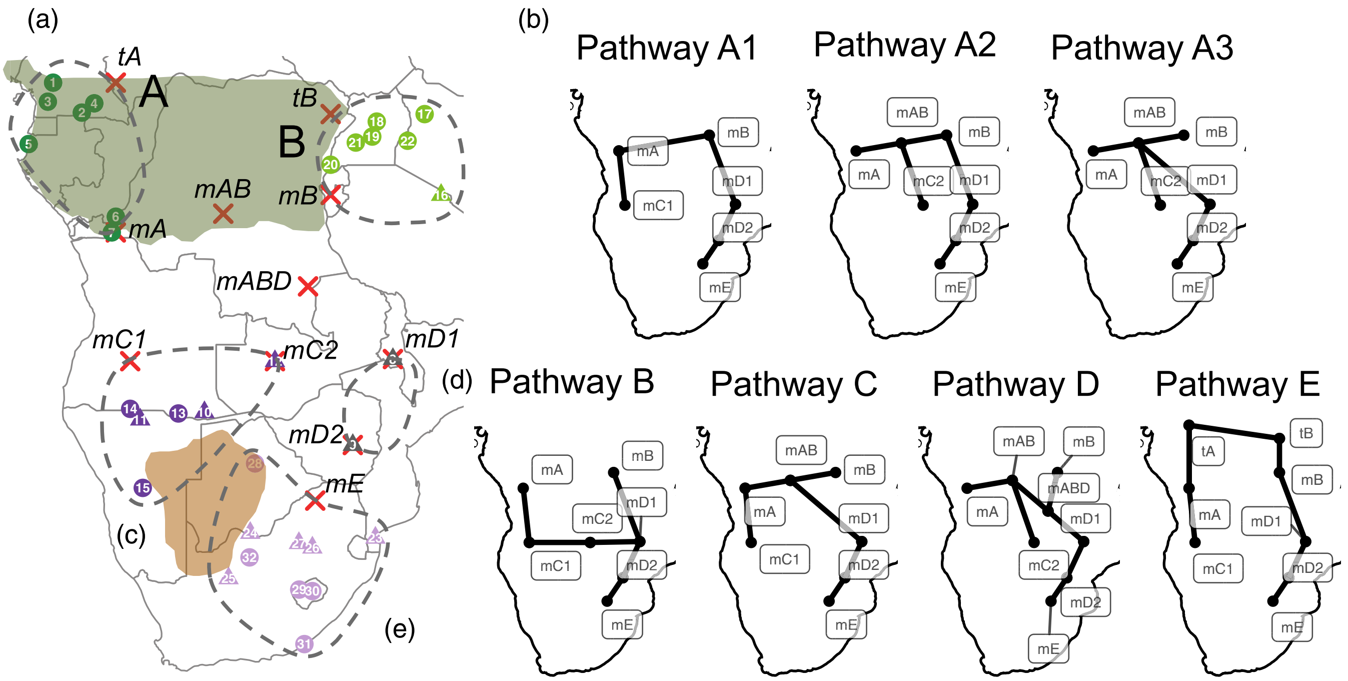

FIGURE 3 (a) Location and composition of the five geographical clusters identified in Figure 1 and waypoints (red crosses) used to estimate the model-based geographical distances; population labels as in Figure 1. (b) Schematic representation of the pathways used for the calculation of the model-based geographical distances.

TABLE 1 Coefficient of determination (r^2) for the correlation between the geographical (great-circle distances or the seven pathways tested) and either the linguistic, non-corrected, or corrected genetic distances.

	Non-corrected genetic distances	Corrected genetic distances	Linguistic distances
GCD	*n.s.*	0.5449	0.3381
Pathway A1	0.0577	0.6290	0.4368
Pathway A2	0.0671	0.5053	0.4742
Pathway A3	0.0490	0.6475	0.4375
Pathway B	*n.s.*	0.4371	0.4315
Pathway C	0.0411	0.7852	0.4008
Pathway D	0.0346	0.5266	0.4953
Pathway E	0.1033	0.5324	0.3596

Note: *n.s.*, nonsignificant.

longitude of each cluster. Additionally, midpoints between some of these waypoints were also identified—*mAB* and *mABD*. By connecting the various clusters via these waypoints we explored how differences in the paths of dispersal across sub-Saharan Africa affected the correlation of linguistic and genetic variation to geography in BSPs.

