Bringing together linguistic and genetic evidence to test the Bantu expansion

Cesare de Filippo¹,²,* Koen Bostoen³–⁵, Mark Stoneking²
and Brigitte Pakendorf¹,†

¹Max Planck Research Group on Comparative Population Linguistics, and ²Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany
³Ghent University, KONGOKING Research Group, Rozier 44, 9000 Ghent, Belgium
⁴Université libre de Bruxelles, 50 avenue F. D. Roosevelt, 1050 Brussels, Belgium
⁵Royal Museum for Central Africa, Leuvensesteenweg 13, 3080 Tervuren, Belgium

The expansion of Bantu languages represents one of the most momentous events in the history of Africa. While it is well accepted that Bantu languages spread from their homeland (Cameroon/Nigeria) approximately 5000 years ago (ya), there is no consensus about the timing and geographical routes underlying this expansion. Two main models of Bantu expansion have been suggested: The ‘early-split’ model claims that the most recent ancestor of Eastern languages expanded north of the rainforest towards the Great Lakes region approximately 4000 ya, while the ‘late-split’ model proposes that Eastern languages diversified from Western languages south of the rainforest approximately 2000 ya. Furthermore, it is unclear whether the language dispersal was coupled with the movement of people, raising the question of language shift versus demic diffusion. We use a novel approach taking into account the spatial and temporal predictions of the two models and formally test these predictions with linguistic and genetic data. Our results show evidence for a demic diffusion in the genetic data, which is confirmed by the correlations between genetic and linguistic distances. While there is little support for the early-split model, the late-split model shows a relatively good fit to the data. Our analyses demonstrate that subsequent contact among languages/populations strongly affected the signal of the initial migration via isolation by distance.

Keywords: human migration; mtDNA; Y chromosome; autosomes; lexical data; Bantu

1. INTRODUCTION

Bantu represents the largest African language family in terms of number of languages (approx. 500), occupied territory (approx. 9 million km²) and number of speakers (approx. 240 million). Bantu languages are generally thought to have originated approximately 5000 years ago (ya) in the Cameroonian Grassfields area neighbouring Nigeria, and started to spread, possibly together with agricultural technologies [1], through Sub-Saharan Africa as far as Kenya in the east and the Cape in the south [2]. The evolution of the Bantu languages has been suggested as fitting a branching-tree model, but this model does not entirely explain the variation observed in the linguistic data [3]. In addition, the modality of spread of Bantu-speaking peoples across Sub-Saharan Africa has been under debate in various disciplines during the last few decades. Contrasting models for the migration of Bantu-speaking groups have been proposed, sometimes even from studies using the same lexicostatistical dataset (cf. [4,5]). The major debate concerns the spatial and temporal dispersal of Bantu languages in Sub-Saharan Africa. One hypothesis states that Bantu languages split at an early stage (approx. 4000 ya) north of the rainforest, from which the Western and Eastern Bantu languages are derived as two primary branches [4,6–9]. A contrasting hypothesis argues that there was a major migration to the south of the rainforest, with a later split of the Eastern Bantu languages from the Western group only approximately 2000 ya [5,10–12]. We here call the former the ‘early-split’ and the latter the ‘late-split’ model (figure 1). For a detailed overview of the historical linguistic and molecular anthropological insights into the Bantu expansions, see [13].

The expansion of Bantu-speaking peoples is not only debated with regard to the underlying events and the route(s) taken, but also whether the spread of the languages took place as the result of ‘demic diffusion’ via an actual movement of people [1], or whether it was rather a cultural diffusion involving the movement of languages via ‘language shift’ [14] without concomitant gene flow. Genetic studies have highlighted the strong demographic impact of the Bantu migration on the gene pool of Sub-Saharan African populations for mtDNA [15,16], the Y chromosome [17,18] and autosomes [19,20]. However, a recent study by Sikora et al. [21] of approximately 2800 autosomal single nucleotide polymorphisms suggested that the spread of Bantu languages to southeastern Africa (i.e. Mozambique) involved language shift rather than a movement of people, thus revitalizing the debate concerning demic diffusion versus language shift.

