WILEY-BLACKWELL

Cladistics (2018) 1–20

Cladistics

10.1111/cla.12353

Revising the Bantu tree

Peter M. Whiteley[a], Ming Xue[a] and Ward C. Wheeler[b,*]

[a]Division of Anthropology, American Museum of Natural History, 200 Central Park West, New York, NY, 10024-5192, USA; [b]Division of Invertebrate Zoology, American Museum of Natural History, 200 Central Park West, New York, NY, 10024-5192, USA

Accepted 15 June 2018

Abstract

Phylogenetic methods offer a promising advance for the historical study of language and cultural relationships. Applications to date, however, have been hampered by traditional approaches dependent on unfalsifiable authority statements: in this regard, historical linguistics remains in a similar position to evolutionary biology prior to the cladistic revolution. Influential phylogenetic studies of Bantu languages over the last two decades, which provide the foundation for multiple analyses of Bantu sociocultural histories, are a major case in point. Comparative analyses of basic lexica, instead of directly treating written words, use only numerical symbols that express non-replicable authority opinion about underlying relationships. Building on a previous study of Uto-Aztecan, here we analyse Bantu language relationships with methods deriving from DNA sequence optimization algorithms, treating basic vocabulary as sequences of sounds. This yields finer-grained results that indicate major revisions to the Bantu tree, and enables more robust inferences about the history of Bantu language expansion and/or migration throughout sub-Saharan Africa. "Early-split" versus "late-split" hypotheses for East and West Bantu are tested, and overall results are compared to trees based on numerical reductions of vocabulary data. Reconstruction of language histories is more empirically based and robust than with previous methods.

Introduction: a problem in historical linguistics

The expansion of Bantu languages and peoples throughout sub-Saharan Africa has been studied extensively (Vansina, 1995; Bastin et al., 1999; Hombert and Hyman, 1999; Ehret, 2001; Nurse and Phillipson, 2003a; Holden and Gray, 2006; Rexová et al., 2006; Pakendorf et al., 2011; de Filippo et al., 2012; Currie et al., 2013; Bostoen et al., 2015; Grollemund et al., 2015; de Luna, 2016). From a homeland generally postulated near the Nigeria–Cameroon border, ancestral Bantu languages are deemed to have spread into central, west-central, eastern and southern Africa over perhaps five millennia (Fig. 1). Estimates of extant Bantu languages range from 300 to 680 (depending on definitions of language vs. dialect). Classificatory relationships, routes and times of ancestral spread, and placement within a more inclusive "Niger–Congo" group are much debated, as noted by Maho (2006, pp. 202–203):

> There is no accepted subclassification of the Bantu languages, at least none that can claim any historical validity. It has proven extremely difficult to establish a stable internal structure of subgroups based on shared innovations, regular sound changes, lexicostatistics, or whatever … In fact, the only currently used classifications are largely or entirely referential ones.

Suggested causes of spread alternatively highlight grain-crop domestication, technological revolution (iron smelting), slow growth of a farming culture and/or Holocene climate change, among others (Maho, 2006; Grollemund et al., 2015). Principal models (Fig. 2) emphasize either an "early split" or "late split". In the former, eastern and western Bantu languages diverged in the homeland: the eastern lineage moved along the 5°N parallel, north of the Congo rainforest, to reach Lake Albert ca. 4000 BP, with descendants expanding south into central and

*Corresponding author:
E-mail address: wheeler@amnh.org

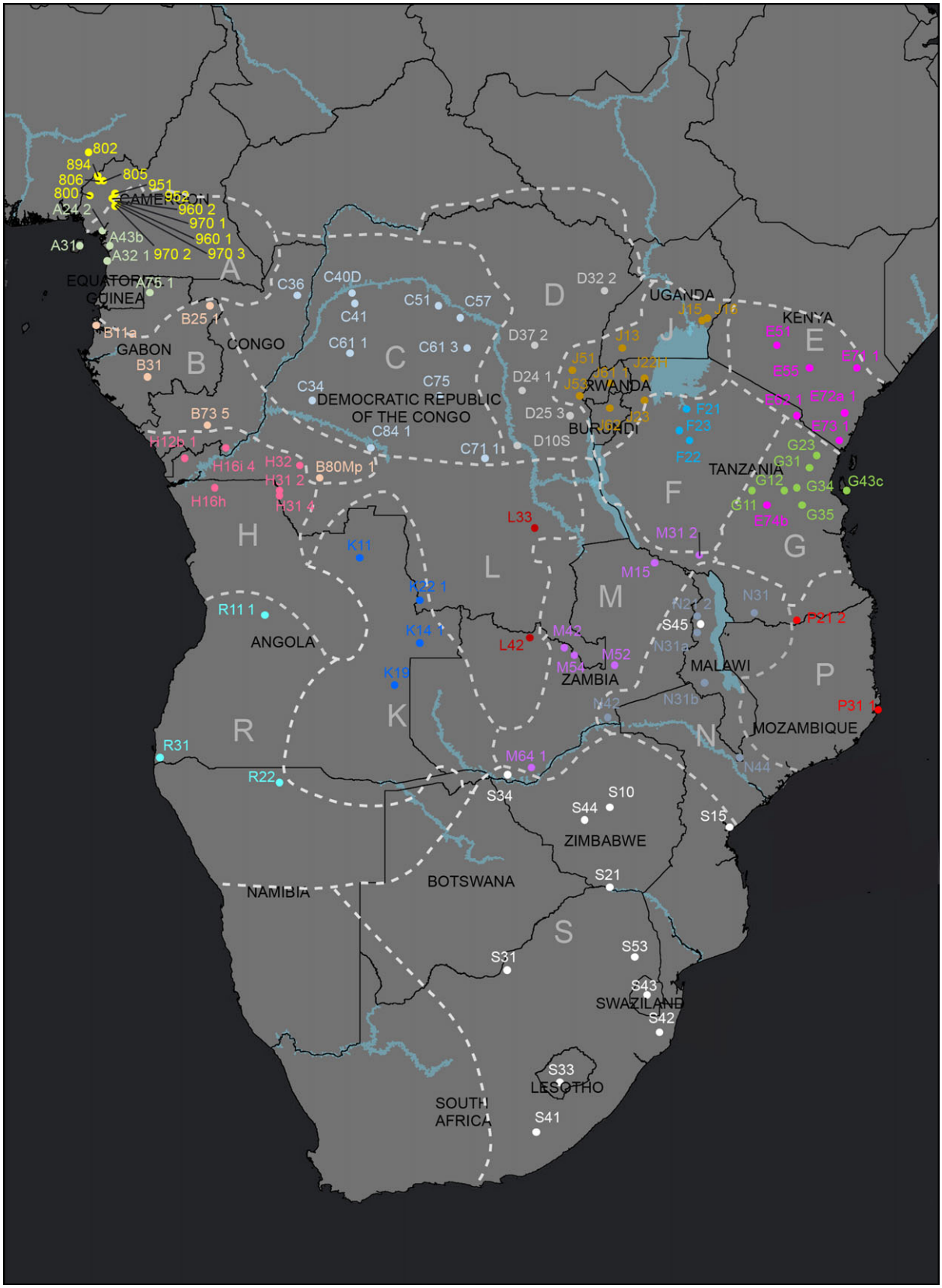

Fig. 1. Guthrie Zones (letter designation of Guthrie), locating 105 sample languages. Languages are coloured by Guthrie zone with the exception of the yellow out-group languages. Map by Thomas Blaber, Nicholas Triozzi, AMNH. Service layer credits: Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors and the GIS community.

southeast Africa; meanwhile, the western lineage moved south through the rainforest to the lower Congo Basin, with some descendants moving eastward into the upstream rainforest, and others into west-central and southwest Africa. According to the "late-split" model, the common ancestor of both east and west Bantu moved south through the rainforest to reach the lower Congo Basin ca. 2000 BP; thereafter, the ancestor of eastern groups moved to the upper Congo west of Lake Tanganyika, and from there, descendants radiated east into the interlacustrine zone, and south into central, east-coast and southeast Africa; meanwhile, the western ancestor spread from the lower Congo southward into west-central and southwest Africa (de Filippo et al., 2012).

The extent to which "East" and "West" Bantu represent monophyletic groups rather than geographical agglomerations is debatable. Moreover, suggested points and directions of diversification in late-split models vary extensively (Rexová et al., 2006; Currie et al., 2013; Grollemund et al., 2015), including possible marine routes (Blench, 2012). Whatever the validity of particular language-spread scenarios, as de Filippo et al. (2012, p. 3256), pointed out, it is "unclear whether the language dispersal was coupled with the movement of people, raising the question of language shift versus demic diffusion" (see also Bostoen et al., 2015; Patin et al., 2017). Whether the best schema for Bantu language evolution is conceived as trees or networks remains in question. Some (notably Vansina, 1995) argue for wave (i.e. horizontal) models, in which intra-regional neighbours serially borrow and re-borrow linguistic forms, obviating vertical transmission signals. Reticulate relationships are certainly important (see, e.g., Schadeberg, 2003; Bryant et al., 2005; Holden and Gray, 2006; Carlo and Good, 2015), but as developed to date, wave models are empirically unsupported and unnecessarily complex, as emphasized by Rexová et al. (2006). Phylogenetic signals for Bantu show clear indications of vertical transmission (e.g. Holden, 2002; Greenhill et al., 2009), so until there are adequate tests for vertical vs. horizontal histories in targeted cases, tree analysis continues to be the expected paradigm.

In short, Bantu language classification and historico-geographical spread remain unresolved, with multiple competing hypotheses. The main reasons for the lack of consensus lie in muddled methods of historical linguistics and resultant bowdlerizing of empirical data, as explained below. The aim here is for a decisive test of existing hypotheses, via direct computational analysis of lexical data as sound sequences, comparable to DNA sequences (as in Wheeler and Whiteley, 2015). The same core group of Bantu languages targeted by previous studies is addressed, but in contrast to prior methods, the underlying data, words themselves, rather than symbolic reductions of cognacy judgements, are the object of analysis. Existing scenarios and inferred migration hypotheses are tested: in particular, early-split vs. late-split models, whether "East" vs. "West" Bantu is a meaningful division, and whether West Bantu forms a monophyletic group with a unique common ancestor not shared with East Bantu. Results of the present analysis share elements with some prior models, but exhibit differences, with new implications for both linguistic relationships and historical expansion.

Language histories: methods and assumptions

Language phylogenies offer a sound baseline for inferring evolutionary patterns in other sociocultural phenomena (Mace and Pagel, 1994), but language phylogenies can only be as good—that is, retrodictively robust—as the suitability of methods and quality of data (Barbançon et al., 2013; List, 2016). Comparison of basic vocabulary remains methodologically central (Bowern and Evans, 2015). Lexical approaches assume the more words, both in form and in meaning, are shared between two languages, the closer their historical relationship (for Bantu in this context, see Marten, 2006). The Swadesh list of 100 words (Swadesh, 1971) that are cross-culturally least susceptible to horizontal replacement was devised for a now outmoded lexicostatistical approach, but remains a robust framework for inferring historical patterns among related languages (Starostin, 2009). For most language families (including Bantu), compiled vocabulary data tend to be far more extensive than other records (e.g. morphosyntax) used to reconstruct language histories. Bantu lexical data have been used independently, and/or combined and contrasted with grammatical data (Nurse and Phillipson, 2003b; Rexová et al., 2006; Dimmendaal, 2011). Standard approaches to lexical data use the comparative method (the central theoretical tool of historical linguistics) to identify sound shifts from presence–absence patterns among paired languages (e.g. Campbell and Poser, 2008). While observed sound patterns are obviously informative, converting these into exceptionless historical laws—a tendency since the Neogrammarians—has created much intractable argument that pre-empts or preconceives further empirical study (see, e.g., Durie and Ross, 1996). In practice, even if inferred sound-shift laws are stated explicitly (often they are not) encoding them into analysis tautologously predetermines results. Such inferences should be conclusions of argument, not premises.