The use of model-based distances led to a generalized increase in the amount of linguistic and corrected genetic variation explained when compared to great-circle distances (Table 1). However, genetics and linguistics were found to behave differently in response to these modifications. The pathway that seemed to better correlate with genetic variation was *pathway C*, with over 55% and 78% of the corrected genetic variation being explained by geography (in the "complete" and "overlap dataset," respectively). Instead, the pathway with the highest correlation with linguistic distances was *pathway D*, with almost 50% of the variation in linguistic distances explained by geography. The two pathways differ in having larger geographic distances between *clusters B/C* and *D/E* in *pathway C* and *clusters A* and *B/C* in *pathway D*.

We highlight here that our methodology is not specifically designed to test the different hypotheses related to the dispersal of Bantu languages. Rather, we aim to provide some indication of how different pathways might produce results that are more strongly associated with linguistic and genetic distances. For this reason, we also tested an additional pathway linking *cluster A* and *cluster B* in the northern part of the forest distribution using waypoints *tA* and *tB* (Figure 3, *pathway E*). This pathway was associated with the largest coefficient of correlation when analyzed with the non-corrected genetic distances, but the coefficient was consistently among the lowest when this pathway was tested with the linguistic and corrected genetic distances (Table 1).

3.5 | Different patterns of linguistic and genetic variation in geography-based clusters

We further explored the patterns of genetic and linguistic variation by considering the variation within and between the identified geographical clusters (Table S7; Figure S11). We found a homogeneity in the degree of geographical variation within the different clusters (Figure S11a)—which seems to indicate that we do not need to control for geography-related heterogeneity within clusters. Heterogeneity was nevertheless observed within clusters for linguistic distances (Figure S11b). Cluster A showed the highest values among all the clusters, significantly different from all except Cluster C ($p < 0.05$ after Bonferroni correction; Figure S2b), possibly due to the small sample size of the latter (only three populations). The corrected genetic distances within each cluster were overall very low and more homogeneous than the non-corrected ones, as noted before (Figure 2c; Figure S11c,d). *Cluster E* showed the largest set of distances between clusters (Figure S11e). On the contrary, no single cluster showed an excess of differentiation from others when linguistic variation was considered (Figure S11f). The geographic distances between populations in *Cluster E* and all the populations in the other clusters were significantly different across pathways, *pathway E* showing the largest number of significant comparisons (5 out of 6, Figure S11g).

On the basis of these observations, we introduced a few modifications to the pathways in order to explore how different specific elements affected our results. Linguistic and genetic distances appeared to respond differently to modifications to the pathways. Linguistically, the strongest correlation with geography was for *pathway D*, r^2 improving to more than 0.6 when comparisons within *clusters A* and *E* were removed (both individually and together). On the other hand, the modifications that yielded the biggest r^2 improvements for the genetic distances were based on the increasing of the geographic distances between *cluster E* and all the other clusters (Table S7)—*pathway C* being the best distance-based model for the "overlap dataset" (distances increased by 1000 km; $r^2 = 0.81$), and *pathway B* for the "complete dataset" (distances increased by 2000 km; $r^2 = 0.60$). Completely removing populations from *cluster E* from the analysis resulted in *pathway B* as the best fit for both genetic datasets (Table S7). Notably, none of the modifications resulted in *pathway E* generating the largest values of the coefficient of correlation (Table S7).

4 | DISCUSSION

Bantu-speaking populations (BSPs) have been shown to share a striking genetic similarity despite their broad distribution across a vast area of most of sub-Saharan Africa (Busby et al., 2016; Tishkoff et al., 2009). In fact, this shared genetic ancestry has been one of the pieces of evidence used to support the Bantu expansion being an actual movement of people across the continent and not just a cultural spread of languages through neighboring populations (Tishkoff et al., 2009). Nonetheless, this extended shared ancestry does not mean uniformity, and it is noteworthy that some degree of differentiation is found among BSPs. While early studies focusing on Bantu speakers highlighted their relative genetic homogeneity, more recent studies have brought to light their heterogeneity (Choudhury et al., 2017; Patin et al., 2017).