In this study, genetic markers and linguistic data are used together for the first time to shed light on the
2. MATERIAL AND METHODS

(a) Genetic data and analyses

We compiled genetic datasets from published data for African populations with a sample size of at least 10 chromosomes. The mtDNA dataset consisted of 5018 sequences of the first hypervariable segment, the Y chromosome dataset contained 5018 sequences of the first Y hypervariable segment, the Y chromosome dataset contained 1248 sequences of the first Y hypervariable segment, the Y chromosome dataset contained 848 sequences of the first Y hypervariable segment, and included a total of 7200 trees, considering the Bantu-speaking groups when testing the models of Bantu migrations. The linguistic dataset used in this study consists of lists of 92 basic words for 412 Bantu languages (figure 2a; electronic supplementary material, table S4), taken largely from Bastin et al. [9]. We calculated linguistic distances among languages as patristic distances from the trees generated with Bayesian phylogenetic methods in order to take into account heterogeneity of replacement rates across words and rate changes of the same word over time [25]. Specifically, BAYESPHYLGENIES v. 1.1 [26] was used to generate a total of 7200 trees, considering a two-state character covarion model, which has the highest likelihood (electronic supplementary material, figure S1). Patristic distances were calculated for each tree and for the consensus tree reflecting the median of the branch-length distributions by means of the Python library DENDROPy [27].

The closest related language/dialect was chosen as corresponding to the genetic data for those populations whose language was not present in the linguistic dataset, resulting in both linguistic and genetic data for 36 populations for.

(b) Linguistic data and analyses

Figure 1. The two main models of Bantu migrations. The current area occupied by Bantu languages is shaded in grey, and the extent of the rainforest is indicated by the darker shading. (a) Early-split and (b) late-split.

Table 1. Number of populations for each ethnolinguistic group and genetic marker considered.

<table>
<thead>
<tr>
<th>groups</th>
<th>autosomes</th>
<th>mtDNA</th>
<th>Y chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bantu</td>
<td>33</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Niger-Congo</td>
<td>17</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Afro-Asiatic</td>
<td>28</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Nilo-Saharan</td>
<td>28</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Khoisan</td>
<td>4</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Pygmy</td>
<td>5</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

dispersal of Bantu languages and peoples in Sub-Saharan Africa. The study addresses two major questions. (i) We examine whether the expansion of Bantu languages was a cultural dispersal (i.e. language shift) or a joint movement of languages and people (i.e. demic diffusion) by using data from mitochondrial DNA (mtDNA), Y-chromosomal and autosomal markers. The demic diffusion model predicts that the genetic distances among populations speaking Bantu languages should be lower than those between Bantu-speaking populations and populations speaking other languages, whereas the cultural dispersal model predicts no consistent differences in the genetic distances among Bantu-speaking populations versus between Bantu-speaking populations and those speaking other languages. (ii) We test the two most commonly cited models for the Bantu expansion: the early-split versus the late-split model. For this purpose, we use an alternative method to previous linguistic studies that reconstructed models of expansion only from trees (or networks) of languages based on a quantitative comparison of Bantu lexical cognates [4,5]. Our method uses the migration distances between populations predicted by the early-split and late-split models, and correlates these distances with both linguistic and genetic distances. A third model of expansion (isolation by distance, IBD) is included as an alternative explanation to linguistic and genetic data for 36 populations for.

The linguistic dataset used in this study consists of lists of 92 basic words for 412 Bantu languages (figure 2a; electronic supplementary material, table S4), taken largely from Bastin et al. [9]. We calculated linguistic distances among languages as patristic distances from the trees generated with Bayesian phylogenetic methods in order to take into account heterogeneity of replacement rates across words and rate changes of the same word over time [25]. Specifically, BAYESPHYLGENIES v. 1.1 [26] was used to generate a total of 7200 trees, considering a two-state character covarion model, which has the highest likelihood (electronic supplementary material, figure S1). Patristic distances were calculated for each tree and for the consensus tree reflecting the median of the branch-length distributions by means of the Python library DENDROPy [27].

The closest related language/dialect was chosen as corresponding to the genetic data for those populations whose language was not present in the linguistic dataset, resulting in both linguistic and genetic data for 36 populations for.

Figure 1. The two main models of Bantu migrations. The current area occupied by Bantu languages is shaded in grey, and the extent of the rainforest is indicated by the darker shading. (a) Early-split and (b) late-split.

Table 1. Number of populations for each ethnolinguistic group and genetic marker considered.

