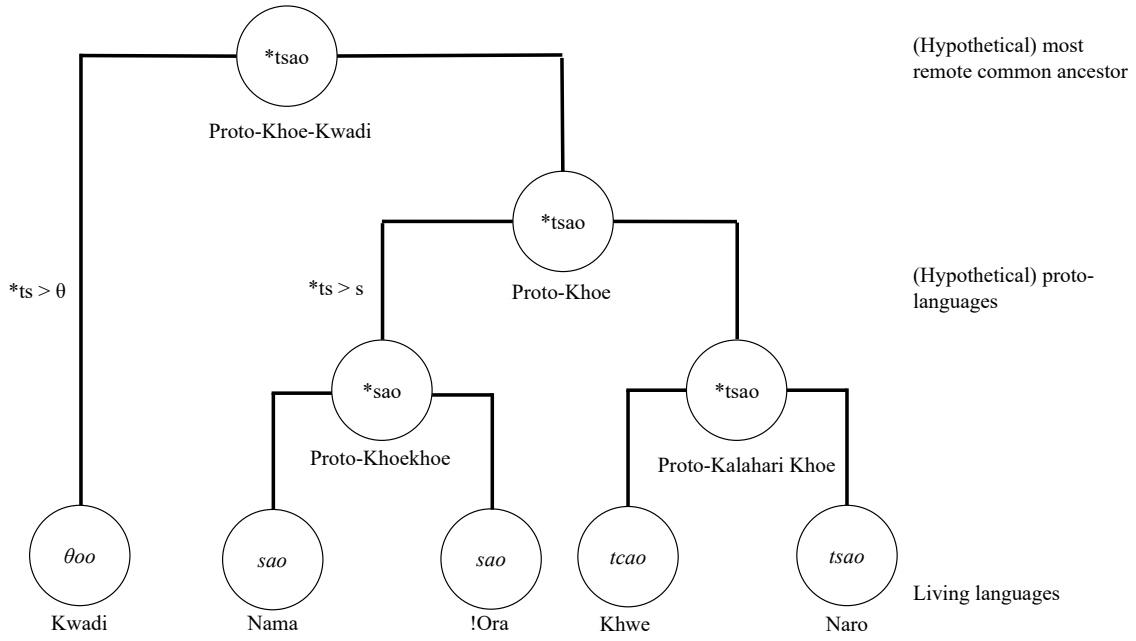
Supplementary Texts

Supplementary Text 1: The comparative method

To establish genealogical relations between languages, discipline of historical linguistics uses the so-called "comparative method" in which regular sound shifts and other kinds of shared innovations are used to derive family relationships, based on common ancestry, rather than superficial resemblance (Rankin, 2017). Following the establishment of cognate sets, i.e., sets of related words from different languages, the identification of regular correspondences between sounds leads to the reconstruction of so-called "roots", i.e. underlying ancestral forms from which all present day reflexes can be derived.

This is exemplified below with a root meaning 'tail, to follow' in the southern African language family Khoe-Kwadi (see Fig. 2C): among living languages, the forms *θoo* (Kwadi), *sao* (Nama), *sao* (!Ora), *tcao* (Khwe) and *tsao* (Naro) are attested (Westphal, no date a; Vossen 1997). If all of these are considered related, the most likely ancestral form is *tsao, whereas the star * highlights the hypothetical status of the reconstructed form. *tsao* was retained in Naro, and – safe for the palatalization (*ts > *tc*) – in Khwe. In Nama and !Ora, the alveolar affricate *ts changed to an alveolar fricative *s*. As the two languages share this innovative sound shift (among others), they can be grouped together in a subfamily we call Khoekhoe (Vossen, 1997). In Kwadi, *ts shifted to an interdental fricative *θ* (pronounced like English <th>), underlining the outlier status of this language within the family which is also supported by other types of data (Güldemann, 2004). Importantly, these sound shifts cannot only be observed for the root *tsao, but also for other Khoe-Kwadi roots which start with the same sound, like *tsoo 'medicine, to cure' or *tsãã 'hot, to burn, to shine'. Hence, we can talk about a regular sound shift. If enough regular sound correspondences between languages can be identified, it is commonly assumed that they form a family whose internal substructure can be established on the basis of shared innovations.