The aim of this work was to evaluate to what extent genetics, geography, and linguistics are related in BSPs. In doing so, we also highlighted some of the elements that affected these correlations the most. The analysis of the distribution of the distances between BSPs showed a strong link between linguistic diversity and geography. As people movements accompanied the Bantu linguistic dispersal, we expected the current genetic variation among its speakers to be strongly defined by geographical proximity and linguistic similarities within the Bantu family, mirroring what is observed when broader ethno-linguistic diversity of African populations was investigated (Busby et al., 2016; Tishkoff et al., 2009). However, the non-corrected genetic distances did not show a significant link to either geography or linguistic distances (Figure 1e).

The dispersal of BSPs and their interaction with local inhabitants of the newly occupied regions have deeply shaped the genetic and cultural variation of sub-Saharan Africa (González-Santos et al., 2015; Patin et al., 2014; Patin et al., 2017; Tishkoff et al., 2009). Admixture dynamics can vary greatly and as a result BSPs show different genetic profiles throughout the continent (Barbieri, Vicente, et al., 2013; Marks et al., 2015; Mitchell, 2002; Montinaro et al., 2017; Patin et al., 2014; Patin et al., 2017; Pickrell et al., 2012; Pickrell et al., 2014; Tishkoff et al., 2009). ADMIXTURE analysis combined with formal tests of admixture pointed to the role that gene flow had in shaping the different patterns of diversity observed in today's BSPs. European/East African admixture was observed in all BSPs from *cluster B* in East Africa. On the other hand, signatures of admixture with Ju/'hoansi/Mbuti are more common in BSPs from Central-West and South-East Africa, in agreement with previous studies (Barbieri, Butthof, et al., 2013; Barbieri, Vicente, et al., 2013; Marks et al., 2015; Rocha & Fehn, 2016). Overall, it is clear that the genetic structure of BSPs is highly influenced by different dynamics of admixture, mostly shaped by geographical proximity with non-Bantu-speaking communities. Isolation by distance dynamics during the Bantu expansion might have further impacted the degree of differentiation observed among these populations.

In order to overcome some of the issues related to admixture, we implemented a method to recover the signal of the ancestral BSPs, before the impact of admixture with native non-Niger-Congo inhabitants. After this correction, there was a general increase in homogeneity across all BSPs, more so for geographically close groups (Figure 2c and Figure S11d). These results were corroborated with a Procrustes analysis in which the correlation of the Procrustes rotation was improved for all comparisons after the correction of the genetic distances (Table S5).

Using model-based distances also led to a general increase in the correlation of genetic and linguistic variation with geography (Table 1; Table S7). Nonetheless, even though both the genetic and linguistic diversity of BSPs are strongly linked with geography, they seem to be shaped by different evolutionary dynamics as the pathways better correlating languages and genetics with geography are different (Table 1; Table S7).

Languages appear to be more affected by the time of their separation, with older clusters of populations (near the Bantu homeland in Central-West Africa) presenting higher levels of language diversity than clusters of populations that settled later (Figure S11b). This seems to be affecting linguistics-geography correlation, as the removal of the comparisons within this region (*cluster A*) led to a generalized increase of the variation explained (Table S7). Linguistically, most pathways improved the correlation with geography by also simultaneously removing comparisons within *cluster E*, in South-East Africa. This effect might be explained by the presence of linguistic structure between the two main groups of southeastern Bantu languages (Nguni and Sotho-Tswana).