Despite institutionalized claims to the contrary, the methods of historical linguistics are often vague,

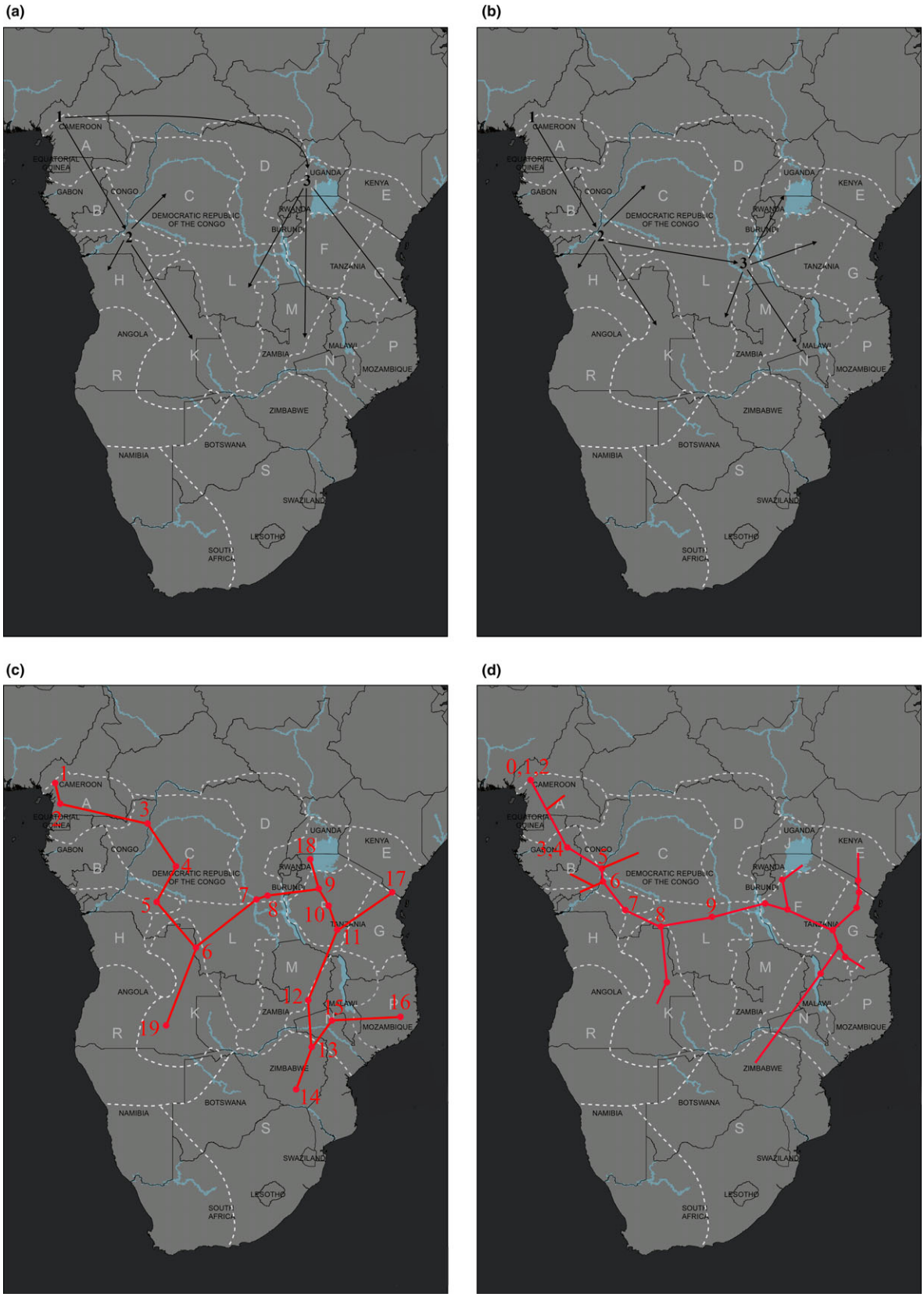

Fig. 2. Hypotheses of Bantu language expansion: (a) early split vs. (b) late split (after Pakendorf et al., 2011, fig. 2); (c) Currie et al. (2013, fig. 2b); (d) Grollemund et al. (2015, fig. 2a primary nodes and branches).

subjective and reliant on expert authority, as Greenberg (2005, p. 153) notably concluded:

> There exists in linguistics in general no coherent theory regarding the genetic classification of languages ... [Any notion that] historical linguistics has an utterly rigorous method, however slow, which reconstructs linguistic history step by step with complete precision is sheer myth.

Bluntly, if the field is to become a genuine science, this situation is unsustainable. Expert judgement is both invaluable and inevitable at certain levels of analysis. For basic vocabulary, expert judgement is indispensable for identifying similarity for both morphemes and meanings (the source data used herein are equally expert products at this level). At least potentially, such similarity judgements are independently verifiable, as representing speech recorded in the field. However, the next common step—establishing underlying cognacy, and reconstructing ancestral proto-forms—is typically more opaque. For example, among Indo-European languages, the gloss "tree" includes English *tree* (IPA/tɹi/), Gothic *triu* (/triu/) and Albanian *dru* (/dʁy/)—all cognates. Standard procedure, invoking historical sound shifts, proposes a common ancestral "root" or starred form: proposals here include /*dóru/, /*dreu-/ or /*derew(o)-/ (Ringe, 2006). The three observed (extant) words are considered "reflexes" or descendants of the proto-form, but here as elsewhere, alternative reconstructions proliferate, reflecting an intrinsic methodological problem: "There is no empirical way of disproving a reconstruction" (Greenberg, 1987, p. 10). Lacking falsifiability, cognacy and reconstruction judgements effectively depend on intuition, guesswork and arguments from authority. While recent automated methods for identifying cognates offer more neutrality, their potential for improving phylogenetic reconstructions of language relationships beyond those based on expert judgements remains debated (List et al., 2017; St. Arnaud et al., 2017; Rama et al., 2018).

A similarly problematic dependency on reconstructed prototypes, derided as "hapless appeals to plesiomorphy" (Rosen et al., 1981, p. 264), preoccupied evolutionary biology prior to the cladistic revolution. Historical analysis was predicated on retrodicting hypothetical proto-forms. Positing proto-languages ancestral to observed speech is the linguistic equivalent of such hapless plesiomorphy and its "futile paleontological searches for ancestors" (Rosen et al., 1981). Statements of relationship between purely hypothetical reconstructions and current languages are ipso facto not testable. A more open and rigorous method is demonstrated here (see also Wheeler and Whiteley, 2015). To explain how this approach differs, the historical and empirical basis of extant phylogenetic analyses of Bantu languages requires describing first.

Data background

Guthrie's classification (Guthrie, 1948, 1967–1971) divides Bantu languages into geographical "Zones" labelled A–S (Fig. 1), internal decimal series (A10, A20, B10, B20, etc.) and individual languages (A11, A23, B32, B44, etc.). Dialects are subdivided: A11a or A11 1, A11b or A11 2, etc., according to different notations. Guthrie's classification was standard before ISO 639-3 (e.g. Simons and Fennig, 2017) and Glottolog coding (Hammarström et al., 2017), and remains widely used (e.g. Maho, 2009). While acknowledging the classification was primarily geographical, Guthrie maintained (1967–1971, II, p. 16) it was also on linguistic grounds: "[the classification and] the zones themselves can scarcely be regarded as of no relevance to genealogical questions." This is firmly rejected by later scholars, however, who argue the classification "... and especially its zones, have little historical reality" (Nurse and Phillipson, 2003b, p. 168). Yet some identified phylogeographical groups, both in previous analyses and in the present one (below) correlate with Guthrie Zones, indicating that they do partly mark historical descent.

Lexicostatistics, developed at mid-century by Swadesh (1971), motivated a long-term Bantu project at the Musée royal de l'Afrique centrale/Royal Museum for Central Africa (RMCA), Tervuren, Belgium. Numerous word lists were collected, beginning in the 1950s. In the 1970s, RMCA reduced Swadesh's 100-word list to 92, removing "I", "you [sing.]", "we", "this", "that", "not", "green", "yellow", and substituting "arm" for "hand" and "leg" for "foot". By 1990 word lists were complete for 530 languages from all Guthrie Zones, and 12 Bantoid languages northwest of Zone A (Bastin et al., 1999). The listings were not entirely consistent (Bastin et al., 1983), and some do not include all 92 words, but many do, and the others have a great majority (for specific numbers, see Bastin et al., 1999). Rendered into a standard orthography, most word lists are posted on RMCA's "lexico" webpage (http://www.africamuseum.be/research/human-sciences/cultsoc/lexico-1/).

This cumulative data set is extraordinarily useful and has been analysed—in reduced form—by RMCA scholars (notably, Coupez et al., 1975; Bastin et al., 1983, 1999; Vansina, 1995) and others. The 542 word lists finalized in 1990, or subsets, provide the underlying basis for most phylogenetic analyses of Bantu languages—Bastin et al. (1999): all 542 languages; Holden (2002): 75 languages; Holden et al. (2005): 95 languages; Holden and Gray (2006): the same 95 languages; Rexová et al. (2006): 87 languages (adding grammatical data to the lexical data); Dunn et al. (2011): 75 languages (the same as Holden, 2002); de Filippo et al. (2012): 412 languages; and Currie et al.

(2013): all 542 languages. As noted above, these primarily linguistic analyses have been correlated with multiple cultural, archaeological, geographical, genetic and other features (Holden and Mace, 2003, 2005; Alves et al., 2007; Pagel, 2009; Walker and Hamilton, 2011; Opie et al., 2014; Bostoen et al., 2015; Guillon and Mace, 2016; Patin et al., 2017).

Apart from Bastin et al. (1999)—and then only at the outset, not in the computational analysis—no subsequent study utilized the vocabularies themselves, although that is not always made clear. For example, "the tree sample is generated from lexical data … [The underlying] phylogenetic trees [were] constructed using basic vocabulary data" (Dunn et al., 2011: p. 2, p. 3; citing Holden, 2002, as the data source). Such statements are not untrue, but occlude the fact that the empirical words had been replaced by numerical codes. This is transparent in the original statement of method by Bastin et al. (1999, p. 8):

> The work of cognation judgment was undertaken by André Coupez and Yvonne Bastin. For each gloss they recognized a number of roots to which they assigned numbers (root-codes), and built up in sections a table with a row for each vocabulary and a column for each gloss, entering in each cell the appropriate root-code or codes.

An ancillary aim here is therefore to explicate exactly what prominent phylogenetic analyses of Bantu languages are based on, and why the results of the present analysis differ. "Lexico.txt", the matrix of root-code numbers (Table 1), was sent from Tervuren by Bastin and Coupez to London for computational analysis by Mann, then at the School of Oriental and African Studies (University of London); Mann never received any word lists or reconstructed proto-forms corresponding to root-codes (M. Mann, personal communication, 2011).

Queried (avowedly, two decades after the fact), Mann indicated his belief that, "the starred form which underlies the claim of cognacy … is also on the Tervuren website, under the title Bantu Lexical Reconstructions" (M. Mann, personal communication, 2011). That turns out mostly not to be the case, however. RMCA's Bantu Lexical Reconstructions ("BLR"), a parallel project with "lexico", has appeared in three iterations: Meeussen (1969) [BLR1], Coupez et al. (1998) [BLR2] and Bastin et al. (2002) [BLR3].