<table>
<thead>
<tr>
<th>groups</th>
<th>autosomes</th>
<th>mtDNA</th>
<th>Y chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bantu</td>
<td>33</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Niger-Congo</td>
<td>17</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Afro-Asiatic</td>
<td>28</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Nilo-Saharan</td>
<td>28</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Khoisan</td>
<td>4</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Pygmy</td>
<td>5</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>
mtDNA, 34 populations for the Y chromosome and 21 populations for the autosomes (figure 2b; electronic supplementary material, table S5). See the electronic supplementary material for more details and explanations.

(c) Model-based geographical distances
In order to calculate the model-based geographical distances between Bantu-speaking populations, we grouped the Bantu languages and populations according to the Bantu classification proposed by Vansina [28], with some modifications. Languages were assigned to 13 different groups (electronic supplementary material, figure S2): Bantoid, Boan-Lebonya, Buneya and North-West representing the 'North' meta-group; East-Central, East-Coastal, East-Great Lakes, East-Kilimanjaro, East-South and East-Southeast representing the 'East' meta-group; West-Coastal, West-Congo Basin and West-Southwest representing the 'West' meta-group. The 'East' and the 'West' meta-groups are more closely related to each other than to the 'North' meta-group [7,11,28]. Based on this classification the geographical distances were calculated as predicted from both early-split and late-split models (see electronic supplementary material, figure S3). Geographical distances based on the IBD model were calculated as great-circle distances between two populations/languages.

(d) Data analysis
Mantel tests (Z) based on Spearman rank correlations were carried out between matrices of linguistic distances and matrices of geographical distances, in order to generate a distribution of correlation coefficients between the linguistic trees and the models of Bantu migrations. Differences between distributions were tested by a two-tailed Mann–Whitney U (MWU) test. Mantel tests with 10 000 permutations were carried out between matrices of linguistic distances (in this case based on the consensus tree; see electronic supplementary material), geographical distances generated according to the early-split, late-split and IBD models, and genetic distances. Spearman rank correlations (ρ) were calculated for the filtered distances (using only the informative pairwise comparisons, as described in the electronic supplementary material), given that these were not matrix-like anymore. In addition, to check for the effect that one model might have on the correlation observed between another model and the lexical distances, partial Mantel tests (Zp) based on Spearman rank correlations with 10 000 permutations were carried out. Unless stated differently, all p-values were corrected for multiple test hypotheses [29].

3. RESULTS
(a) Demic diffusion versus language shift
After a demic diffusion, the genetic differences among populations speaking the same or related language(s) are expected to be lower than those among individuals or populations speaking different or only distantly related language(s). Conversely, with cultural dispersal and language shift, the genetic differences among populations speaking the dispersing language(s) are not expected to be lower than genetic differences between such populations and geographically proximate populations speaking different languages. With this expectation, we
tested whether the expansion of Bantu languages was accompanied by an actual movement of people. Figure 3 shows the distributions of genetic distances among Bantu populations alone, and among Bantu versus any of the other major African linguistic phyla and Pygmies. The distributions of genetic distances among all Bantu populations are significantly lower than those between Bantu and any of the other major linguistic phyla for all genetic markers (all MWU tests with one tail \( p < 10^{-3} \)), consistent with a demic diffusion. Different patterns were observed when this approach was applied to the other linguistic phyla for all markers, with some exceptions (see electronic supplementary material, figures S4–S6).

(b) Patterns of genetic diversity as distance from the Bantu homeland

When population diversity levels were calculated without taking into account heterogeneity in population size, no significant trends for all genetic markers were observed as a function of distance from the Bantu homeland. In contrast, significant trends for all genetic markers were observed as a function of distance from the Bantu homeland. Taking into account heterogeneity in population size, no significant trend for any genetic markers was observed as a function of distance from the Bantu homeland.