Isolation and drift (combined with admixture) appear to play a major role in the genetic differentiation among BSPs, with demographic fluctuations possibly influenced by factors such as overexploitation of resources, pandemics, and climate change (Wotzka, 2006). The pathway that explained the highest amount of genetic variation (*pathway C*) was among those with higher geographical distances with populations from *cluster E* (Figure S11g). Similarly, modifications to the pathways that produced the highest increases in the correlations were those involving increases in the geographical distances to *cluster E* (Table S7). All this seems to indicate that the pairwise F_{ST} involving populations from *cluster E* might be higher than expected based solely on geography. This may indicate that the most supported pathway might be one that is maximizing the geographical distances to these populations to accommodate this. When populations from *cluster E* were removed from the analysis the pathway explaining most of the genetic variation was *pathway B* (56.7% and 68.7% for the "complete" and "overlap dataset," respectively; Table S7). These observations are in line with the reported decrease in genetic diversity moving South along Eastern Africa highlighted in Mozambican and South African populations (Semo et al., 2020). Overall, our results showed that corrections taking in consideration variation in the paths of dispersal and gene-flow increase the degree of correlations between genetics, linguistics and geography, and that these variables should be properly considered when investigating BSPs relationships. However, it should be noted that none of the tested pathways provided the strongest correlation with geography for both genetics and linguistics, as probably too simplistic in their representation of the dispersal of BSPs. We also would like to stress that we did not intend to directly test the different hypothesis for the "Bantu dispersal" but instead explore how such a line of investigation should consider in a more direct way the role played by gene-flow in affecting the biological relationships of populations. The observation that the pathway that mirrored in a simplified way the early split model (*pathway E*) never generated the largest values of correlation (Table S7), except when non-corrected genetic distances were considered (Table 1), should therefore not be interpreted as definitive in rejecting this model.

Our results, while confirming the need for taking in consideration more realistic dispersal patterns when exploring correlations with genetic distances (Ramachandran et al., 2005), also call for more sophisticated simulations integrating appropriate modeling of the biological and cultural dynamics affecting genetics and linguistics as well as the processes shaping the spatial dispersal of BSPs, as all are necessary to explicitly test the support for the different scenarios proposed for the "Bantu expansion."

Our findings highlight how the interactions with inhabitants of the regions where they settled shaped the variation of several BSP populations, influencing their overall similarity and the extent of the correlation between their genetic variation, linguistic diversity and spatial distribution. Usually considered as a homogeneous group, the population structure of BSPs should instead be properly taken into consideration, in particular in biomedical and ancestry investigations.

We note that no corrections were attempted here for the linguistic data but it is reasonable to assume that languages too might have been affected by other dynamics during their evolution and dispersal. An important aspect to take in consideration for example is that often individuals speak more than one Bantu language and that therefore languages do not often operate as "barriers" to gene flow between groups. If so, a simple model of sequential splits followed by isolation might prove unrealistic and inappropriate to explore the relationship between languages and other variables, unless corrected for additional sources of variation. It is also worth mentioning that the dynamics of dispersal of genes and languages might be so different that their full reconciliation over a geographical model might prove complicated. In addition, different linguistic features might be tracking different time-depths as well as being more appropriate to explore intra or inter variation of linguistic families (Matsumae et al., 2021). The integration of archeological and linguistic data—and possibly the molecular analyses of ancient remains (Lipson et al., 2020)—with more complex demographic models is probably essential for a better understanding of the cultural and demic processes through which languages and people spread across Africa as part of the "Bantu expansion." These considerations apply to investigations focusing on similar events in other parts of the world (Creanza et al., 2015).

AUTHOR CONTRIBUTIONS

Miguel Gonzalez-Santos: Conceptualization (lead); data curation (lead); formal analysis (lead); writing – original draft (lead); writing – review and editing (lead). **Francesco Montinaro:** Formal analysis (supporting); writing – review and editing (supporting). **Rebecca Grollemund:** Formal analysis (supporting); writing – review and editing (supporting). **Davide Marnetto:** Formal analysis (supporting); writing – review and editing (supporting). **Masharip Atadzhanov:** Resources (supporting); writing – review and editing (supporting). **Celia A. May:** Resources (supporting); writing – review and editing (supporting). **Nedio Mabunda:** Resources (supporting); writing – review and editing (supporting). **Pierre de Maret:** Resources (supporting); writing – review and editing (supporting). **Ockie Oosthuizen:** Resources (supporting); writing – review and editing (supporting). **Erica Oosthuizen:** Resources (supporting); writing – review and editing (supporting). **Cristian Capelli:** Conceptualization (lead); data curation (lead); formal analysis (lead); funding acquisition (lead); writing – original draft (lead); writing – review and editing (lead).