Supplementary Text 2: Bayesian Phylolinguistics

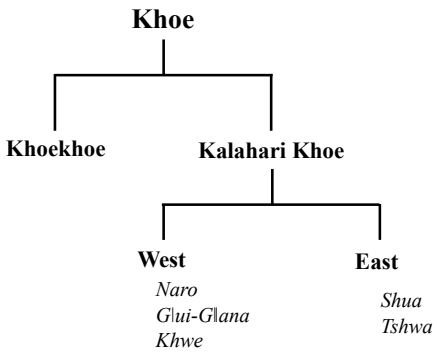
In recent years, computational methods have increasingly been applied to linguistic data. Especially Bayesian phylogenetic approaches have become popular tools to explore the substructure of linguistic family trees, assess the probability of individual subclusters, add dates and locate the data in a geographical framework, including the assessment of places of origin and migration routes (Dunn, 2014; Greenhill, Heggarty and Gray, 2020). The input file for the analysis of lexical data consists of binary coded cognate sets in which individual languages are coded for the presence or absence of a particular lexical root covering a specified meaning. This process – which presupposes extensive knowledge of regular sound shifts in the data to be analyzed – is exemplified below for the meaning 'thorn' in the Khoe-Kwadi language family (Westphal, no date a; Vossen 1997) (Fig. 2C):

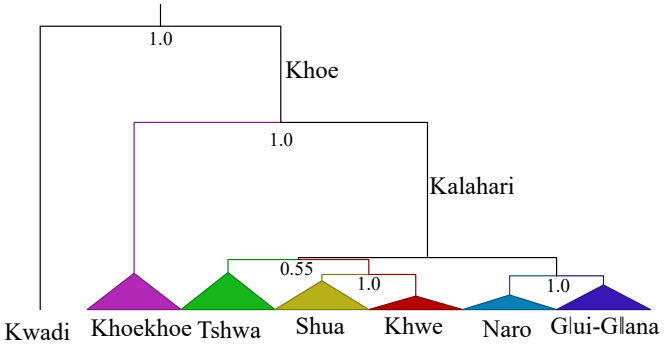
'thorn'

Kwadi	*fuu*	*ǁxuu
Nama	*ǁxuu*	
!Ora	*ǁxuu*	
Naro	*gǁam*	*gǁam
Khwe	*gǁam*	

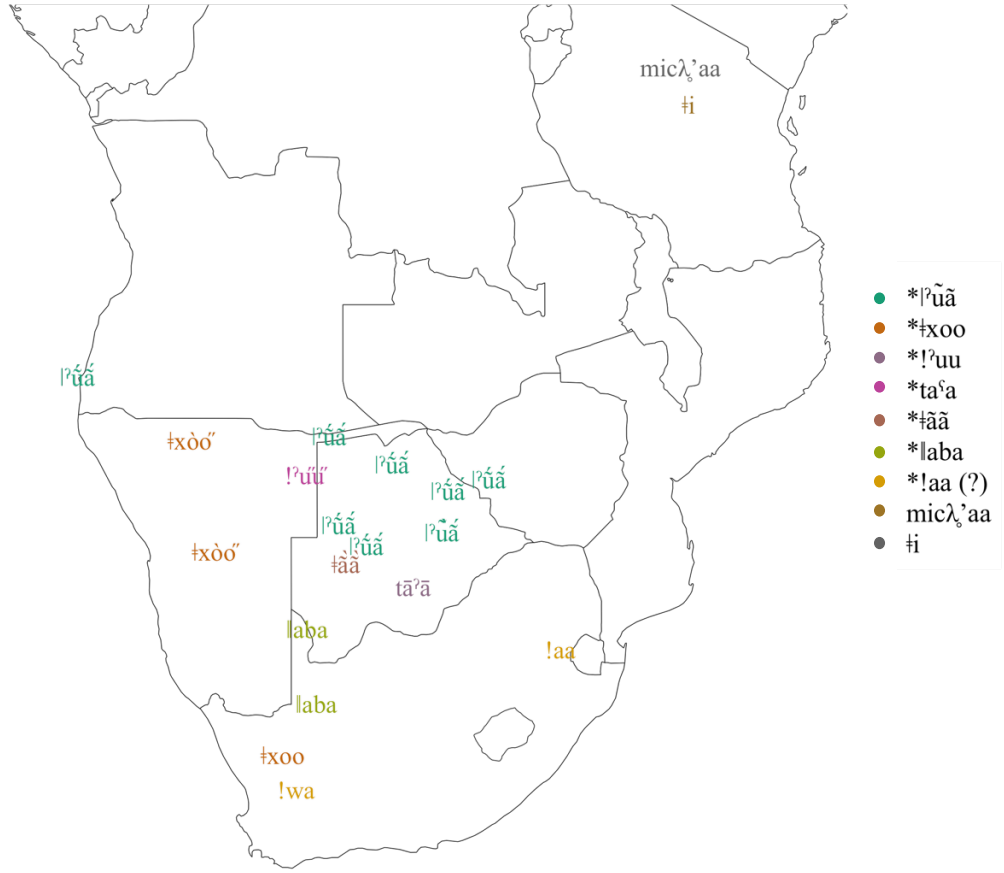
	Kwadi	Nama	!Ora	Naro	Khwe
*ǁxuu	1	1	1	0	0
*gǁam	0	0	0	1	1

