Anne-Maria Fehn

Tracing invisible footsteps: towards a multidisciplinary model for the spread of Khoe-Kwadi languages in pre-colonial southern Africa

Ana Beatriz Silva Amorim

Mestrado em Biodiversidade, Genética e Evolução

Departamento de Biologia

2020

Orientador

Dr. Jorge Macedo Rocha, Professor Associado, FCUP

Coorientador

Dr. Anne-Maria Fehn, Investigador Auxiliar, CIBIO-InBIO

Dr. Magdalena Gayà Vidal, Investigador Auxiliar, CIBIO-InBIO

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Assinado por : **ANTÓNIO EMÍLIO FERRAND DE ALMEIDA MÚRIAS DOS SANTOS**
Num. de Identificação: BI073858331
Data: 2021.01.21 16:25:48 +0000

CARTÃO DE CIDADÃO

Porto, _____/_____/_____

"Num universo de sim ou não, branco ou negro, eu represento o talvez."
— Pepetela (Mayombe)

Acknowledgements

I first want to acknowledge the tremendous help from my supervisors that allowed me to finish this thesis with confidence and pride in my work. That being so, thank you, Professor Jorge; your enthusiasm and broad knowledge of genetics and anthropology impress and encourage me to improve myself as a scientist. You showed me what it is like to be passionate about your work. Thank you, Anne; without you, I would not know that the ancestors of the Ts'ixa were lions, and, most importantly, I would not have read any Neil Gaiman books. You made me feel like I belonged here, talking about these people whom I have never visited. I also want to thank Magda for always being available to help in all fields and empathizing with my pain of trying to sequence mtDNA.

I would like to acknowledge the help of the HUMANEVOL group: Dr. Bérenice Alard, who was responsible for making the ancestry-assignment analysis that allowed to make the plots displayed here, and for sharing her knowledge on not only clans but also on cool Netflix shows. I thank Dr. Sandra Oliveira, as well, who was responsible for making the masking analysis and the PCAs, helped to develop our IBD analysis and was always ready to answer my questions.

I want to thank Dr. Dang Liu, from the MPI in Leipzig, who, without ever meeting me, helped me with my IBD analysis significantly. The same goes for Dr. Simon Greenhill, from the MPI in Jena, for answering my very long questions on computational linguistics.

I would like to acknowledge Admire Phiri (Department of Linguistics - University of Stellenbosch) and Dineo Peke (San Research Institute – University of Botswana). Although I still did not get the chance to visit you in Botswana and Zimbabwe, I want to thank you for all the time we spent here in Porto, and all the knowledge you brought me on Khoisan culture.

Even though my work in the lab was cancelled, I want to send my love to the CTM lab people for sharing their experience and answering my many questions. Especially Susana, who supported me emotionally and was ever-present in helping to develop our proto-col.

I would not be here without the support of my friends. In particular, Valeria, who always told me that I was worth it and that I could do this. Thank you for being the person that slaps me every time I say something stupid. Also, Sofia, thank you for being my R master and showing me the magic of a for loop and the countless meals at your apartment. Finally, Dani, my SPN soulmate, thank you for the editing, and for reporting missing commas.

I want to mention the support from my co-workers at Marques Soares, who always believed in me, even when I did not. Also, thank you for letting me use my computer at work.

I also want to recognize the emotional support of my family. Thank you, dad, mom, Laika, and especially Filipa, for pretending to be interested in my work and keeping me company during skype calls.

Finally, I want to express gratitude to the music of Phoebe Bridgers, Father John Misty and SZA for emotional support during the darker times.

Resumo

Existem três camadas populacionais associadas a diferentes linhagens linguísticas na Africa Austral pré-colonial. A camada mais antiga, pertence a caçadores-recolectores que falam as línguas Kx'a e Tuu, comummente designadas por "Khoisan". Por volta de 1500BP, duas ondas independentes de agropastoralistas da idade do Ferro, com origem na África Ocidental, introduziram línguas da família Bantu, hoje o filo linguístico mais difundido na África Austral. Uma terceira camada surgiu há cerca de 2000BP, estando associada a grupos de pastores com uma origem presumida na África Oriental, que se pensa terem introduzido as línguas da família Khoe-Kwadi na região.