ACKNOWLEDGMENTS

Miguel González-Santos, Francesco Montinaro and Cristian Capelli would like to thank St. Hugh's College for hosting them during part of the project; the High Performance Computing Facility of Oxford University for computational resources; and all the people who donated their DNA to make this study possible. Cristian Capelli was supported by the Leverhulme Trust Research Project Grant ("The genetic landscape of southern Africa human populations"; RPG-2013-298), the Wenner-Gren Foundation, the University of Oxford Boise Fund, and the John Fell Oxford University Press (OUP) Research Fund. Miguel González-Santos was supported by the Fundação para a Ciência e Tecnologia (SFRH/BD/90648/2012); Francesco Montinaro was supported by European Union through the European Regional Development Fund (project no. 2014-2020.4.01.16-0030); Davide Marnetto was supported by European Union through the European Regional Development Fund (project no. 2014-2020.4.01.16-0024, MOBTT53).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

New data here analysed is available at https://capelligroup.wordpress.com/data/. Information on previously published data is available in Table S1.

ORCID

Miguel González-Santos https://orcid.org/0000-0002-1489-3503

Cristian Capelli https://orcid.org/0000-0001-9348-9084

REFERENCES

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Barbieri, C., Butthof, A., Bostoen, K., & Pakendorf, B. (2013). Genetic perspectives on the origin of clicks in bantu languages from southwestern Zambia. *European Journal of Human Genetics*, *21*(4), 430–436.

Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S. W., Stoneking, M., & Pakendorf, B. (2013). Ancient substructure in early mtDNA lineages of southern Africa. *American Journal of Human Genetics*, *92*(2), 285–292.

Bastin, Y., Coupez, A., & de Halleux, B. (1983). Classification lexicostatistique des langues bantoues (214 relevés). *Academie Royale des Sciences d'outre-mer*, *27*(2), 173–199.

Bastin, Y., Coupez, A., & Mann, M. (1999). Continuity and divergence in the bantu languages: Perspectives from a lexicostatistic study. *Annales, Sciences Humaines*, *162*, 315–317.

Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M. F., … Villems, R. (2010). The genome-wide structure of the Jewish people. *Nature*, *466*(7303), 238–242.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521.

Bostoen, K., Clist, B., Doumenge, C., Grollemund, R., Hombert, J. M., Muluwa, J. K., & Maley, J. (2015). Middle to Late Holocene Paleoclimatic change and the early bantu expansion in the rain forests of Western Central Africa. *Current Anthropology*, *56*(3), 354–384.

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., Amenga-Etego, L. N., Enimil, A., Apinjoh, T., Ndila, C. M., Manjurano, A., Nyirongo, V., Doumba, O., Rockett, K. A., Kwiatkowski, D. P., Spencer, C. C. A., Malaria Genomic Epidemiology Network (2016). Admixture into and within sub-Saharan Africa. *eLife*, *5*, e15266.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., … Pepper, M. S. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, *8*, 2062.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., & Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1265–1272.

Currie, T. E., Meade, A., Guillon, M., & Mace, R. (2013). Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. *Proceedings of the Royal Society B*, *280*(1762), 20130695.

de Filippo, C., Bostoen, K., Stoneking, M., & Pakendorf, B. (2012). Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B*, *279*(1741), 3256–3263.

de Maret, P. (2013). Archaeologies of the bantu expansion. In P. Mitchell & P. J. Lane (Eds.), *The Oxford handbook of African archaeology* (pp. 319–328). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199569885.013.0043

González-Santos, M., Montinaro, F., Oosthuizen, O., Oosthuizen, E., Busby, G. B., 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 Biology and Evolution*, *7*(9), 2560–2568.

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. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(43), 13296–13301.

Heine, B. (1973). Zur genetischen Gliederung der Bantu-Sprachen. *Afr Uebersee*, *56*(3), 164–185.