(c) Languages and models of migrations

Three different matrices of great circle geographical distances among languages were calculated to represent the three models of the Bantu migration investigated here, namely the early-split, late-split and the simple IBD model. Figure 5 shows an example of geographical distances between an Eastern and a Western Bantu language calculated according to the three models. Overall there was a high correlation among the models (mean \( Z = 0.70, \ all \ p < 10^{-3} \); see electronic supplementary material, table S6) because the early-split and late-split models do not have different predictions for 43 per cent of the pairwise comparisons (see electronic supplementary material, figure S3c). Although it is difficult to avoid collinearity among models, with ‘informative’ distances, which show large differences between the early-split and late-split models (such as those shown in figure 5), the correlation between these two models was significantly reduced \( (p = 0.31, \ p < 0.001; \ electronic \ supplementary \ material, \ table S6).\)

To test which model, if any, best fits the linguistic differences observed among 412 Bantu languages, we performed correlations between 7200 matrices of linguistic distances and three matrices of geographical distances as predicted by the three models. The results (table 2; electronic supplementary material, figure S7) indicate that whereas all models are significantly correlated with the linguistic distances, the IBD model consistently
Table 2. Correlations between the distribution of linguistic distances and models of Bantu migration using all (Mantel test $Z$) and only ‘informative’ (Spearman correlation $\rho$) pairwise comparisons. All pairwise comparisons of correlation distributions (electronic supplementary material, figure S6) are significantly different (all MWU tests with $p < 10^{-16}$).

<table>
<thead>
<tr>
<th>correlation</th>
<th>IBD</th>
<th>late-split</th>
<th>early-split</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$ (90% CI)</td>
<td>0.602 (0.550–0.658)</td>
<td>0.510 (0.446–0.618)</td>
<td>0.492 (0.425–0.569)</td>
</tr>
<tr>
<td>$\rho$ (90% CI)</td>
<td>0.603 (0.499–0.656)</td>
<td>0.436 (0.371–0.521)</td>
<td>0.303 (0.186–0.394)</td>
</tr>
</tbody>
</table>

Figure 6. Correlations (Mantel tests $Z$) of genetic distances with the geographical distances predicted by the migration models (IBD in black, early-split in grey, late-split in light grey) and with the linguistic distances (in white). The bars of each correlation coefficient represent the 95% CIs generated with 1000 bootstraps. On top of the bar, the asterisk indicates $p$-values less than 0.05 after correction for multiple tests.

Table 3. Partial Mantel test between lexical and genetic distances by keeping constant one model of Bantu migrations.

<table>
<thead>
<tr>
<th>genetic distance</th>
<th>IBD</th>
<th>late-split</th>
<th>early-split</th>
</tr>
</thead>
<tbody>
<tr>
<td>mt-$\Phi_{ST}$</td>
<td>0.113</td>
<td>0.240**</td>
<td>0.418**</td>
</tr>
<tr>
<td>Y-$R_{ST}$</td>
<td>0.144</td>
<td>0.209</td>
<td>0.222**</td>
</tr>
<tr>
<td>Y-$F_{ST}$</td>
<td>0.207*</td>
<td>0.196*</td>
<td>0.272**</td>
</tr>
<tr>
<td>A-$R_{ST}$</td>
<td>0.003</td>
<td>0.041</td>
<td>0.183</td>
</tr>
</tbody>
</table>

*p < 0.01 without correction for multiple tests.
**p < 0.01 with correction for multiple tests.

(d) Genes, languages and models of Bantu migration

Mantel tests were also carried out between the geographical distances predicted by the three models of Bantu expansion and genetic distances. The results (figure 6) indicate that mt-$\Phi_{ST}$ and A-$R_{ST}$ distances best correlate with the IBD model, while Y-$R_{ST}$ and Y-$F_{ST}$ distances best correlate with the late-split model. The early-split model exhibits the lowest correlation for all genetic markers—similar to the pattern observed for the linguistic data—and no significant correlation for Y-chromosomal distances.

All three genetic markers show significant correlations with the linguistic distances (figure 6), with the highest correlation for mt-$\Phi_{ST}$ ($Z = 0.54$), intermediate for A-$R_{ST}$ ($Z = 0.37$), and lowest for both Y-$F_{ST}$ and Y-$R_{ST}$ ($Z = 0.27$ and 0.25, respectively). These correlations of ‘genes’ with languages as well as with geographical distances further support the hypothesis of a demic diffusion of Bantu-speaking people.

Table 3 reports all partial Mantel tests that were carried out between linguistic and genetic distances by keeping constant as a control a matrix of the geographical distances predicted by a particular model of Bantu migration. Both mtDNA and the Y-chromosomal distances are significantly
correlated with linguistic distances, in contrast to the autosomal distances, which are not correlated at all with linguistic distances, irrespective of the model of migration.