Asked whether the roots or root-codes for lexico.txt were represented in BLR3, co-author Schadeberg (personal communication, 2016) pointed out:

> … only a relatively small part of the cognation sets represented by symbols in BCM1999 [Bastin et al., 1999] can be linked to BLR3 entries (nor in any older version such as BLR2). How come? Linguists may be convinced that E[nglish] fire/G[erman] Feuer/D[utch] vuur are cognate WITHOUT postulating a reconstruction.

Here are seen philosophical differences in linguistic reconstruction. Mann (personal communication, 2011) understood each root-code in the matrix to stand for an actual root or starred form, which is "a form in some ancestral language, or simply … a way of summing up a set of (fairly) regular sound-correspondences between words in related languages … [although] different linguists will take different views of the status of "starred" forms …" For Schadeberg (above), the root-codes represent convictions about cognation not actual reconstructions. The difference may seem trivial—roots/starred forms are based on cognate sets correlated to sound shifts—but it confirms the critique (above) of linguistic plesiomorphy. Following Schadeberg, the root-codes in effect represent amorphous inferences: summary expressions of working hypotheses based on expert assumptions that are left unstated.

As a practical matter, Mann's understanding of Coupez and Bastin's procedure must be more nearly correct. Across the 542 vocabularies, most of the 92 glosses were assigned more than 20 root-codes, with one-third (30 glosses) more than 40 root-codes and three glosses ("good", "lie down" and "say") more than 60 each. So many root-codes for individual glosses indicate that comparing specific word entries must have entailed positing provisional roots in lexical form, even if these did not achieve some Platonic ideal as starred proto-forms (Mann [personal communication, 2011] plausibly inferred Coupez and Bastin built the root-code lists per gloss "on the fly"). Without lexical prototypes of some sort, correlation—already a "mind-boggling task" (M. Mann, personal communication, 2011)—would have been impossible. The root-codes in lexico.txt, it is inferred, must represent a continuum from provisional roots to actually posited starred forms. However, following Schadeberg's point, it is important to recognize that the data for Bastin et al. (1999) and all subsequent analyses are even more abstract than plesiomorphic proto-forms: many root-codes symbolize cognation judgements without posited roots. None of this means the judgements were necessarily wrong, simply that their epistemology is typically opaque, and thus they are largely authority statements.

The 3035 total root-codes of lexico.txt minimally reduce the data by about 94%: even treating individual words as units (rather than sound sequences), hypothetically (i.e. discounting entries with no data), $542 \times 92 = 49\,864$ observations. Compared to the word lists, root-codes in many instances encompass significant lexical variety. For instance, of 33 total root-codes for "tree", #1 includes those in Table 2.

Table 1
Modified lexico.txt excerpt (Courtesy M. Mann. Glosses translated, but shown in original French order; double entries = alternative root-codes; language labels as in original)

Guthrie Zone	Code	Name	Tree	Sit	Many	White	Drink	Good	Mouth	Arm	Burn	Ashes
8	00	Ejagham	1	10–28	3	11	1	55	1	1	2	13
8	02	Tiv	13	3	8	8	1	22	10	1	1	5
8	06	Ambele	1	21	0	0	1	4	8	1	13	1
8	94	Asumbo	1	6	0	15	1	55	1	1	13	5
8	05	Amasi	32	21	0	16	1	11	9	1	0	9
9	51	Bangangte	1	19	15	8	1	26	9	1	28	2
9	00a	Mifi	1	4–19	3	6	1	15	9	1	13–28	2
9	00b	Bandjoun	1	4–19	12	8	1	15	9	1	13	2
9	00c	Dschange	1	4–20	3	8	1	15	9	1	13	2
9	70a	Fe'fe'	1	4–19	2	8	1	56	9	1	13	2
9	70b	Bafang	1	4–19	3	8	1	56	9	1	13	2
9	70c	Fefe	1	4–19	1	8	1	56	9	1	13	2
A	15g	Mbo	22	15	17	8	1	4	3	12	2	6
A	24	Duala	31	1–4	8	12	1	4	3	12	2	9
A	26	Pongo	20	1	8	12	1	4	3	12	2	9
A	27	Limba	1	1	8	11	1	4	3	14	2	2
A	31	Bubi	1	1	1	8	4	8	6	1	0	5
A	32b	Puku	1	1	0	11	1	4	3	13	2	2
A	32c	Tanga	1	1	8	11	1	4	3	13	2	2

Table 2
Selected entries corresponding to lexico.txt root-code #1 for *arbre*/tree

Language	Word
806 Ambele	gégyít
A31 Bubi	bótté
B11a Mpongwe	erere
C51 Mbesa	mòté
D37 2 Kumu	mé
E72a 1 Giryama	muhi
F21 Sukuma	-'ti
L42 Kaondeb	ki-chi
M15 Mambwe	icimuti
P31 2 Makwa	mwéré
R31 Herero	omuti
S21 Venda	mu-ri

In Table 2, some correspondences seem intuitive, others (e.g. Ambele, Kumu and Makwa) much less so. However, with no explicit statements or hypotheses of sound shifts, decoding the cognation judgements (and exclusions from these of entries corresponding to the other 32 root-codes) is functionally impossible. In sum, the ontological status of data analysed in influential phylogenetic studies of Bantu languages is "lexical" only via several removes of abstraction. The root-codes that served as the object of analysis are essentially authority statements (of unclear meaning) rather than scientific propositions.

Phylogenetic analyses of lexico.txt

For initial analysis of lexico.txt, Mann developed a correlation matrix to measure similarities based on the number of shared root-codes (M. Mann, personal communication, 2011), resulting in multiple trees and "heterograms" (Bastin et al., 1999). However, Mann's trees were explicitly "phenetic" or similarity-based, and therefore "not fully appropriate to the historical reconstruction of language evolution" (Rexová et al., 2006). Using lexico.txt and Mann's derived correlation matrix, Holden (2002) developed the first phylogenetic treatment (using parsimony). Holden selected 73 Bantu cases from all Guthrie Zones except one, and rooted these using two Bantoid languages (Tiv and Ejagham) as out-groups. The same data set was revisited with Bayesian methods (Holden et al., 2005), expanding the sample by 20 Bantu languages (notably seven from previously omitted Zone G). The resultant phylogeny broadly confirmed Holden's first analysis. The deepest tree splits appeared in the northwest, with groupings designated East, Southwest and Southeast appearing as monophyletic, and with West Bantu comprising a grade of languages with major internal clades ("Savannah/Southwest" and "Equatorial/Forest West"—see Supporting Information), and a "Central" group split between East and West. The same 95 languages were re-investigated (Holden and Gray, 2006) with network analysis and a Bayesian majority-rule tree to address horizontal borrowing vs. vertical descent. This majority-rule tree produced a designation of four major regional groups—West, East, Southwest and Central (the latter two intermediate between West and East). Within West Bantu, which is monophyletic on this tree, several major languages were inferred as diverging simultaneously, while borrowing and dialect continua appeared important for East Bantu. This

analysis collapsed the long-held division between a "Northwest" group and all other languages (Vansina, 1990; Holden, 2002; Holden et al., 2005).

Using lexico.txt for 87 different languages and combining the root-codes with grammatical data, Rexová et al. (2006), unlike other phylogenetic analyses to date, supported "a monophyletic superclade containing all the Bantu languages found in the territories south and east of the rainforest areas of Congo-Kinshasa." They thus disputed a West Bantu taxon encompassing northern and southern areas (contra Holden, 2002), and challenged both the early-split model and migration proposals of Vansina (1995) and Holden (2002). Rexová et al. (2006) concluded, "The main phylogenetic signal of our data favours the colonization of Angola, SW Congo-Kinshasa and surrounding territories from the more eastern source areas." Lexico.txt was utilized by several subsequent analyses, including Dunn et al. (2011), de Filippo et al. (2012) and Currie et al. (2013): the last has underpinned new models of sociocultural evolution (Opie et al., 2014; Guillon and Mace, 2016).

In short, since the first analysis of lexico.txt by Bastin et al. (1999), a succession of Bantu language phylogenies addressing (subsets of) the same, highly processed, numerical data has been developed, in turn grounding phylogenetic analysis of sociocultural patterns.

Current analysis: data and methods

The 95 languages selected by Holden et al. (2005) and Holden and Gray (2006) are targeted here, with the addition of ten Bantoid out-groups. Rather than using lexico.txt's root-codes, however, the actual word lists provide the data. Relabelling of some RMCA lexico language identifiers since Bastin et al. (1999) was adjusted to ensure complete concordance (see Supporting Information). Individual words were rendered into LATEX TIPA 1.3 (Rei, 2004) as resistant to interference across platforms. Evident prefixes and suffixes were eliminated (as advised by A. Mbeje, personal communication, 2015) and in cases where more than one word is listed in the same cell of a word list only the first word was retained. The words are treated strictly as comparable sequences of sounds represented by individual letters and diacritics, analogous to DNA sequences (see Wheeler and Whiteley, 2015). The data are posted at: https://wardwheeler.wordpress.com/data-sets/ and http://www.amnh.org/our-research/anthropology/research.

To Tiv and Ejagham, the two out-groups used by Holden, the ten additional Bantoid cases are as follows: Amasi (805), Ambele (806), Asumbo (894), Atsang (952), Bangang (951), FeFe 1 (970 1), FeFe 2 (970 2), FeFe 3 (970 3), Ghomala 1 (960 1) and Ghomala 2 (960 2) (see Bastin et al., 1999). Word lists for the ten new cases were entered directly from RMCA fieldnotes files. For those (e.g. FeFe, Ghomala) requiring additional tonal phonemes lacking in standard TIPA symbol sets, LATEX entries were adapted following TIPA conventions as far as possible. In total, 287 distinctive sounds were recorded for all 105 languages, encompassing 9288 words [372 cases lack entries in RMCA's records; total entries = 9288 ($92 \times 105 = 9660$-372)].

Phylogenetic analyses were conducted with POY5 (Wheeler et al., 2015) using Direct Optimization techniques (e.g. Wheeler, 2003; Varón and Wheeler, 2013). The 9288 sound sequences were run through a binary csv parser to produce (*.fastc) files for each word, separating sounds to ensure accurate identification of sequences and to aid in identifying any errors deriving from typographic mistakes in data entry. Because of the unknowable nature of relative sound transformation cost, several scenarios were examined in a sensitivity analysis context (Wheeler, 1995):

1. "1–1", where all changes in sound were equally costly (cost = 1), including the gain and loss of sounds;
2. "1–2", where the gain and loss of sounds costs 2, but all other transformations (e.g. vowel to vowel, consonant to consonant, vowel to consonant) cost 1;
3. "vcn", where intravowel and intraconsonant transformations cost 1, transformations between vowel and consonant cost 2, and the gain and loss of sounds costs 2;
4. "vcn2", where intravowel and intraconsonant transformation cost 1, transformations between vowel and consonant cost 2, and the gain and loss of sounds costs 4;
5. "all5", where the gain and loss of sound costs 5, and costs of other sound transformations are based on differences in production from 1 to 4.

To calculate the sound transformation costs in the five scenarios, aspects of the 287 sounds that appear in the language sample were lined out (Table 3). For instance, from sound (i.e. LATEX TIPA entry) E to á, there are two aspects of sound difference (vowel articulation and presence/absence of diacritics), so the transformation cost in "all5" from E to á is 2. For each Swadesh-list word, a sound transformation matrix was created for all sounds that appear in the 105 vocabularies. Then in each cell, the transformation cost calculated from sound aspect differences was input (Table 4). The analysis reconstructs hypothetical proto-forms for all nodes and all words based on the heuristically best tree and cost scenario, thereby eliminating expert retrodiction.