Heine, B., Hoff, H., & Vossen, R. (1977). Neuere Ergebnisse zur Territorialgeschichte der Bantu. In W. J. G. Möhlig, F. Rottland, & B. Heine (Eds.), *Zur Sprachgeschichte und Ethnohistorie in Afrika* (pp. 57–72). Dietrich Reimer.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751.

Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(44), 17758–17764.

Henrici, A. (1973). Numerical classification of the Bantu languages. *Afrikaans Language Studies*, *14*, 82–104.

Holden, C. J. (2002). Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society B*, *269*(1493), 793–799.

Holden, C. J., & Gray, R. D. (2006). Rapid radiation borrowing and dialect continua in the Bantu languages. In P. Forster & C. Renfrew (Eds.), *Phylogenetic methods and the prehistory of languages* (pp. 19–31). McDonald Institute for Archaeological Research.

Holden, C. J., Meade, A., & Pagel, M. (2005). Comparison of maximum parsimony and Bayesian Bantu Language trees. In R. Mace, C. J. Holden, & S. Shannan (Eds.), *The evolution of cultural diversity: A phylogenetic approach* (pp. 53–65). University College London Press.

Johnston, H. (1886). *The Kilima-njaro expedition: A record of scientific exploration in eastern equatorial Africa, and a general description of the natural history, languages, and commerce of the Kilima-njaro district*. Kegan Paul & Trench.

Lavachery, P. (2001). The Holocene archaeological sequence of Shum Laka Rock Shelter (Grassfields, Western Cameroon). *African Archaeological Review*, *18*(4), 213–247.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*(1), e1002453.

Li, S., Schlebusch, C., & Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society B*, *281*(1793), 20141448.

Lipson, M., Ribot, I., Mallick, S., Rohland, N., Olalde, I., Adamski, N., Broomandkhoshbacht, N., Lawson, A. M., Lopez, S., Oppenheimer, J., Stewardson, K., Asombang, R. N., Bocherens, H., Bradman, N., Culleton, B. J., Cornelissen, E., Crevecoeur, I., de Maret, P., Fomine, F. L. M., … Reich, D. (2020). Ancient West African foragers in the context of African population history. *Nature*, *577*(7792), 665–670.

Liu, H., Prugnolle, F., Manica, A., & Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *American Journal of Human Genetics*, *79*(2), 230–237.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873.

Marks, S. J., Levy, H., Martinez-Cadenas, C., Montinaro, F., & Capelli, C. (2012). Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. *Molecular Ecology*, *21*(20), 4958–4969.

Marks, S. J., Montinaro, F., Levy, H., Brisighelli, F., Ferri, G., Bertoncini, S., Batini, C., Busby, G. B., Arthur, C., Mitchell, P., Stewart, B. A., Oosthuizen, O., Oosthuizen, E., D'Amato, M. E., Davison, S., Pascali, V., & Capelli, C. (2015). Static and moving frontiers: The genetic landscape of southern African Bantu-speaking populations. *Molecular Biology and Evolution*, *32*(1), 29–43.

Matsumae, H., Ranacher, P., Savage, P. E., Blasi, D. E., Currie, T. E., Koganebuchi, K., Nishida, N., Sato, T., Tanabe, H., Tajima, A., Brown, S., Stoneking, M., Shimizu, K. K., Oota, H., & Bickel, B. (2021). Exploring correlations in genetic and cultural variation across language families in Northeast Asia. *Science Advances*, *7*(34), eabd9223.

Mitchell, P. (2002). *The archaeology of Southern Africa*. Cambridge University Press.

Montinaro, F., Busby, G. B., González-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V. L., & Capelli, C. (2017). Complex ancient genetic structure and cultural transitions in southern African populations. *Genetics*, *205*(1), 303–316.