4. DISCUSSION

(a) Demic diffusion of Bantu-speaking peoples

In this study, we consider for the first time both haploid and diploid markers to test whether the dispersal of Bantu languages in Sub-Saharan Africa was coupled with the movement of people (demic diffusion) or whether it was the result of language shift. Our comparison of the genetic distances among Bantu populations with those of Bantu versus all other linguistic and ethnic groups (figure 3) indicates that even geographically distant Bantu-speaking populations are closely related to each other, as expected with demic diffusion, and argues against a major role for language shift in the Bantu expansion.

The other African ethnolinguistic groups, on the other hand, do not show a similar degree of genetic proximity between populations regardless of the geographical distance separating them (see electronic supplementary material, figures S4–S6). This probably reflects complex demographic histories represented by various events of demic diffusion, language shift and/or language/population contact with other groups. Given that these represent much older phylogenetic units than the relatively young Bantu language family, there has been more time for such demographic events to obscure signals of relationship in these phyla than in the Bantu family.

In addition, the positive correlations ascertained between linguistic and genetic (as well as geographical) distances support the view that the Bantu expansion was a combined movement of languages and genes through Sub-Saharan Africa. Such associations among genetic, linguistic and geographical distances were also found at a finer geographical scale among Austronesian-speaking populations in the eastern Indonesian island of Sumba [30] and in Northern Island Melanesia [31]. Our results thus confirm the general view of anthropological genetics that the Bantu expansion took place via a demic diffusion [1,15,18–20], although they contrast with the results of Sikora et al. [21]. However, this discrepancy could be explained by the fact that Sikora et al. used a very different dataset (only one Eastern and two Southeastern Bantu populations) than we do here.

Further support for a demic diffusion is found in the significant reduction of genetic diversity that is proportional to the distance from the Bantu homeland (figure 4). These reductions of diversity levels were stronger for mtDNA than for the Y chromosome, suggesting a different history of men and women during the Bantu migrations. The lack of reduction in autosomal diversity probably reflects the severe lack of Bantu-speaking populations in this dataset, especially in south-central Africa (cf. figure 2b). Interestingly, the stronger reduction in mtDNA diversity contradicts the idea that the patrilocal Bantu-speaking peoples would have incorporated local women during their migrations (cf. [13]). However, given that mtDNA and the Y chromosome are basically single loci with different mutation rates it is difficult to disentangle the effects of stochasticity from demographic events. In addition, the sparse sampling of Bantu populations in general, and the mismatch between the availability of mtDNA versus Y-chromosomal data in particular (figure 2), precludes any definitive conclusions about the nature of the Bantu migrations. More research is thus needed to disentangle the complex history of the Bantu expansion.

(b) Model of expansion and contact among Bantu

The spatial and temporal dynamics of the Bantu expansion have also been debated. Based on phylogenetic trees constructed from lexical data, different theories have been suggested for the dispersal of Bantu languages through Sub-Saharan Africa (cf. [4,5]). These theories can be summarized via two core models that we called here the early-split and late-split models (figure 1). Contrary to previous studies, we also include in our analyses a simple IBD model, which is expected to hold if there is ongoing contact between populations that is mediated by the geographical distance between them. Our results clearly indicate that the IBD model best explains the observed pattern of linguistic differences, while the early-split model shows the worst fit (table 2 and electronic supplementary material, figure S7). However, both the late-split and early-split models correlate with linguistic distances even after controlling for the effect of IBD. Nevertheless, in all of our analyses the early-split model consistently receives the least support, as also found by Alves et al. [32]. Furthermore, there is much variation in the linguistic and genetic distances that is not explained when including the early-split distances in the analyses (table 3), and the early-split model shows no significant correlation with the Y-chromosomal data. However, these results should be viewed with caution given that there is controversy on the validity of partial Mantel tests [33,34].

This apparent difference among genetic markers does not reflect differences in the populations sampled for each type of marker. Indeed, when using only the same languages as the populations considered for the Y chromosome, the correlation between linguistic distances and migration models (electronic supplementary material, table S7) showed the same pattern as for the entire linguistic dataset (table 2). In addition, when using only the 21 populations for which both mtDNA and Y-chromosomal data are available, the highest correlations were again with the IBD and late-split model, respectively (see electronic supplementary material, table S8).