Este estudo foca-se nos vários grupos etnolinguísticos que falam Khoe-Kwadi, com o propósito de testar hipóteses anteriormente propostas sobre as migrações de povos pastoris originários da África Oriental, e a consequente difusão de línguas, práticas culturais e património genético em diversos locais da África Austral. Seguindo uma abordagem multidisciplinar, foram usados dados linguísticos, recolhidos durante trabalho de campo em Angola, Namíbia, Botswana e Zimbabwe, assim como marcadores genéticos uni-parentais e dados de variação genómica de populações sul-africanas, incluindo informação ainda não publicada sobre comunidades do deserto do Namibe em Angola.

Uma análise filogenética Bayesiana usando dados lexicais, que cobrem toda a família de línguas Khoe-Kwadi, revelou uma árvore filogenética com ramificação tripartida, representando os sub-grupos Kwadi, Kalahari Khoe e Khoekhoe. Comparando estas informações com a distribuição geográfica destas línguas, verificou-se que o seu centro de difusão se encontra na fronteira entre a Namíbia e o Botswana, contradizendo hipóteses anteriores que colocavam este centro no nordeste do Botswana ou no Zimbabwe.

Apesar de as línguas Khoe-Kwadi descenderem de um ancestral comum, as populações Khoe-Kwadi não formam um grupo geneticamente diferente dos restantes grupos que habitam a região. Uma análise detalhada da variação genética dos povos falantes de Khoe-Kwadi revelou perfis genéticos semelhantes aos Kx'a, Tuu e aos seus vizinhos Bantu. Alguns grupos que habitam a periferia da bacia do Kalahari apresentaram níveis elevados de miscigenação, partilhando segmentos genómicos com populações "Khoisan" e Bantu. No entanto, os grupos Khoe-Kwadi do centro do Kalahari foram os menos afetados pela migração Bantu.

Apesar da elevada miscigenação, foram detetados vestígios de uma contribuição genética com origem na África Oriental, que se encontra em quantidades variáveis em todas as populações falantes de Khoe-Kwadi. Esta contribuição genética também foi observada nos Kwepe, antigos falantes de Kwadi, e noutros grupos de menor dimensão do Namibe de Angola, cujo perfil genético é partilhado com falantes de Bantu da mesma área. Este

componente genético da África Oriental foi encontrado em pastores da mesma região e, adicionalmente, numa amostra de DNA antigo proveniente de um local arqueológico associado a comunidades pastoris que viviam na Tanzânia há 3100BP. O padrão deste componente genético nas atuais populações sul-africanas sugere uma introdução do pastoralismo na região da África Austral por um grupo migrante oriundo do Este.

Em suma, as evidências apontam para uma associação entre as línguas Khoe-Kwadi, a prática de pastoralismo na África Austral e vestígios de um património genético da África Oriental. Após a chegada à fronteira entre a Namíbia e o Botswana, as comunidades pastoris, com origem inicial na África Oriental, divergiram e interagiram com grupos falantes de Kx'a e Tuu, assim como com os agricultores que chegavam das migrações Bantu. Embora o perfil completo dos pastores migrantes tenha sido mais bem preservado nos pastores Khoekhoe da África do Sul e Namíbia, as interações entre populações na zona de contacto na África Austral produziram várias combinações de traços culturais, genes e línguas nas populações atuais que falam Khoe-Kwadi.

Palavras-chave

Pré-história Africana; Estudos do autossoma; Genética humana; Khoe-Kwadi; Linguística; Filogenética; África Austral.

Abstract

For pre-colonial southern Africa, three major population layers associated with distinct linguistic lineages can be distinguished: the oldest layer is constituted by hunter-gatherers speaking Kx'a and Tuu languages. By 2000BP, Late Stone Age pastoralists with a presumed eastern African origin arrived in the area; they are thought to have introduced languages of the Khoe-Kwadi family. Around 1500BP, two independent waves of Iron Age agropastoralists from western Africa introduced languages of the Bantu family, which now constitutes the most widespread linguistic phylum in southern Africa.

In this work, we focus on ethnically diverse populations speaking languages of the Khoe-Kwadi family in order to test longstanding hypotheses about a pastoralist migration from eastern into southern Africa, leading to the diffusion of languages, cultural practices, and genetic material. Following a multidisciplinary approach, we use linguistic data collected during fieldwork in Angola, Namibia, Botswana, and Zimbabwe, as well as a comprehensive dataset of genome-wide and uniparental markers from southern African populations, including unpublished data on newly genotyped populations from the Angolan Namibe.