Contrary to traditional distance-based lexicostatistics, Bayesian methods are capable of distinguishing retentions from innovations and are therefore particularly well suited to complement the historical-comparative method. Furthermore, they are capable of providing posterior probabilities for individual clusters, thereby highlighting uncertainty which may arise from horizontal transfer within a language family, or from non-tree like processes of diversification. In the example below, a tree of the southern African language family Khoe based on a traditional historical-comparative analysis (Vossen, 1997) is compared to a lexicon-based Bayesian analysis of more recent data from the same set of languages, using the remote relative Kwadi as an outgroup (Fehn et al., in prep). Note how the internal classification of Kalahari Khoe differs from the classical analysis: the Bayesian consensus tree does not show a division between Eastern and Western Kalahari Khoe, but strongly supports a link between Naro and G|ui-G‖ana spoken in the Central Kalahari (1.0), as well as between Khwe and Shua spoken along the northern Kalahari Basin fringe (1.0).

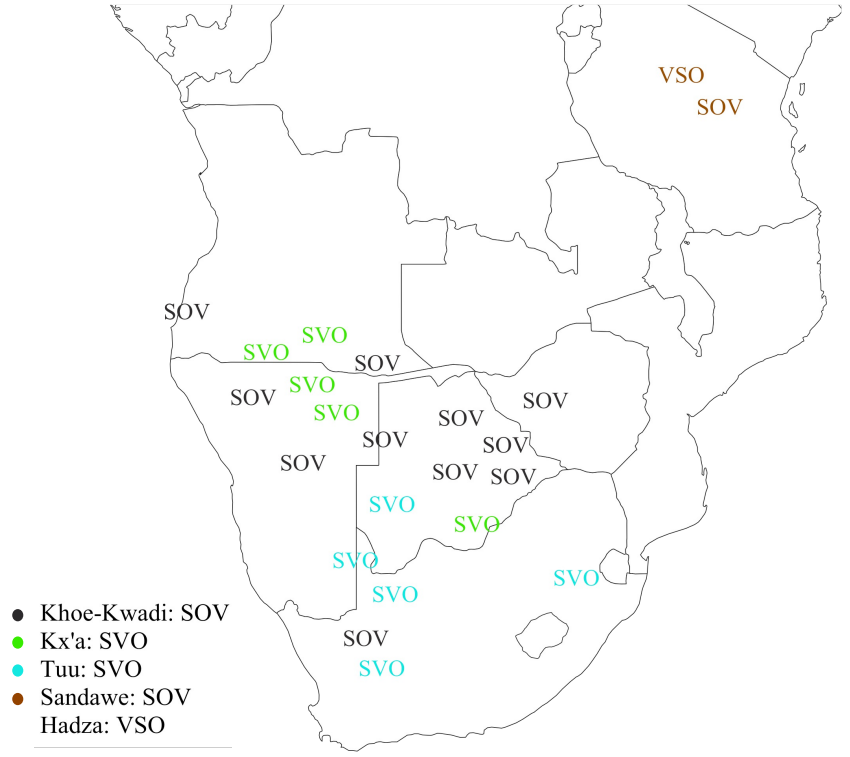




Supplementary Figures