Interestingly, there were no significant correlations between mt-D_{mt} and either Y- F_{RST} (Z = 0.13, p = 0.15) or Y-R_{ST} (Z = 0.01, p = 0.42). This could indicate that Bantu-speaking women have a different history than Bantu-speaking men, as also suggested by the different correlations between the genetic markers and the migration models, and in accordance with a previous study [35] that found a higher effective population size and migration rate of females than males in Bantu-speaking populations.

The finding that the linguistic and genetic data best fit the IBD model might indicate that the diffusion of the Bantu languages was a gradual process of dialectal diversification, with changes taking place at a small scale and spreading locally through contact. However, this scenario of a slow and gradual process of diversification over short distances does not fit well with the evidence for a demic
migrations apparent in the reduction of diversity values for the genetic markers, nor with the low genetic distance values among geographically distant Bantu groups. An alternative explanation could be that the signal of the initial migration(s) might have been attenuated by later contact among peoples and languages, as suggested by previous genetic studies of mtDNA \([35,36]\), Y-chromosomal \([18]\) and autosomal \([32]\) markers. This hypothesis finds some support in our data, since the linguistic distances between Eastern and Western languages finds some support in our data, since the linguistic distances belonging to adjacent Guthrie’s zones \((such \ as \ K \ versus \ M)\) are significantly smaller compared with those of non-adjacent zones \((such \ as \ H-R \ versus \ G-P, \ K \ versus \ M)\).

Since lexical data might be very susceptible to contact pressure, while structural features are considered to be more stable \((cf. \ summary \ in \ [37])\) \([38]\), we applied our method to eight structural features reported for 66 Bantu languages \([38]\). However, we do not find any correlation between the structural linguistic distances and the geographical distances predicted by any model of migration. This probably reflects the paucity of structural data available for the Bantu languages together with the absence of variation in such data \([38]\). Therefore, more informative structural data, comparable with those collected by Reesink et al. \([39]\), would be needed to investigate the complex expansion history of the Bantu languages.

### 5. CONCLUSIONS

Our analyses primarily indicate that the dispersal of Bantu languages was coupled with the movement of people \((i.e. \ demic \ diffusion)\), as demonstrated by the lower genetic distances among Bantu populations when compared with those between Bantu and all the other major ethnolinguistic groups, as well as by the reduction of mtDNA and Y-chromosomal diversity proportional to the distance from the Bantu homeland. Furthermore, we found strong correlations between genetic and linguistic \((as \ well \ as \ geographical) \ distances \ as \ additional \ evidence \ of \ a \ demic \ diffusion.\) With regard to the geographical routes underlying the Bantu expansion, our analyses indicate that the early-split model, which postulates an initial split into Eastern and Western Bantu languages approximately 4000 ya, finds little support in the lexical and genetic data. In contrast, the late-split model of the Bantu expansion, which proposes a relatively recent development of the Eastern Bantu languages out of Western Bantu, fits the data better. However, as evidenced by the strong correlation of the IBD model with both the lexical and the genetic data, it is clear that subsequent contact between Bantu languages \([2,3,7,40]\) and populations \([18]\) has strongly affected the signal of the initial migration. Furthermore, the low coverage of Bantu-speaking populations—especially from eastern, central and southern Africa—makes the conclusions tenuous; it would therefore be highly desirable to include genetic data from more Bantu-speaking populations to obtain more firmly supported insights into the demographic processes at play in this historical event.

We thank Roger Mundry, Michael Dannemann, Hernan A. Burbano and Sean Myles for statistical/programming advice and comments on the manuscript, and Yvonne Bastin, Michael Mann, Chiara Barbieri and Nina Kottenhagen for sharing linguistic and genetic data. We are also grateful to two anonymous reviewers, whose comments helped improve the manuscript. This work was funded by the Max Planck Society (C.d.F., M.S. and B.P.) and the Belgian Science Policy (K.B.).

### REFERENCES

11. Heine, B., Hoff, H. & Vossen, R. 1977 Neuere Ergebnisse zur Territorialgeschichte der Bantu. In Zur Sprachgeschichte und Ethnomierie in Afrika \((eds \ W. \ J. \ Möhlig, \ F. \ Rottland \ & \ B. \ Heine)\), pp. 52–72. Berlin, Germany: Dietrich Reimer.


22 Klieman, K. A. 2003 ‘The Pygmies were our compass’: Bantu and Batwa in the history of west central Africa, early times to c. 1900 CE. Portsmouth, NH: Heinemann.