Each cost scenario was examined via the "search" option during a series of runs of 100, 200, 400 and 800 h (when results stabilized) on 64 CPU cores (AMD Opteron™ Processor 6380 at 2500 MHz) for a total of 96 000 CPU hours for each scenario or 480 000 CPU hours overall. Goodman–Bremer (Goodman et al., 1982; Bremer, 1990) support values were estimated (upper bounds) via a round of TBR rearrangement of the heuristically best tree ("swap(all, tbr, visited: 'sample'", "report('bremer.pdf', graphsupports: bremer 'sample'))"). Jackknife support values (Farris et al., 1996) were calculated based on 128 replicates with 36% delete resampling (of words in this case). Each replicate consisted of a single Wagner build followed by TBR branch swapping.

Results and discussion

The five cost scenarios described above resulted in five trees (Figs 3 and 4). The "1–1" all equal costs scenario yielded two equally parsimonious trees at cost 29 767. The "1–2" scenario produced a single tree at cost 38 763. Cost regime "vcn" also yielded a single tree at cost 42 500. A single tree at cost 59 673 was produced by the "vcn2" costs. The "all5" regime yielded a single tree at weighted cost 88 576.

The "all5" scenario was the only cost regime that yielded a monophyletic in-group (and convex outgroup). On this basis, the "all5" analysis tree was chosen as heuristically "best", support values were calculated (Fig. 4), and this tree forms the basis of subsequent discussion.

Based on their distribution on the "all5" tree, Goodman–Bremer supports are characterized as "low" if below 200, "medium" from 200 to 299, "high" from 300 to 399 and "very high" if over 400.

Several clades correspond to geographical (Guthrie) groups and support historical associations of languages in these areas. In several instances, the results accord with trees based on lexico.txt; in others, they are distinctly different. Propinquities on the all5 tree in many cases correlate closely with network adjacencies in Mann's "heterograms" of geographical linkages among languages (Bastin et al., 1999) and correlate also with G. P. Murdock's map of ethnolinguistic groups (Murdock, 1959). Particularly noteworthy clades emerge at node 15 and encompass groups 1–5 (Fig. 4).

Table 4
A section of the sound transformation matrix for the Swadesh-word "all"

	E	N	O	S	\'@	\'E	\'O	\'a	\'e
E	0	3	3	3	3	1	3	2	2
N	3	0	3	3	3	3	3	3	3
O	3	3	0	3	3	3	1	3	3
S	3	3	3	0	3	3	3	3	3
\'@	3	3	3	3	0	3	3	3	3
\'E	1	3	3	3	3	0	3	2	2
\'O	3	3	1	3	3	3	0	3	3
\'a	2	3	3	3	3	2	3	0	2
\'e	2	3	3	3	3	2	3	2	0

Results and conclusions (set out in detail below) may be summarized as follows:

1. There is no support for an "early split" between East and West Bantu, nor associated eastward migration north of the Congo rainforest. This result corroborates previous analyses based on lexico.txt (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006; Rexová et al., 2006; de Filippo et al., 2012; Currie et al., 2013).
2. A large clade of languages south and east of the rainforest, with inferred common descent from an ancestor in the northwest, is supported (Rexová et al., 2006).
3. "West Bantu", a group comprising languages of Guthrie Zones A, B, C, H and parts of D (Holden and Gray, 2006), is not supported.
4. "Southwest Bantu", a group combining Guthrie Zones R and K (Holden, 2002; Holden and Gray, 2006), is not supported.
5. Neither "East Bantu" nor proposed subdivisions into "East Africa" (comprising all E, F, G, J, two M, one P, one D) and "Southeast Africa" (comprising all S, N, and one P) (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006) are supported. There is no support for a single spread from Lake Victoria throughout East Africa and then southeast Africa (Holden, 2002).
6. Tree results compared to physical geography depict an ancestral spread south and southeast

Table 3
Example sound aspects

Latex	Type	Articulation	Articulation2	Rounding	Voicing	Diacritics
E	Vowel	Open-mid	Front	Unrounded	–	n/a
\'a	Vowel	Open	Front	Unrounded	–	High tone
S	Consonant	Fricative	Postveolar	–	Voiceless	n/a
\'@	Vowel	Mid-central	–	n/a	–	High tone

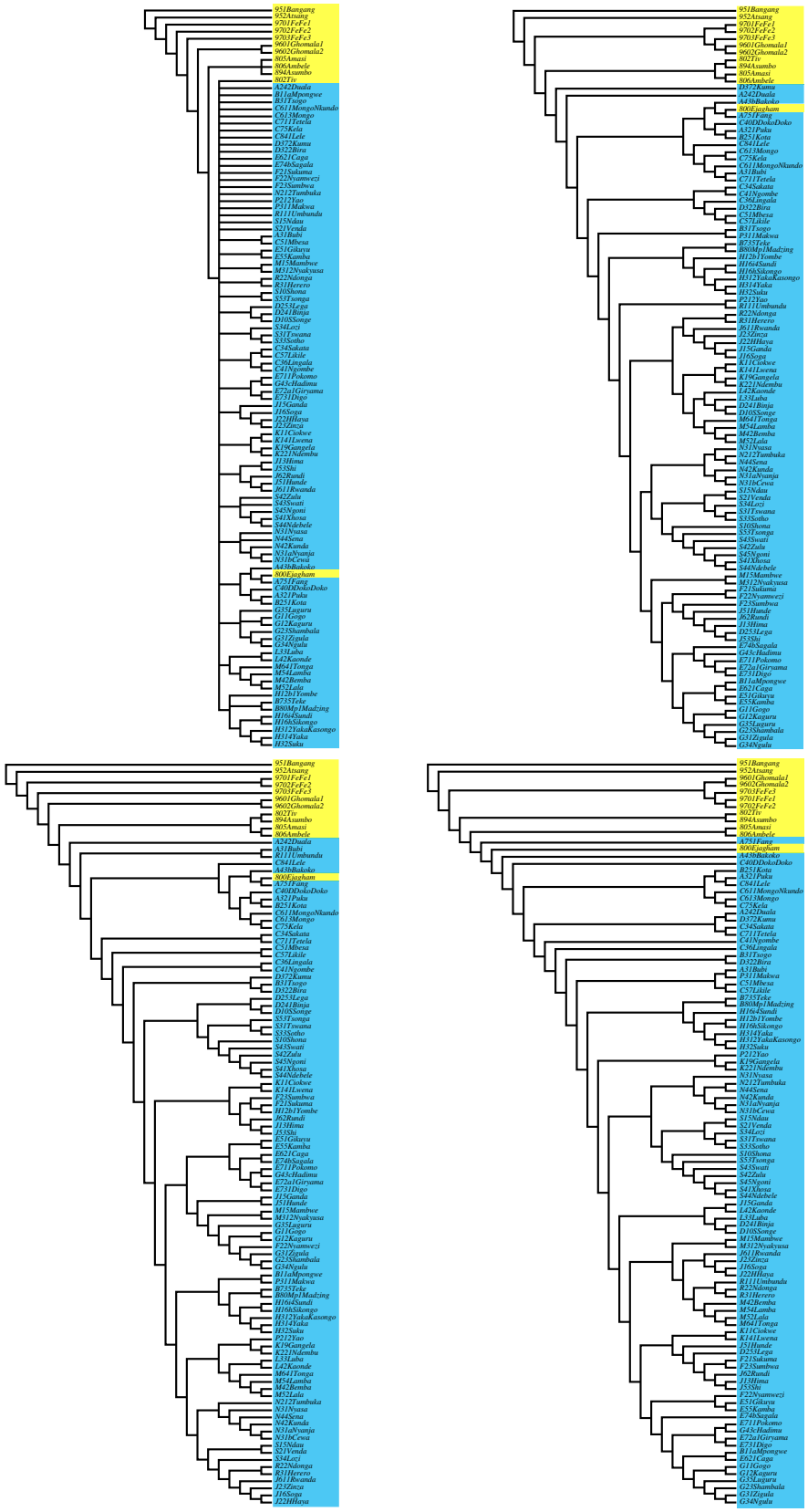

Fig. 3. Heuristically "best" trees derived from different cost regimes. Out-group languages in yellow, in-group languages in blue. Upper left, "1–1", all changes in sounds (including gain and loss) equally costly; upper right, "1–2", gain and loss of sounds costs 2, but all other transformations cost 1; lower left "vcn", intravowel and intraconsonant transformations cost 1, transformations between vowel and consonant cost 2, and the gain and loss of sounds costs 2; lower right, "vcn2", intravowel and intraconsonant transformation cost 1, transformations between vowel and consonant cost 2, and the gain and loss of sounds costs 4. Out-taxon languages begin with numbers (800–9703), whereas in-group languages begin with their Guthrie Zone designation (A–S)

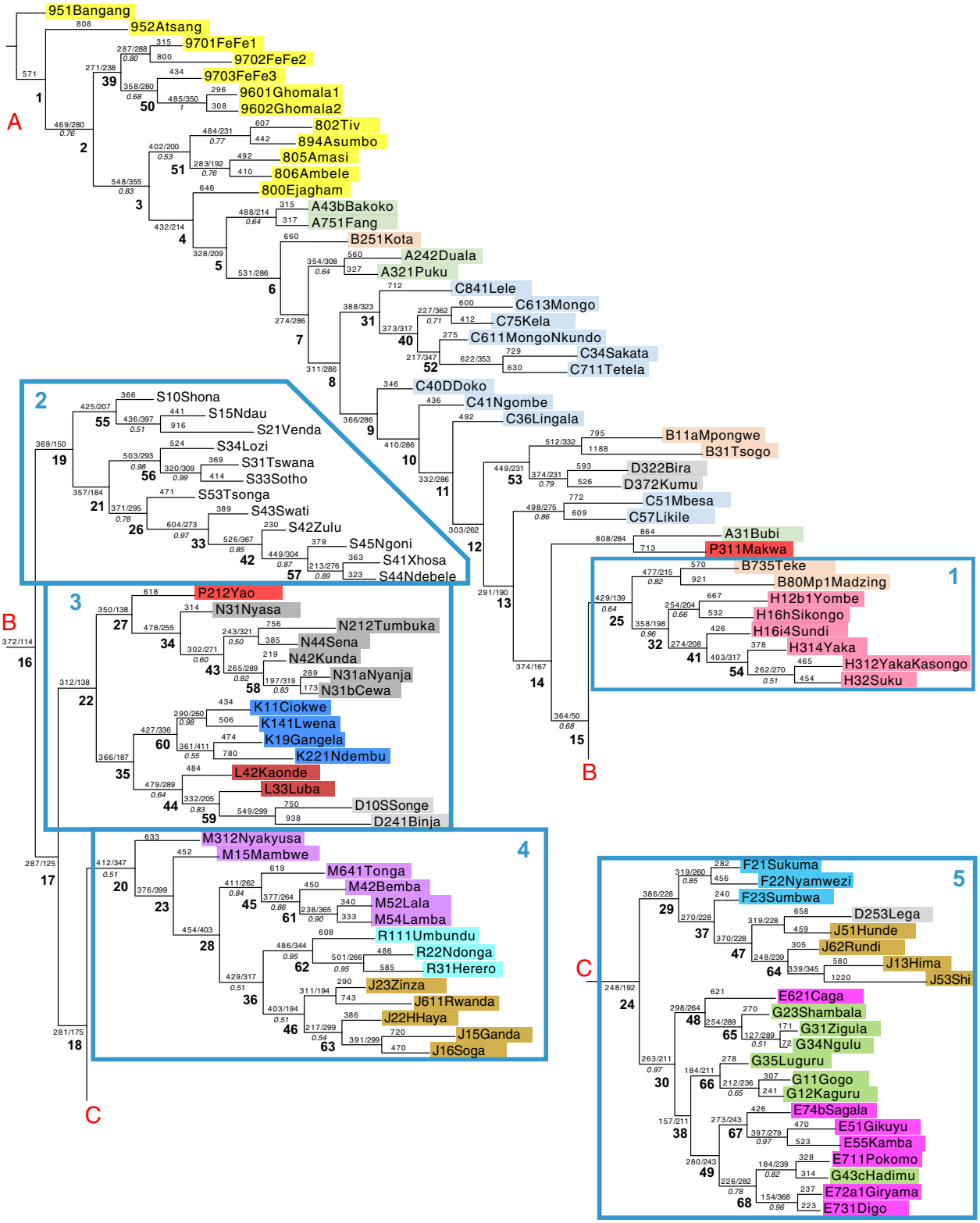

Fig. 4. Bantu–Bantoid language tree based on "all5" cost scenario (gain and loss of sound costs 5, and costs of other sound transformations are based on differences in production from 1 to 4). Clades are numbered as in the text. Terminal languages are coloured based on Guthrie Zones (Fig. 1). Branches are labelled with up to three values: branch length (in weighted sound transformations), Goodman–Bremer support (Goodman et al., 1982; Bremer, 1990; to the right of "/"), and below the branch, jackknife values if >0.50. Terminal and basal-most branches do not have support values, only branch lengths are displayed.

from the lower to mid-Congo River, following rivers and river valleys, with migrations into the East African Plateau via the land bridge between Lakes Tanganyika and Malawi. More derived groups are also broadly consistent with migrations via river systems (notably Kasaï, Zambezi and Ruvuma) in relation to other topographic features. Prior models showing early eastward migration from the Congo River into the interlacustrine East Africa zone across the Albertine Rift or Lake Tanganyika are not supported.