Nychka, D., Furrer, R., Paige, J., & Sain, S. (2015). *Fields: Tools for spatial data*. http://www.image.ucar.edu/fields

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2017). *Vegan: Community ecology package. R package version* 2.4–3. https://CRAN.R-project.org/package=vegan

Ongaro, L., Scliar, M. O., Flores, R., Raveane, A., Marnetto, D., Sarno, S., Gnecchi-Ruscone, G. A., Alarcon-Riquelme, M. E., Patin, E., Wangkumhang, P., Hellenthal, G., Gonzalez-Santos, M., King, R. J., Kouvatsi, A., Balanovsky, O., Balanovska, E., Atramentova, L., Turdikulova, S., Mastana, S., … Montinaro, F. (2019). The genomic impact of European colonization of the Americas. *Current Biology*, *29*(23), 3974–3986.

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Romero, I. G., Ayub, Q., Mehdi, S. Q., Thomas, M. G., Luiselli, D., Bekele, E., Bradman, N., Balding, D. J., & Tyler-Smith, C. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian Gene Pool. *American Journal of Human Genetics*, *91*(1), 83–96.

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Van der Veen, L., Hombert, J. M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., & Quintana-Murci, L. (2009). Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics*, *5*(4), e1000448.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J. M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*(6337), 543–546.

Patin, E., Siddle, K. J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Regnault, B., Lemee, L., Gravel, S., Hombert, J. M., Van der Veen, L., Dominy, N. J., Perry, G. H., Barreiro, L. B., Verdu, P., Heyer, E., & Quintana-Murci, L. (2014). The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications*, *5*, 3163.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093.

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the mantel test. *Oecologia*, *129*(2), 169–178.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Guldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P.-R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., … Pakendorf, B. (2012). The genetic prehistory of southern Africa. *Nature Communications*, *3*, 1143.

Pickrell, J. K., Patterson, N., Loh, P. R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., & Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(7), 2632–2637.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967.

Prugnolle, F., Manica, A., & Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. *Current Biology*, *15*(5), R159–R160.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15942–15947.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494.

Rexová, K., Bastin, Y., & Frynta, D. (2006). Cladistic analysis of Bantu languages: A new tree based on combined lexical and grammatical data. *Naturwissenschaften*, *93*(4), 189–194.

Rocha, J., & Fehn, A.-M. (2016). Genetics and demographic history of the Bantu. In *eLS*, (pp. 1–9). John Wiley & Sons, Ltd.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.

Seidensticker, D., Hubau, W., Verschuren, D., Fortes-Lima, C., de Maret, P., Schlebusch, C. M., & Bostoen, K. (2021). Population collapse in Congo rainforest from 400 CE urges reassessment of the bantu expansion. *Science Advances*, *7*(7), eabd8352.

Semo, A., Gaya-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., Prista, A., Damasceno, A., Fehn, A.-M., Schlebusch, C., & Rocha, J. (2020). Along the Indian Ocean coast: Genomic variation in Mozambique provides new insights into the bantu expansion. *Molecular Biology and Evolution*, *37*(2), 406–416.

Simons, G. F., & Fennig, C. D. (2018). *Ethnologue: Languages of the world, twenty-first edition*. SIL International http://www.ethnologue.com

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*(12), 1540–1542.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), 1035–1044.

Vansina, J. (1979). Bantu in the crystal ball, I. *History in Africa*, *6*, 287–333.

Vansina, J. (1990). *Paths in the rainforests: Toward a history of political tradition in equatorial Africa*. The University of Wisconsin Press.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Wotzka, H. P. (2006). Records of activity: Radiocarbon and the structure of iron age settlement in Central Africa. In H. P. Wotzka (Ed.), *Grundlegunge: Beiträge zur europäischen und afrikanischen Archäologie für Manfred K H Eggert* (pp. 271–289). Francke.

Wright, S. (1949). The Genetical structure of populations. *Annals of Human Genetics*, *15*(1), 323–354.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: González-Santos, M., Montinaro, F., Grollemund, R., Marnetto, D., Atadzhanov, M., May, C. A., Mabunda, N., Maret, P. d., Oosthuizen, O., Oosthuizen, E., & Capelli, C. (2022). Exploring the relationships between genetic, linguistic and geographic distances in Bantu-speaking populations. *American Journal of Biological Anthropology*, 1–14. https://doi.org/10.1002/ajpa.24589