Supplementary Fig. 1 - Words conveying the meaning 'bone' across Greenberg's "Khoisan family". Rather than going back to a single common ancestor, at least nine different roots can be identified. Forms given with a star * are historical reconstructions based on modern reflexes which existed in a hypothetical proto-language. Sources: Khoe-Kwadi: Westphal (no date a); Meinhof (1930); Haacke and Eiseb (2002); Vossen (1997); Phiri (2019); own data; Kx'a: Dickens (1994); König and Heine (2008); Gerlach (2016); Tuu: Bleek (1956); Traill (2018); Sands and Jones (2022); Sandawe: ten Raa (2012); Hadza: de Voogt (1992).



Supplementary Fig. 2 - Clausal word order patterns as an example for a typological feature setting Khoe-Kwadi apart from Kx'a and Tuu while providing a link with Sandawe (S – Subject, V – Verb, O – Object). Source: Vossen (2013); own data.

A

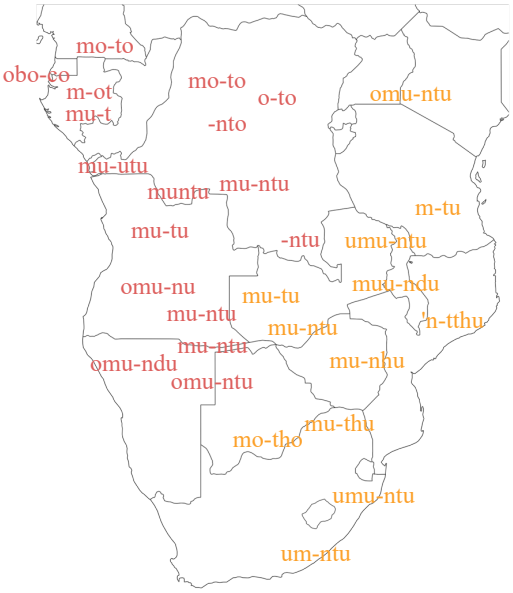
West Bantu

Duala (A24): mo-to
Bubi (A31): obo-co
Fang (A75): m-ot
Kele (B22): mu-t
Vili (B503): muu-tu
Lingala (C36): mo-to
Mongo (C61): -nto
Tetela (C71): o-to
Lega (D25): mon-to
Kimbundu (H21): mu-tu
Kikongo (H16): mu-ntu
Nyemba (K10): mu-ntu
Lozi (K21): mu-tu
Kwangali (K33): mu-ntu
Luba-Kasai (L31): mu-ntu
Sanga (L35): -ntu
Umbundu (R11): omu-nu
Ndonga (R22): omu-ntu
Herero (R31): omu-ndu