The base of the tree shows a grade of all out-group cases (nodes 1–4), and the "900" group appears paraphyletic—but due to the rooting of the cladogram (within the "950" languages), this could well be due to their basal out-group status and no conclusion should be drawn from these data alone. Four of the five "800" groups (Tiv, Asumbo, Amasi and Ambele) appear as a clade (node 51) with relatively medium Goodman–Bremer (200) and marginal (53%) jackknife support. Ejagham (800) is placed as sister to the remainder of the node 4 subtree (with medium Goodman–Bremer and below 50% jackknife support). The two lineages are geographically proximate (Fig. 1), and the topology confirms the overall concordance of the treatment of words as sound sequences with previous, geography-based results. The most basal in-groups (nodes 5–13) comprise a paraphyletic grade of some Guthrie A, B, C and D languages with internal clades of limited but geographically concordant inclusion, notably (node 31) conjoining C30, 60, 70 and 80 cases (i.e. C34 Sakata to C84 1 Lele) with high Goodman–Bremer support (yet <50% jackknife). The overall pattern here with relatively large branch lengths indicates substantial linguistic diversity. Paraphyly in the Bantu northwest agrees with prior analyses (Holden, 2002; Holden et al., 2005; Rexová et al., 2006; Currie et al., 2013), but on the all5 tree, this grade-like pattern continues into more derived components. The result refutes a unitary "West Bantu" clade (proposed by Holden and Gray, 2006) comprising all A, B, C, H and two D (Bira and Kumu) geographical zones (Fig. 1), and the alternative proposal (Holden, 2002; Holden et al., 2005) of two sister clades comprising "Forest West Bantu" (two B, C, two D, H) and "Southwest Bantu" (K and R).

Node 14 (first branch) presents a striking anomaly, indicating A31 Bubi and P31 1 Makwa as a sister pair: branch lengths are substantial but not markedly different from some other terminal sister pairs (with medium Goodman–Bremer but below 50% jackknife). The same sister-pair appears with other analysis scenarios (e.g. vcn2). Node 14 marks the origin of a series of well-supported groups, several in clear geographical groups. The Bubi–Makwa branch (sister to node/clade 15), however, pairs languages spoken on islands at opposite sides of the continent: Bioko and the Island of Mozambique. Bubi and Makwa also appear anomalous in other weighting schemes, albeit differently than here. Holden (2002) and Holden et al. (2005) place Bubi as the most basal in-group, in accordance with Vansina's hypothesis (e.g. Vansina, 1990) that it represents the earliest split from proto-Bantu, although as an island group with multiple non-Bantu influences over recent centuries, that inference seems odd. For Holden and Gray (2006), Bubi is the first derived clade from their most basal in-group, sister-pair C51 Mbesa–C57 Likile; however, that pair is found in the middle of the continent at the north-eastern fringe of Zone C [in northern Democratic Republic of the Congo (DRC)], far distant from Bubi, and with multiple A, B and C Zone cases intervening geographically. Makwa's placement on lexico.txt trees (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006) is also somewhat enigmatic, grouping with the N languages (but not with its most proximate neighbour, P21 2 Yao), either sister to a clade of N and S languages (Holden, 2002) or as forming a terminal sister-pair with N21 2 Tumbuka (of northern Malawi) on the N clade (Holden et al., 2005; Holden and Gray, 2006). Yet geographically, Tumbuka is the most distant N case of all from Makwa, more than 800 km airline, and across Lake Malawi (the Bubi–Makwa pairing is re-examined below).

Node 15 defines a large clade (low Goodman–Bremer, 68% jackknife support), encompassing all the major groups higher on the tree, consisting of the H, J, K, L, M, N, R and S languages, plus two B and one P. This clade does not appear in the most directly comparable analyses of lexico.txt (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006), but agrees quite closely—notwithstanding the different languages sampled, and the addition of grammatical data—with the group south and east of the rainforest obtained by Rexová et al. (2006).

Clade 1 (node 25, box 1 in Fig. 4, with low Goodman–Bremer and 64% jackknife support) is sister to a group encompassing Clades 2–5. Clade 1, in the lower Congo River Basin below the southern fringes of the rainforest, includes all H Zone languages (with low Goodman–Bremer, but high 96% jackknife support) and two nearby B70 and B80 languages (with medium Goodman–Bremer and 82% jackknife support). Within Clade 1, the H languages ally as a monophyletic subgroup (node 32), sister to the two B languages (B73 5 Teke and B80 Mp1 Madzing) that themselves form a terminal sister-pair. Thus, Clade 1 appears both as a clade and as a reasonable geographical unit. Certainly as regards the H subclade, Guthrie's zonal grouping is supported as a monophyletic unit.

All remaining components of the tree compose a large clade (node 16) of languages further south, southwest, southeast and east of the rainforest. Within this, Clade 2 (node 19, box 2 Fig. 4), consisting of all S cases (corresponding with Murdock, 1959; "South-eastern Bantu" peoples), branches off as a coherent group, suggesting Zone S comprises a meaningful linguistic as well as geographical unit (similar unitary Zone S clades appear on the three directly comparable lexico.txt trees). In addition, within Clade 2 are three subclades that each constitute a geographically and/or historically plausible grouping, consistent with their Guthrie-code proximities, network adjacencies (Mann's heterograms) and adjoining ethnolinguistic territories (Murdock, 1959, Map 17): (a) all three S10 and S20 cases (node 55); (b) all three S30 cases (node 56); and (c) all six S40 and S50 cases (node 26). S45 Ngoni's geographical incongruity (in Malawi) vis-à-vis its tree locus clearly reflects the Ngoni migration (ca. 1825–1835) from southern Africa during the Zulu wars (e.g. Thompson, 1981). The propinquities of S45 Ngoni and S42 Zulu (the latter immediately basal to the former—node 42) on the all5 tree (and in earlier lexico.txt analyses) corroborate a characterization of Ngoni as a "Zulu dialect" (Gowlett, 2003, p. 610). Most of these internal S clades are supported by medium Goodman–Bremer and relatively high jackknife values.

Clade 3 (node 22, box 3) comprises two subclades: (a) on which all N languages group with one of the two P cases (P21 2 Yao) at node 27 (with low Goodman–Bremer and below 50% jackknife support); (b) grouping both L, two D and all four K at node 35, and splitting into two smaller clades—K (node 60 with high Goodman–Bremer and below 50% jackknife) and L-D (node 44 with medium Goodman–Bremer and 64% jackknife). Both sister pairs K11 Ciokwe–K14 1 Lwena and K19 Gangela–K22 1 Ndembu seem plausible geographically. The D sister-pair (D10S Songe and D24 1 Binja) and its branching with L33 Luba and L42 Kaonde (adjoining groups on Murdock, 1959, Map 17) also reflect geographical proximities. The N languages (node 34, medium Goodman–Bremer support) form a subgroup (in contrast to prior analyses, except Holden and Gray, 2006) correlative with a close geographical group into which only one language intrudes on the ground (Fig. 1) that is not a member of the clade: S45 Ngoni (from clade 2, between N31a Nyanja and N21 2 Tumbuka), for the historical reasons noted above. That the N languages group (node 27) with the only nearby P language (P21 2 Yao) is again geographically congruent, with Yao appearing next on the tree to N31 Nyasa (and directly adjacent on Mann's heterograms). This contrasts sharply with Yao's location on prior trees, where it either comprises a terminal sister-pair with M31 2 Nyakyusa (Holden, 2002; Holden et al., 2005) or is immediately basal to a group in which Nyakyusa groups with six G Zone languages (Holden and Gray, 2006; see also Rexová et al., 2006)—in both instances set apart from the N languages. On these prior trees, rather than Yao, it is Makwa (the other P case in the sample) that groups with the N group (Makwa's anomalous position is noted above). Yao's position on the tree appears more congruent with geography, and more so than Makwa's proposed propinquity with the N group, either as a grade or as a sister-pair with distal N21 2 Tumbuka (Holden et al., 2005).

Clade 4 (node 20, box 4) conjoins three separate geographical units: R, M and five J languages, respectively (with high Goodman–Bremer and marginal jackknife support). The R languages of Angola and Namibia, and J languages (J15 Ganda, J16 Soga, J22 Haya, J23 Zinza and J 61 1 Rwanda) of Uganda, Tanzania and Rwanda, form a subclade that splits evenly into separate subgroups as R (node 62) and J (node 46). Apart from Rwanda, all these J cases are from Guthrie's original Zone E, sometimes categorized now as "JE" (Bastin, 2003; Maho, 2009).

The M languages group, but as a grade (node 20, first branches, with high Goodman–Bremer support) that generally follows a NNE–SSW geographical arc from southwestern Tanzania through central Zambia to northern Zimbabwe. M52 Lala and M54 Lamba (in the Central and Copperbelt Provinces of Zambia) form a terminal sister-pair: they are the two closest M languages according to Guthrie codes, Mann's heterogram adjacencies and Murdock's ethnolinguistic map (Murdock, 1959, Map 17). To that sister-pair's northeast, M15 Mambwe and M31 2 Nyakyusa are located beyond several intervening languages/peoples (Bastin et al., 1999; Murdock, 1959, Map 17). Geocoordinates listed for the M42 Bemba vocabulary (Bastin et al., 1999, p. 21) lie within Lamba territory, although three societies to the northeast (Aushi, Unga and Bisa) separate Lamba from Bemba on Murdock's map: this may explain the large branch lengths and branching patterns at node 61 (with high Goodman–Bremer and high jackknife values through the clade). M64 1 Tonga lies to the southwest of Lala and Lamba beyond several intervening ethnolinguistic groups (Kaonde, Lenje, Ila, Nsengia; Murdock, 1959, Map 17). In other words, the monophyletic Lala-Lamba clade is geographically concordant, and the paraphyletic descent pattern of the whole M group correlates with the known geography of ethnolinguistic distribution, in a generally north-south vector. The R language clade (node 62) represents the southwesternmost Bantu cases. Geographically, they are most proximate to each other (and within Murdock, 1959, "South-western Bantu" culture province), with adjacent or close network links on Mann's heterograms, and with R31

Herero and R22 Ndonga, the most geographically proximate, forming a terminal sister-pair.