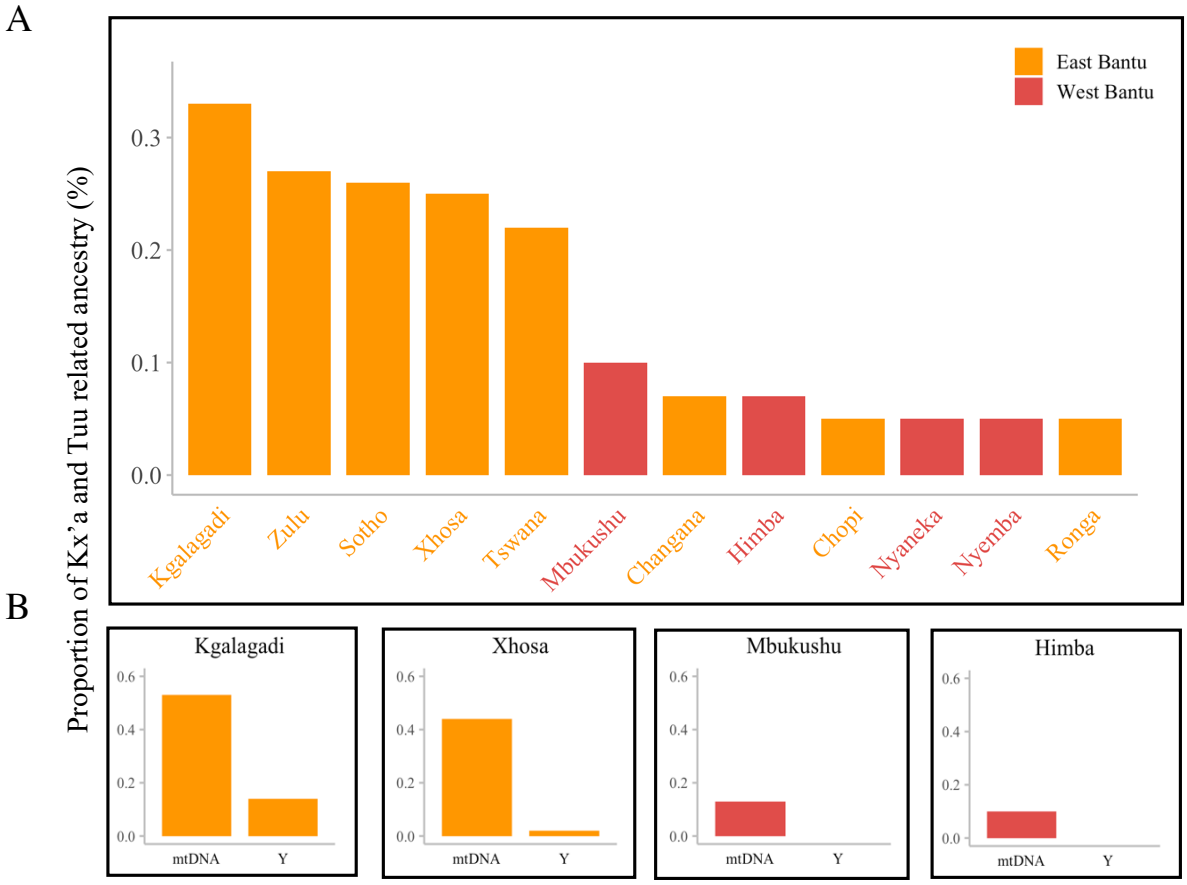
East Bantu

Kikuyu (E51): mo-ndo
Swahili (G42): m-tu
Rundi (JD62): umu-ntu
Ganda (JE15): omu-ntu
Bemba (M42): umu-ntu
Tonga (M64): mu-ntu
Yao (P21): muu-ndu
Makhuwa (P31): 'n-tthu
Shona (S11): mu-nhu
Venda (S21): mu-thu
Tswana (S31): mo-tho
Xhosa (S41): um-ntu
Zulu (S42): umu-ntu
Tsonga (S53): mu-nhu

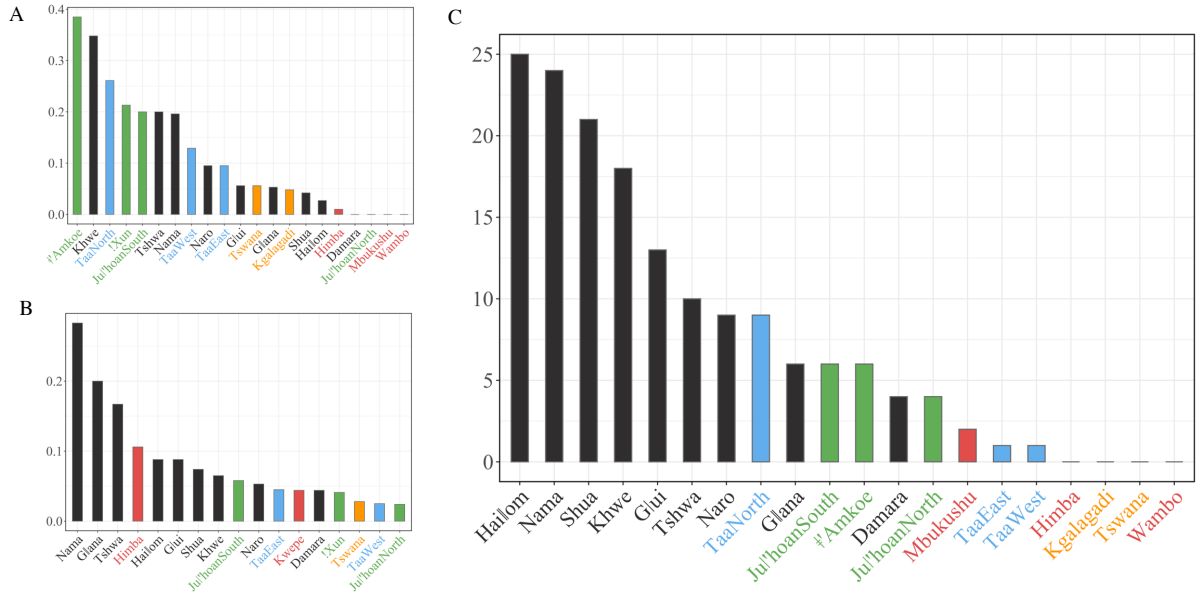
B



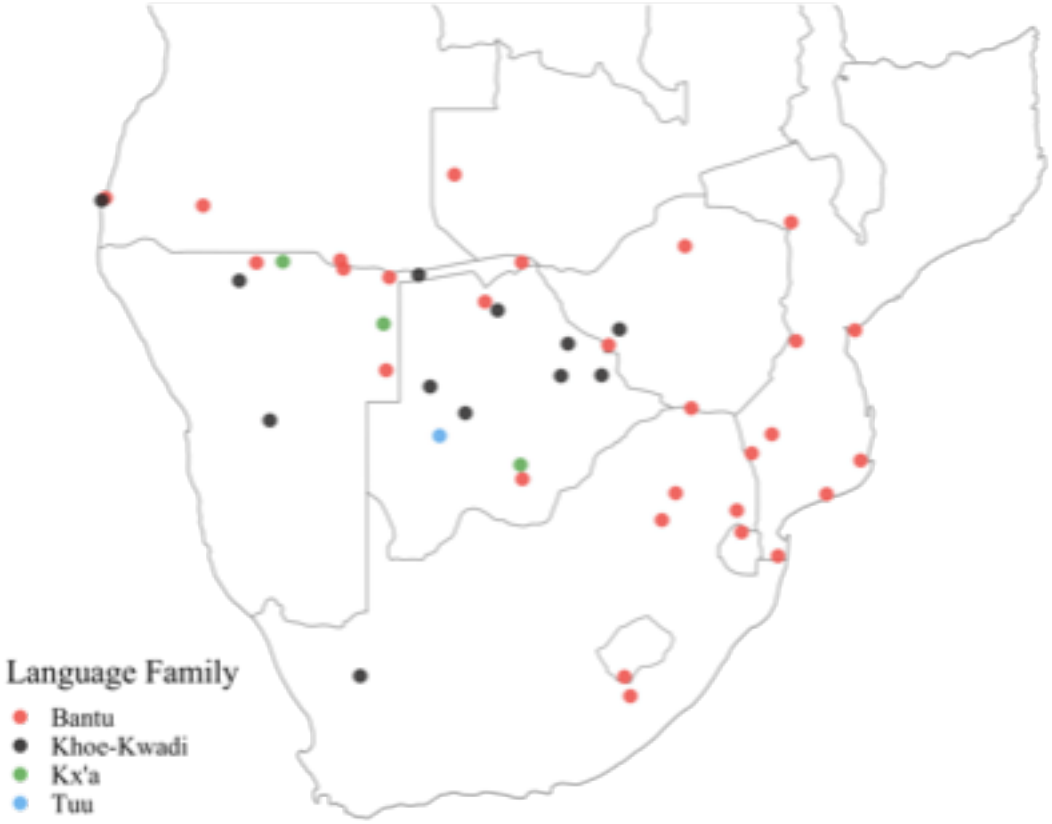
Supplementary Fig. 3 - (A) Examples for reflexes of the Proto-Bantu root *ntu ‘person’ in modern West (red) and East (orange) Bantu languages. The code following each language refers to the so-called Guthrie classification of Bantu which assigns an alphanumerical code to each language, according to its geographical location. **(B)** Geographical distribution of the reflexes of *ntu. Source: Grollemund et al. (2015).