Clade 5 (node 24, box 5) shows a group comprising all G and E languages (with medium Goodman–Bremer and high–97%–jackknife support), all three F, four J (J 13 Hima, J51 Hunde, J53 Shi and J62 Rundi) and one D (D25 3 Lega) that is geographically close to the four J cases (especially Shi and Rundi). Except for Hima—which exhibits some distinctions as influenced by proximate Nilo-Saharan languages (Bastin, 2003)—all these J languages lie in Guthrie's original Zone D (now sometimes categorized as "JD"), and their grouping with D25 3 Lega (that forms a terminal sister-pair with J51 Hunde) on the all5 tree appears salient here: see below. F23 Sumbwa (node 37) is immediately basal on the tree to this J-D subclade (with medium Goodman–Bremer support), consistent with the notion that "Sumbwa is most likely an original member of J" (Nurse and Philippson, in Bastin, 2003, p. 251). F23 Sumbwa derives from node 29, so the grade of F (Guthrie Zone) cases (located in a close geographical network southeast of Lake Victoria) appears "ancestral" here to the J-D subclade. Tree topology vis-à-vis geography suggests the separate J groups (of Clades 4 and 5) may correlate with adaptive and/or historical differences. If F is genuinely "ancestral" to the Clade 5 J cases, this may imply a migration northwest by the latter into the Albertine Rift highlands from southeast of Lake Victoria, distinct from and later in time than the migration of the Clade 4 J group around the Victoria lakeshore.

The E and G languages, of southern Kenya, northeastern Tanzania and one (G43c Hadimu) of Zanzibar, split into their own clade (node 30) with relatively short branch lengths. This E-G subclade constitutes a close geographical connection, in which there are no intervening cases not part of the clade (Mann's heterograms). Both the subclade and its geography support a common origin, with one main exception: E74b Sagala, whose geocoordinates place it in Tanzania (in Zone G), south or southwest of all seven G cases, and separated by them from all the other E cases. The coordinates (36.3°E, 6.9°S; Bastin et al., 1999, p. 18) are clearly in error, however, referring to language G39 Sagala (not included in the RMCA lexico data set), rather than E74b Sagala. E74b Sagala's position on the tree with E51 Gikuyu and E55 Kamba is more intelligible if assigned to its proper location in southeastern Kenya, i.e. 38.5°E, 3.6°S, slightly northwest of E73 1 on Fig. 1. Appearance at this position on the tree suggests that migration of Bantu languages into this region of East Africa represents a relatively late phase of pan-continental differentiation. Some sister pairs suggest close historical and geographical relationships: especially G31 Zigula–G34 Ngulu, E51 Gikuyu–E55 Kamba and E72a 1 Giryama–E73 1 Digo. The Gikuyu–Kamba pair approximates the northern limit of Bantu languages in East Africa, where they abut Maasai and Samburu (Nilo-Saharan languages) to the west and northwest, and lie close to the tip of a northward pointing Bantu "peninsula".

In short, the Bantu language clades numbered 1–5 above, with a series of internal subclades, largely correspond with evident geographical patterns that in turn point towards common origins among their language-bearing populations.

Topology, topography and migration

When treated as a model of historical development of the Bantu family, the all5 tree points towards serial migrations/expansions across the sub-Saharan landscape (Fig. 5). Obviously, as with all trees, these are hypotheses—and subject to revision via inclusion of larger samples and possibly also via analysis of reticulate vs. vertical relationships. However, as is, the all5 tree offers several new perspectives on Bantu language history.

In common with phylogenetic analyses based on lexico.txt (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006; Rexová et al., 2006; de Filippo et al., 2012; Currie et al., 2013; see also Grollemund et al., 2015), an "early split" between East and West Bantu, and associated eastward migration north of the rainforest, is refuted here. A "superclade" of languages south and east of the rainforest (Rexová et al., 2006) is corroborated, indicative of common descent from an ancestor in the northwest. The paraphyletic pattern among in-groups at the base of the all5 tree disconfirms "West Bantu" (Holden and Gray, 2006) as a monophyletic unit (including A, B, C, H and parts of D), or "Southwest Bantu" as conjoining R with K (Holden, 2002; Holden and Gray, 2006). Neither do Clades 2–5 support "East Bantu" as a clade, or proposed subdivisions into "East Africa" (comprising all E, F, G, J, two M, one P, one D), and "Southeast Africa" (comprising all S, N, and one P) (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006). The idea of "a single long spread of East Bantu languages from Lake Victoria into the rest of East Africa and then to southeast Africa" (Holden, 2002, p. 798) is disconfirmed.

Guthrie's A Zone is unsupported as an integral entity on the all5 tree, especially with the anomalous pairing of A31 Bubi with P31 1 Makwa. Bubi is generally regarded as an "isolate" of unknown historical affiliation (Bastin and Piron, 1999, p. 153), so the argument—purportedly affirmed by its basal-most position on lexico.txt trees—that Bubi represents the earliest split from proto-Bantu seems quixotic. Bubi and Makwa account for two of three island

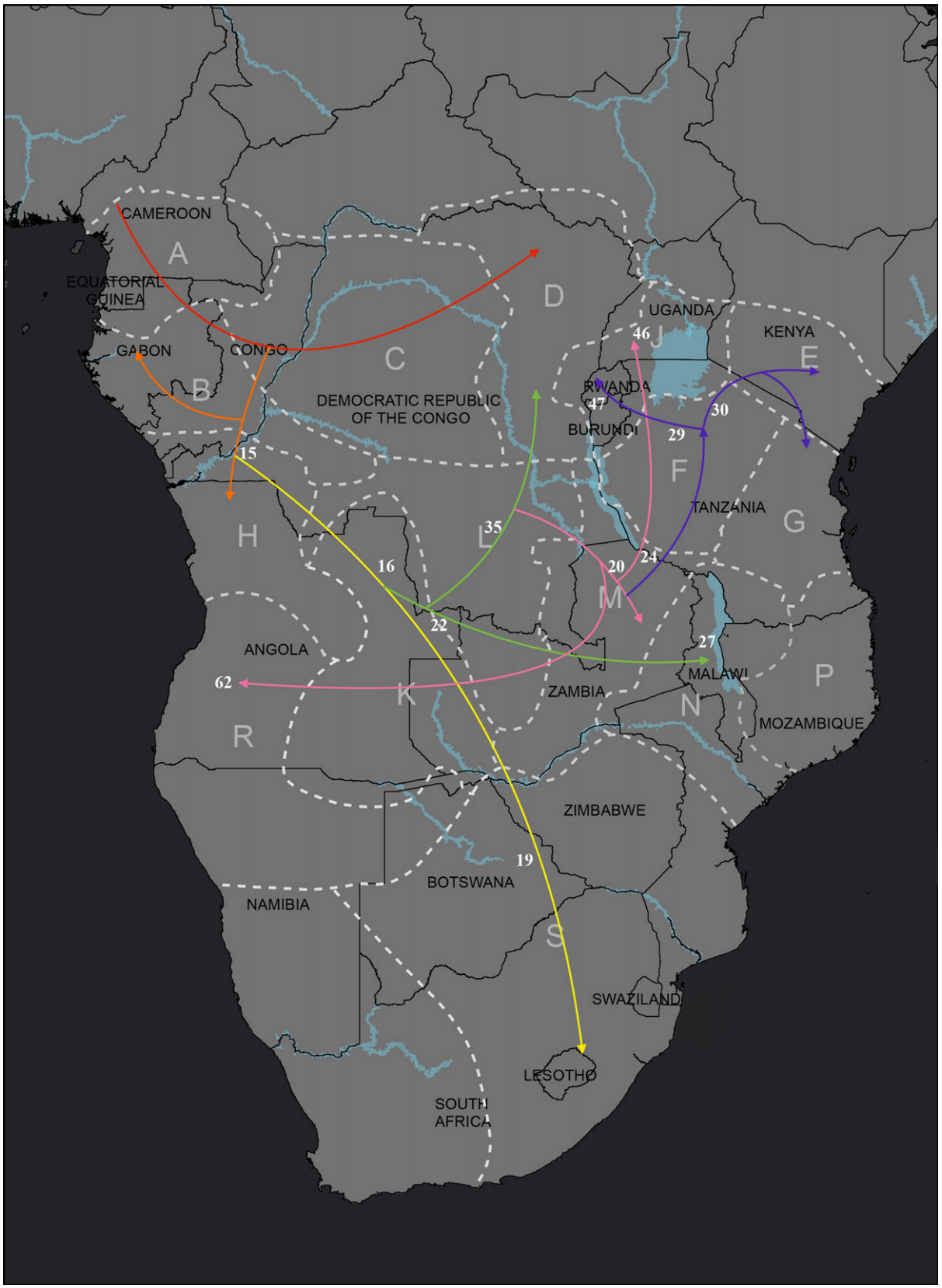

Fig. 5. Migration trajectories based on language tree of Fig. 4 and zones of Fig. 1

vocabularies in the sample (the third is G43c Hadimu, Zanzibar). Both Bubi and Island Makwa experienced extensive Portuguese and other colonial influence over five centuries (e.g. Vansina, 1990, pp. 137–146; Sundiata, 1994; Harries, 2016). Bubi, on Bioko 100 km south of the Nigerian coast, absorbed influxes of refugee slaves (Sundiata, 1994), such that in the 1880s, Bubis described a major part of their population as "Potugi" (Portuguese), i.e. descendants of indigenous Bubi and refugees from São Tomé e Príncipe (Baumann, 1888). Bubi phonology and morphology show influence by Spanish and English (Rurangwa, 1989, p. 77, p. 90), so Portuguese (present earlier) may also have had effects. The Makwa vocabulary was recorded on the Island of Mozambique, capital of Portuguese East Africa from the 16th to 20th centuries (Newitt, 1995). Makwa is either heavily influenced by or even a dialect of Swahili, some of whose vocabulary "borrowed massively from Arabic and later Portuguese" (Blench, 2012). While island geography and long-term Lusophone influences are intriguing, our data set alphabet does not include unique phonological accretions from either Bubi or Makwa *per se*, and neither word list evidences plausible morphemic matches with equivalent Portuguese glosses. Whatever the cause, the all5 tree's topology does not support the notion that Bubi represents the "earliest split" from "proto-Bantu".

Paraphyly among the basal in-groups on the all5 tree contrasts with the clear clades with relatively short branch lengths in geographically peripheral areas, notably Zones H, S, N, K, R and combined G-E. This is consistent with Sapir's "centre of gravity" principle (e.g. Sapir, 1949 [1916], p. 455), in which the greatest degree of in-group diversity (in our case the diversity is cladistic) is held to coincide with the origin of language-family dispersals, in contrast to the greater degree of uniformity in distinct peripheral areas. By this principle, the Bantu centre of gravity clearly lies in the northwest. Where clades and grades on the all5 tree are coordinate with Zones, they support Guthrie's argument that his classification was in part "genealogical" (i.e. monophyletic) notwithstanding standard rejection of that position.

Node 15 on the all5 tree (where Clade 1 divides from Clades 2–5) represents the hypothetical ancestral language-bearing group that emerged south of the rainforest somewhere in the lower to mid-Congo River region (compare Currie et al., 2013, Fig. 2b, area 5). Clade 1 (node 25) reflects a hypothetical ancestral group that remained *in situ* or moved either down- or upriver within Zone H. Just as for more derived groupings on the tree, rivers and river valleys seem obvious natural corridors of migration (Kouerey et al., 1989, pp. 185–186; partly contra Grollemund et al., 2015, p. 13 298). Hypothetical Clade 2–5 groups moved south, perhaps up the Kasaï River and its tributaries from near the Kasaï–Congo confluence, and split at node 16 between hypothetical language Clades 2 and 3–5: a reasonable geographical proxy is the upper Kasaï near the divide between the Congo and Zambezi River systems (near the southwest DRC border with Angola).