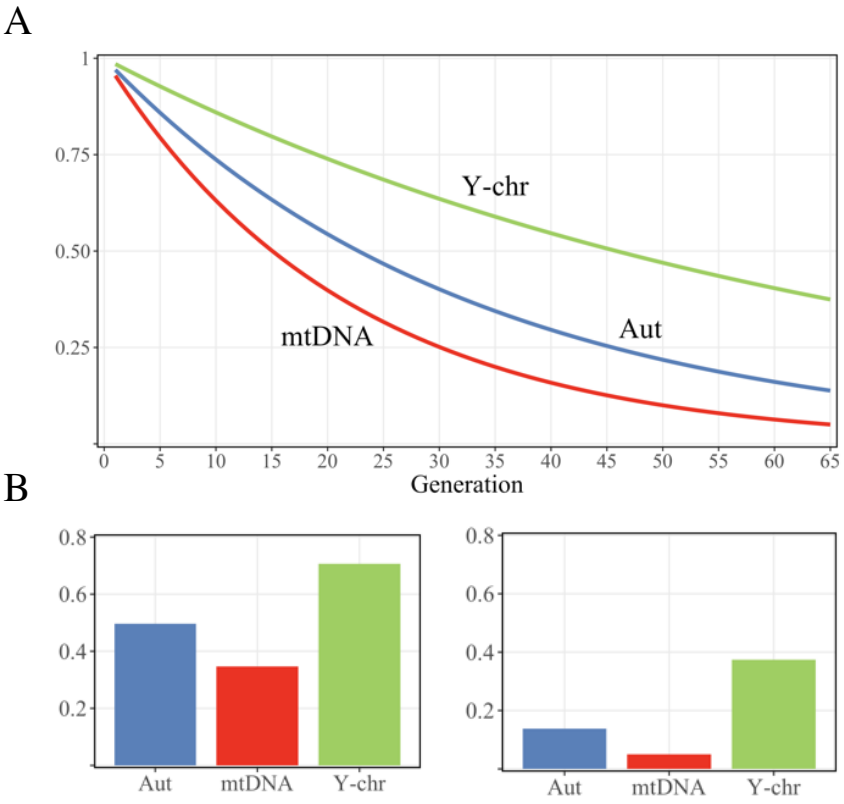
Supplementary Fig. 4 - (A) Autosomal ancestry related to Kx'a and Tuu-speaking groups in diverse East and West Bantu-speaking populations from southern Africa. **(B)** Kx'a and Tuu-related mtDNA and Y-chromosome ancestry in four Bantu groups from southern Africa. Sources: Pickrell et al. (2012); Barbieri et al. (2014); Marks et al. (2015); Bajić et al. (2018); Semo et al. (2020); Sengupta et al. (2021).



Supplementary Fig. 5 - Proportions of eastern African ancestry in southern African populations (colours according to Fig. 4). **(A)** Y-chromosome haplogroup E1b1b. **(B)** Lactase persistence -14010C mutation. **(C)** Autosomes. Sources: Pickrell et al. (2014); Breton et al. (2014) Macholdt et al. (2014); Pinto et al. (2016); Schlebusch et al. (2017); Bajić et al. (2018); Oliveira et al. (2019).



Supplementary Fig. 6 - Geographical distribution of the root *guu 'sheep' in Khoe-Kwadi, Kx'a , Tuu and Bantu. Sources: Khoe-Kwadi: Westphal (no date a, b); Dornan (1917); Meinhof (1930); Visser (2001); Haacke and Eiseb (2002); Kilian-Hatz (2003); Nakagawa (2014); Phiri (2019); own data; Kx'a: Dickens (1994); König and Heine (2008); Gerlach (2016); Tuu: Traill (2018); Bantu: Johnston (1919).