Clade 2's hypothetical ancestral language group (node 19) moved south-southeast to eventually occupy Zone S. Linguistic correlations with the archaeological record, while pervasive in the literature, are ipso facto untestable. However, it may be worth noting that the inferred migration for the linguistic group (Clade 2) is not dissimilar to the trajectory proposed for Early Iron Age spread of Western Stream or Kalundu Tradition ceramics (Huffman, 1989, fig. 36). Within Clade 2, internal descent patterns correlate in southeast Africa with an initial northwest-east/south movement (node 55: Shona, Ndau, Venda), and a straight north–south movement (node 56: Lozi, Tswana, Sotho). Partly reversing that trend, the higher numbered branches, although consistent with initial north–south descent (node 26: Tsonga, Swati, Zulu), suggest a later movement (node 42) from Zulu to both the north (Ngoni, Ndebele) and southwest (Xhosa). This maps surprisingly well onto the 19th-century diaspora from Zululand (the "Mfecane"), entailing northward migrations of Ngoni to Malawi (Thompson, 1981) and Ndebele to Zimbabwe (Rasmussen, 1978). In short, branching patterns within Clade 2 conform to the general northwest–southeast trajectory from the lower Congo Basin to southeast Africa, suggested by splits at nodes 15, 16 and 19. Where that directional pattern shifts among S cases near the tips of the tree, it is consistent with known historical population dispersal.

After separation from Clade 2, Clade 3's hypothetical ancestral group (node 22) spread east, northeast and southeast within central Africa between the headwaters of the Kasaï and the upper Lualaba River (the main upper Congo tributary) to the east, or the approximate area of former Katanga Province (DRC). Clade 3 splits into two subclades. One (node 35) correlates with those on the Kasaï (K11, K22) and Lualaba (D10S, D24 1, L33, L42), and nearby headwaters of the Kasaï divide with the Zambezi (K14, K19, L42; in the border region of DRC, Angola and Zambia). The other group (node 27) radiated eastward, probably via the Zambezi and its northern tributaries (notably Luangwa) towards Lake Malawi, and in two cases (N31 and P21 2) around the lake's north end to the eastward-flowing Ruvuma River (that empties into the Indian Ocean). N31's sister group with nearby P21 2 Yao (farther down the Ruvuma at the border between Zones N and P) appears consistent with geography, notably vis-à-vis the drainage system. Although not consistent with the placement of P31 1 Makwa on the

all5 tree, it seems plausible that P groups to the east and southeast of Yao reflect a further eastward migration from the N Zone: a hypothesis that requires testing minimally with more P cases. At all events, the topology does not suggest P descends from interlacustrine groups to the north, in contrast to lexico.txt trees.

Clade 4 divides into three geographical groups (M, R and five J). The northnortheast–southsouthwest M arc from southern Tanzania through southern Zambia begins (M15, M31 2) near the eastern Congo–Zambezi drainage divide between Lakes Tanganyika and Malawi, on northern tributaries to the middle Zambezi (the Luangwa and Songwe, the latter emptying into northern Lake Malawi). The topology suggests M may represent the earliest split from Clade 3, whose subclades (K-L-two D to the west, N-one P to the southeast) are geographically bisected by the M arc. M's descendant clade R moved west into Angola and northern Namibia, perhaps via the upper Zambezi and its tributaries (notably, Lungué-Bungo) or crossing to the Okavango River system, and thence towards westward-flowing rivers draining into the Atlantic [including the Cuvo (R11 1 Umbundu), and Cunene (R31 Herero)]. M's second descendant, the ancestor of Clade 4's J subclade, moved north from M into the East African Plateau (east of Lake Tanganyika), probably via the land bridge between Lake Tanganyika and Lake Malawi (near the Ufipa Plateau and Lake Rukwa), and north to the west side of Lake Victoria. This northward movement contrasts sharply with other reconstructions of Bantu expansion in East Africa depicting a north–south flow. A similar northward movement and division appears reflected at node 24 on the all5 tree, in which Clade 5 split from Clade 4, later but perhaps also near the drainage divide between Lakes Malawi and Tanganyika.

Zone J was a later superimposition (by A. E. Meeussen, Bastin et al., 1999, p. v) on Guthrie's (1948) original zones, carved out from his easternmost D and westernmost E. However, the two separated J clades on the all5 tree—in contrast to their unity on previous trees (Holden, 2002; Holden et al., 2005; Holden and Gray, 2006)—suggest Guthrie's original grouping has persistent historical value. On Clade 4, all J cases except Rwanda belong to Guthrie's original Zone E, as on Clade 5, all except Hima belong to original Zone D. Clade 4's J subclade corresponds with contiguous woodland–savannah lakeshore cases around northern (J16 Soga, J15 Ganda) and western (J23 Zinza, J22 Haya) Lake Victoria, with Rwanda (J61 1) immediately to the west of Zinza and Haya. In contrast, Clade 5's J (plus one D) subclade occupies the Great Lakes highlands of the Albertine Rift, west of the woodland–savannah. Except for J51 Hunde (on the western slopes of the Mitumba Mountains northwest of Lake Kivu), all Clade 5 J cases lie above 1500 m (the highlands demarcation), approximately 300 m higher on average than the J cases of Clade 4 (except Rwanda). A distinction between lakeshore/woodland–savannah (JE) vs. highland (JD) suggests variant economic adaptations and/or resource competition. Further, from the all5 tree's topology, the highland groups represent a separate, later migration into the region, although probably along a similar pathway from southern Lake Tanganyika.

Clade 5's two subclades are consistent with a geographical separation near the south-eastern shores of Lake Victoria, where from northern Zone F, the J-D subsubclade (node 47) radiated northwest to the Albertine Rift highlands, and the G-E subclade (node 30) eastward into the Kenya highlands and north-eastern Tanzania via river systems that drain into the Indian Ocean (notably the Tana, Athi and Galana rivers for Zone E, and Pangani and Wami rivers for Zone G).

Depictions in late-split models of eastward migration from west of central Lake Tanganyika into the interlacustrine zone appear oblivious to topographic and hydrological factors (e.g. Pakendorf et al., 2011, fig. 2; Currie et al., 2013, fig. 2b; Grollemund et al., 2015, fig. 2a). These models imply seemingly unimpeded ascent up and across the Albertine Rift escarpment and mountain system (including the Rwenzori and Mitumba Mountains) or laterally across Lake Tanganyika (the world's longest freshwater lake, its second largest and deepest). Clearly, the intent is to depict general trajectories, but plausible historical reconstruction should not ignore topographic reality. The phylogenetic signal in the all5 tree suggests a barrier to eastward entry into Zone J and easternmost D that correlates on the landscape with the upthrust landforms and lakes of the Albertine Rift (that extends from Lake Albert in the north to the south end of Lake Tanganyika). Tree topology in relation to late Holocene topography suggests Bantu migration into the interlacustrine region flowed from the south, via the opening between Lakes Malawi and Tanganyika, with its fewer natural barriers to population movement.

Naming the clades by geographical descriptors is not undertaken here, even though some may warrant these: for example, Southeast Bantu for Zone S, Southwest Bantu for Zone R, and Northeast Bantu for Zones E–G. The tree's topology and implications for migration patterns obviate larger groupings by geographical region. The analysis and resultant clades show spread patterns that correlate more closely with topography.

Conclusion

The results of this study move beyond existing phylogenetic models of Bantu languages that depend on untestable authority statements. Concentration on empirically recorded words as sound sequences offers a more testable evidence-based method that: (a) is agnostic regarding sound shifts inferred by the comparative method—phonological differences are addressed by computational parsing of phonemes and phoneme sequences expressed by a standard orthography, with minimal a priori assumptions; (b) avoids a priori specification of plesiomorphic proto-forms dependent on cognation judgements and authority statements (instead reconstructing hypothetic proto-forms for each node and each word via explicit methods); and (c) is not correlated to dates of historical branching, via glottochronology, linguistic palaeontology or any other models (e.g. molecular clock, Grollemund et al., 2015). While a few potential correlations with the archaeological record are mentioned, these are untestable scientifically with any methods known.

The all5 tree (and the trees based on alternative cost scenarios) contain some incongruent aspects, and some inferences require further investigation. It will be important also to test reticulate patterns against vertical transmission targeted here. Nonetheless, the present analysis yields more fine-grained and robust results than existing approaches, and significantly clarifies the historico-geographical spread of Bantu languages consistent with salient topography and hydrology of sub-Saharan Africa.

Acknowledgements

This inquiry was funded by NSF ("Explaining Crow-Omaha Kinship Structures with Anthro-informatics", BCS-0925978, 2009-2012), and DARPA SIMPLEX ("Integrating Linguistic, Ethnographic, and Genetic Information of Human Populations: Databases and Tools", DARPA-BAA-14-59 SIMPLEX TA-2, 2015-2018). We thank: Theodore Powers (now University of Iowa) for initial steps; AMNH Anthropology interns Julia Broach, Claire Feuer, Sara Schwerd and Nicole Striepe, for preliminary transcriptions; Yvonne Bastin, Muriel Garsou and Jacky Maniacky (RMCA) for gracious responses and for word lists; Koen Bostoen (Ghent University) regarding language-label concordances; Mark Pagel (University of Reading) and Simon Greenhill (now Max Planck Institute) for help locating source data; Lutz Marten (School of Oriental and African Studies), who connected us with Michael Mann (SOAS, retired)—who generously provided lexico.txt, and was enormously helpful in explaining research methods; Thilo Schadeberg (University of Leiden), regarding RMCA's Bantu Lexical Reconstructions; Audrey Mbeje (University of Pennsylvania) for advice on Bantu morphology; Jeff Good (University at Buffalo) and Herrmann Jungraithmayr (University of Frankfurt) for responses on Bantoid tonal systems; and Rebecca Grollemund (University of Missouri) for generously sharing her own data sets and for sound advice. We thank AMNH Anthropology Division artist Kayla Younkin, and Thomas Blaber, Nels Nelson North American Archaeology Lab, Steven Thurston of AMNH Invertebrate Zoology for figures, and a *Cladistics* Associate Editor for helpful comments on the manuscript. We remain solely responsible for all content.

References

Alves, I., Coelho, M., Gignoux, C., Damasceno, A., Prista, A., Rocha, J., 2007. Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations. Hum. Biol. 83, 13–38.

Barbançon, F., Warnow, T., Evans, S.N., Ringe, D., Nakhleh, L., 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30, 143–170.

Bastin, Y., 2003. The interlacustrine zone (Zone J). In: Nurse, D., Philippson, G. (Eds.), The Bantu Languages. Routledge, New York, NY, pp. 501–528.

Bastin, Y., Piron, P., 1999. Classifications lexicostatistiques: bantou, bantou et bantoïde: de l'intérêt des "groupes flottants". In: Hombert, J.-M., Hyman, L.M. (Eds.), Bantu Historical Linguistics: Theoretical and Empirical Perspectives. CSLI, Stanford, pp. 149–163.

Bastin, Y., Coupez, A., de Halleux, B., 1983. Classification lexicostatistique des langues bantoues (214 relevés). Bulletin des séances de l'Academie Royale des Sciences d'Outre-Mer, nouvelle série 27, 173–199.

Bastin, Y., Coupez, A., Mann, M., 1999. Continuity and Divergence in the Bantu Languages: Perspectives from a Lexicostatistic Study. Number 162 in Annales du Musée Royalde l'Afrique Centrale, Sciences Humaines. Musée Royal de l'Afrique Centrale, Tervuren.