Supplementary Fig. 7 - (A) Over time decrease of eastern African ancestry in different genome compartments of an incoming Khoe-Kwadi group due to accumulated female-biased gene flow from a resident group. In each generation the incoming group receives 3% of its genes from the local group with a female-to-male ratio of 3:1. **(B)** Proportions of eastern African ancestry retained in autosomes (Aut), mtDNA and Y-chromosomes (Y-chr) after 23 (left-graphic) and 65 generations (right-graphic).

Supplementary References

Bleek DF (1956) A Bushman dictionary. New Haven: American Oriental Society.

de Voogt AJ (1992) Some phonetic aspects of Hatsa and Sandawe clicks. MA thesis, Rijksuniversiteit te Leiden.

Dickens P (1994) English-Ju|'hoan, Ju|'hoan-English dictionary. Rüdiger Köppe, Cologne.

Dornan SS (1917) The Tati Bushmen (Masarwas) and their language. J R Anthropol Inst 47(1):37-112.

Dunn M (2014) Language phylogenies. In: Bowern C, Evans B (eds) The Routledge Handbook of Historical Linguistics, Routledge, London, p. 190–211.

Gerlach L (2016) N!aqriaxe. The phonology of an endangered language of Botswana. Harrassowitz, Wiesbaden.

Greenhill SJ, Heggarty P, Gray RD (2020) Bayesian phylolinguistics. In: Janda RD, Joseph BD, Vance BS (eds) The Handbook of Historical Linguistics, Vol. 2, Wiley-Blackwell, Hoboken, New Jersey, p. 226-253.

Güldemann T (2004) Reconstruction through 'de-construction': the marking of person, gender, and number in the Khoe family and Kwadi. Diachronica 21(2):251-306.

Haacke WHG, Eiseb E (2002) A Khoekhoegowab Dictionary. Gamsberg Macmillan, Windhoek.

Johnston HH (1919) A comparative study of the Bantu and semi-Bantu languages. Clarendon Press, Oxford.

Kilian-Hatz C (2003) Khwe dictionary with a Supplement on Khwe Place-Names of West Caprivi by Matthias Brenzinger. Rüdiger Köppe, Cologne.

König C, Heine B (2008) A concise dictionary of Northwestern !Xun. Rüdiger Köppe, Cologne.

König C, Heine B (2008) A concise dictionary of Northwestern !Xun. Rüdiger Köppe, Cologne.

Meinhof C (1930) Der Koranadialekt des Hottentottischen. Zeitschrift für Eingeborenen-Sprachen, Beiheft 12. Dietrich Reimer, Berlin.

Nakagawa H (2014) G|ui Dictionary. Unpublished manuscript.

Phiri A (2019) Tjwao fieldnotes. Unpublished manuscript.

Rankin RL (2017) The comparative method. In: Joseph BD, Janda RD (eds) The handbook of historical linguistics, Wiley-Blackwell, Hoboken, New Jersey, p. 181-212.

Sands B, Jones K (2022) N|uuki – Namagowab – Afrikaans – English dictionary. African Sun Media.

ten Raa E (2012) A dictionary of Sandawe. The lexicon and culture of a Khoesan people of Tanzania, ed. by Christopher & Patricia Ehret, in collaboration with Edward D. Elderkin. Rüdiger Köppe, Cologne.

Traill A (2018) A trilingual !Xóõ dictionary. Rüdiger Köppe, Cologne.

Visser H. 2001. Naro Dictionary: Naro-English, English-Naro. Naro Language Project, D'Kar.

Vossen R (1997). Die Khoe-Sprachen. Ein Beitrag zur Erforschung der Sprachgeschichte Afrikas. Rüdiger Köppe, Cologne.

Vossen R (2013) (ed) The Khoesan Languages. Routledge, London.

Westphal EOJ (no date a) Kwadi fieldnotes and recordings. University of Cape Town Archives.

Westphal EOJ (no date b) G‖abak'e fieldnotes and recordings. University of Cape Town Archives.