Bastin, Y., Coupez, A., Mumba, E., Schadeberg, T.C. (Eds.), 2002. Bantu Lexical Reconstructions 3. Royal Museum for Central Africa, Tervuren. http://www.africamuseum.be/collections/browsecollections/humansciences/blr/.

Baumann, O., 1888. Eine afrikanische tropen-insel, Fernando Póo und die Bube, dargestellt auf Grund einer Reise im Auftrage der K. K. Geographischen Gesellschaft in Wien. Hölzel, Vienna.

Blench, R., 2012. Two vanished African maritime traditions and a parallel from South America. Afr. Archaeol. Rev. 29, 273–292.

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. Curr. Anthropol. 56, 354–384.

Bowern, C., Evans, B. (Eds.), 2015. The Routledge Handbook of Historical Linguistics. Routledge, New York.

Bremer, K., 1990. Combinable component consensus. Cladistics 6, 369–372.

Bryant, D., Filimon, F., Gray, R.D., 2005. Untangling our past: Languages, trees, splits and networks. In: Mace, R., Holden, C.J., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. Institute of Archaeology, University College, London, pp. 69–85.

Campbell, L., Poser, W., 2008. Language Classification: History and Method. Cambridge University Press, New York, NY.
Carlo, P.D., Good, J., 2015. Comments. Curr. Anthropol.56, 368.
Coupez, A., Évrard, E., Vansina, J., 1975. Classification d'un échantillon de langues bantoues d'aprés la lexicostatistique. Africana Linguistica 6, 131–158.
Coupez, A., Bastin, Y., Mumba, E. (Eds.), 1998. Reconstructions Lexicales Bantoues 2. Musée Royal de l'Afrique Centrale, Tervuren. CD-Rom.
Currie, T.E., Meade, A., Guillon, M., Mace, R., 2013. Cultural phylogeography of the Bantu languages of sub-Saharan Africa. Proc. R. Soc. B 280, 1–8.
Dimmendaal, G., 2011. Historical Linguistics and the Comparative Study of African Languages. John Benjamins, Amsterdam.
Dunn, M., Greenhill, S.J., Levinson, S.C., Gray, R.D., 2011. Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473, 79–82.
Durie, M., Ross, M. (Eds.), 1996. The Comparative Method Reviewed: Regularity and Irregularity in Language Change. Oxford University Press, New York.
Ehret, C., 2001. Bantu expansions: Re-envisioning a central problem of early African history. Int. J. Afr. Hist. Stud. 34, 5–41.
Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.
de Filippo, C., Bostoen, K., Stoneking, M., Pakendorf, B., 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc. R. Soc. B 279, 3256–3263.
Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zoologica Fennica 169, 19–35.
Gowlett, D., 2003. Zone S. In: Nurse, D., Philippson, G. (Eds.), The Bantu Languages. Routledge, New York, pp. 609–638.
Greenberg, J.H., 1987. Language in the Americas. Stanford University Press, Stanford.
Greenberg, J.H., 2005. Indo-Europeanist practice and American Indianist theory in linguistic classification. In: Croft, W. (Ed.), Genetic Linguistics: Essays on Theory and Method, by Joseph H. Greenberg. Oxford University Press, Oxford, pp. 153–189. Original publication date 1990.
Greenhill, S., Currie, T., Gray, R., 2009. Does horizontal transmission invalidate cultural phylogenies? Proc. R. Soc. Lond. B Biol. Sci. 276, 2299–2306.
Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., Pagel, M., 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc. Natl Acad. Sci. USA 112, 13296–13301.
Guillon, M., Mace, R., 2016. A phylogenetic comparative study of Bantu kinship terminology finds limited support for its co-evolution with social organisation. PLoS One 11, 1–15.
Guthrie, M., 1948. The Classification of the Bantu languages. International African Institute, Oxford University Press, London, UK.
Guthrie, M., 1967–1971. Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. Gregg International, Brookfield, VT, Vol. 4.
Hammarström, H., Forkel, R., Haspelmath, M., 2017. http://glottolog.org.
Harries, P., 2016. Mozambique Island, Cape Town and the organisation of the slave trade in the south-west Indian Ocean, c. 1797-1807. J. South Afr. Stud. 42, 409–427.
Holden, C.J., 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. Proc. R. Soc. B 269, 793–799.
Holden, C.J., Gray, R.D., 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In: Forster, P., Renfrew, C. (Eds.), Phylogenetic Methods and the Prehistory of Languages, McDonald Institute Monographs. McDonald Institute for Archaeological Research, Cambridge, pp. 19–31.
Holden, C.J., Mace, R., 2003. Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. Proc. R. Soc. B 270, 2425–2433.
Holden, C.J., Mace, R., 2005. The cow is the enemy of matriliny: using phylogenetic methods to investigate cultural evolution in Africa. In: Mace, R., Holden, C.J., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. Institute of Archaeology, University College, London, pp. 217–234.
Holden, C.J., Meade, A., Pagel, M., 2005. Comparison of maximum parsimony and Bayesian Bantu language trees. In: Mace, R., Holden, C.J., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. Institute of Archaeology, University College, London, pp. 53–66.
Hombert, J.-M., Hyman, L.M., 1999. Bantu Historical Linguistics: Theoretical and Empirical Perspectives. CLIS, Stanford.
Huffman, T.N., 1989. Iron Age Migrations: The Ceramic Sequence in Southern Zambia. Witwatersrand University Press, Johannesburg.
Kouerey, G.A., Mba, E.E., Mombo, J.-B., Boukoussou, V.M., Ndangoó, P.-C.M., Simon, I., Walter, P., 1989. Unité et diversité du monde Bantu. In: Obenga, T. (Ed.), Les peuples Bantu: migrations, expansion, et identité culturelle. L'Harmattan/CICIBA, Paris/Libreville, pp. 167–186.
List, J., 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. J. Lang. Evol. 1, 119–136.
List, J., Greenhill, S.J., Gray, R.D., 2017. The potential of automatic word comparison for historical linguistics. PLoS One 12, e0170046.
de Luna, K.M., 2016. Collecting Food, Cultivating People: Subsistence and Society in Central Africa. Yale University Press, New Haven.
Mace, R., Pagel, M., 1994. The comparative method in anthropology. Curr. Anthropol. 35, 549–564.
Maho, J.F., 2006. Proto-Bantu. In: Brown, K. (Ed.), Encyclopedia of Language and Linguistics, 2nd edition. Elsevier Science, Amsterdam, Vol. 10, pp. 198–205.
Maho, J.F., 2009. NUGL online: the web version of the new updated Guthrie list, a referential classification of the Bantu languages. http://goto.glocalnet.net/mahopapers/nuglonline.pdf.
Marten, L., 2006. Bantu classification, Bantu trees and phylogenetic methods. In: Forster, P., Renfrew, C., (Eds.), *Phylogenetic Methods and the Prehistory of Languages, MacDonald Institute Monographs*. MacDonald Institute for Archaeological Research, Cambridge, UK, pp. 43–55.
Meeussen, A.E., 1980 (1969). Bantu Lexical Reconstructions. Number 27 in Archives d'Anthropologie. Musée Royal de l'Afrique Centrale, Tervuren.
Murdock, G.P., 1959. Africa: Its Peoples and their Culture History. McGraw-Hill, New York.
Newitt, M., 1995. A History of Mozambique. Indiana University Press, Bloomington.
Nurse, D., Phillipson, G., 2003a. The Bantu Languages. Routledge, New York.
Nurse, D., Phillipson, G., 2003b. Towards a historical classification of the Bantu languages. In Nurse, D., Philippson, G. (Eds.), The Bantu Languages. Routledge, New York, pp. 164–181.
Opie, C., Shultz, S., Atkinson, Q.D., Currie, T., Mace, R., 2014. Phylogenetic reconstruction of Bantu kinship challenges main sequence theory of human social evolution. Proc. Natl Acad. Sci. USA 111, 17414–17419.
Pagel, M., 2009. Human language as a culturally transmitted replicator. Nat. Rev. Genet. 10, 405–415.
Pakendorf, B., Bostoen, K., de Filippo, C., 2011. Molecular perspectives on the Bantu expansion: A synthesis. Lang. Dynam. Change 1, 50–88.
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., Van der Veen, L., Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J-M., Quintana-Murci, L., 2017. Dispersals and

genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356, 543–546.
Rama, T., List, J.-M., Wahle, J., Jäger, G., 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? arXiv:1804.05416v.
Rasmussen, R.K., 1978. Migrant Kingdom: Mzilikazi's Ndebele in South Africa. Collings, London.
Rei, F., 2004. Tipa Manual, Version 1.3. Graduate School of Humanities and Sociology, University of Tokyo, Tokyo.
Rexová, K., Bastin, Y., Frynta, D., 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. Naturwissenschaften 93, 189–194.
Ringe, D., 2006. From Proto-Indo-European to Proto-Germanic. Oxford University Press, New York.
Rosen, D.E., Forey, P.L., Gardiner, B.G., Patterson, C., 1981. Lungfishes, tetrapods, paleontology, and plesiomorphy. Bull. Am. Mus. Nat. Hist. 167, 159–276.
Rurangwa, I., 1989. Enquête linguistique sur le Bubi, langue bantu insulaire de Guinée Equatoriale. In: Obenga, T. (Ed.), Les Peuples Bantu: Migrations, Expansion, et Identité Culturelle. L' Harmattan/CICIBA, Paris/Libreville, pp. 76–100.
Sapir, E., 1949. Time perspective in aboriginal American culture: a study in method. In: Mandelbaum, D.G. (Ed.), Selected Writings in Language, Culture and Personality, by Edward Sapir. University of California Press, Berkeley, pp. 389–462. Originally published in 1916.
Schadeberg, T.C., 2003. Historical linguistics. In: Nurse, D., Philippson, G. (Eds.), The Bantu Languages. Routledge, New York, pp. 143–163.
Simons, G.F., Fennig, C.D., 2017. Ethnologue: Languages of the World, 20th edition. SIL International, Dallas.
St. Arnaud, A., Beck, D., Kondrak, G., 2017. Identifying cognate sets across dictionaries of related languages. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics. (Copenhagen), pp. 2509–2518. Stroudsburg, PA.
Starostin, G., 2009. Review, language classification: history and method, by Lyle Campbell and William J Poser. J. Lang. Relat. 2, 158–174.
Sundiata, I., 1994. State formation and trade: The rise and fall of the Bubi polity, c. 1840-1910. Int. J. Afr. Hist. Stud. 27, 505–523.
Swadesh, M., 1971. The Origin and Diversification of Language. Aldine, Chicago.
Thompson, T., 1981. The origins, migration, and settlement of the northern Ngoni. Soc. Malawi J. 34, 6–35.
Vansina, J., 1990. Paths in the Rainforest: Toward a History of Political Tradition in Equatorial Africa. University of Wisconsin Press, Madison.
Vansina, J., 1995. New linguistic evidence and "the Bantu expansion". J. Afr. Hist. 36, 173–195.
Varón, A., Wheeler, W.C., 2013. Heuristics for the general tree alignment problem. BMC Bioinform. 14, 66.
Walker, R.S., Hamilton, M.J., 2011. Social complexity and linguistic diversity in the Austronesian and Bantu population expansions. Proc. R. Soc. B. 278, 1399–1404.
Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.
Wheeler, W.C., 2003. Implied alignment. Cladistics 19, 261–268.
Wheeler, W.C., Whiteley, P.M., 2015. Historical linguistics as a sequence optimization problem: The evolution and biogeography of Uto-Aztecan languages. Cladistics 31, 113–125.
Wheeler, W.C., Lucaroni, N., Hong, L., Crowley, L.M., Varón, A., 2015. POY version 5: Phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31, 189–196.

Supporting Information

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

Data S1. All data notes and files as well as POY input scripts are available at https://wardwheeler.wordpress.com/data-sets/.