A Bayesian phylogenetic analysis of lexical data covering the full geographic range of the Khoe-Kwadi language family reveals a tripartite branching pattern associated with the Kwadi, Kalahari Khoe, and Khoekhoe subgroups, respectively. When geography is taken into account, the centre of diffusion in southern Africa is placed in the border area between Namibia and Botswana, contradicting previous hypotheses which located the split-off point in north-eastern Botswana or Zimbabwe.

Even though Khoe-Kwadi populations are united by their linguistic ancestry, they do not share an apparent genetic patrimony. An in-depth analysis of autosomal data from modern Khoe-Kwadi speakers reveals genetic profiles similar to those of their Kx'a, Tuu ("Khoisan"), and Bantu-speaking neighbours. Some groups from the Kalahari Basin fringe display high amounts of admixture involving sharing of ancestral segments with both "Khoisan" and Bantu populations, pointing to not detected intensive contact in Khoe-Kwadi speakers from the central Kalahari who are exclusively "Khoisan". However, an eastern African contribution is present at varying amounts in all Khoe-Kwadi speaking populations, including in the formerly Kwadi-speaking Kwepe and other small-scale populations from the Angolan Namibe. They otherwise share a genetic profile with Bantu speakers from the same area. A similar genetic component was found in modern pastoralists from eastern Africa, as well as in an ancient DNA sample from a herding site in Tanzania, suggesting its introduction to southern Africa by a migrant group.

Taken together, our results show an association between Khoe-Kwadi languages, pastoralism, and an eastern African genetic heritage. After their initial migration into southern

Africa, the pastoralists started to diverge and interact with Kx'a and Tuu-speaking groups, as well as with incoming Bantu farmers from the eastern and western branches of the Bantu migrations. While the full package (pastoralism, eastern African ancestry, and Khoe-Kwadi linguistic patrimony) was primarily retained in the Khoekhoe herders of South Africa and Namibia, interaction in a contact zone led to the diverse combinations of culture, genes, and languages characteristic of modern Khoe-Kwadi speakers.

Keywords

African Prehistory; Autosomal studies; Human genetics; Khoe-Kwadi; Linguistics; Phylogenetics; southern Africa

Table of Contents

List of Tables

Supplemental tables:

List of Figures

Supplemental Figures:

List of Abbreviations

ABC	Approximate Bayesian Computation
AD	*Anno Domini*
BC	Before Christ
BP	Before Present
DNA	Deoxyribonucleic Acid
HBD	Homozygous by Descent
IBD	Identity by Descent
logML	Logarithmic Marginal Likelihood
LP	Lactase Persistence
LSA	Late Stone Age
ML	Maximum Likelihood
mtDNA	Mitochondrial DNA
NRY	Non-recombining Region of the Y-chromosome
SNP	Single Nucleotide Polymorphism
stdev	Standard Deviation
SVO	Subject–Verb–Object
TMRCA	Time for the Most Recent Common Ancestor
UV	Ultraviolet
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

[Page intentionally left blank]

Chapter 1: Introduction

Archaeological and fossil records support that modern humans evolved in Africa during the Middle Stone Age, between 300,000 and 190,000 years ago, having lived continuously on the African continent longer than in any other geographic region (Behar *et al.*, 2008; Campbell & Tishkoff, 2010; Salas *et al.*, 2002; Schlebusch *et al.*, 2017). Within Africa, the eastern part of the continent is assumed to have been the starting point for the so-called "Out of Africa" migration, with modern humans crossing the Bab-el-Mandeb strait near the Red Sea and subsequently moving along the coastline of south-east Asia towards Australia (Campbell & Tishkoff, 2010; Mellars, 2006; Quintana-Murci *et al.*, 1999; Stanyon *et al.*, 2009). This model became widely accepted when it was shown that the mitochondrial DNA variation present in Eurasia, Oceania and the Americas was merely a subset of the variation found in Africa (Campbell & Tishkoff, 2010; Cann *et al.*, 1987; Ingman *et al.*, 2000; Mellars, 2006; Quintana-Murci *et al.*, 1999; Stanyon *et al.*, 2009; Tishkoff *et al.*, 2009).

African populations not only hold the greatest genetic diversity among all modern humans but also display considerable linguistic, environmental and cultural variation (Campbell & Tishkoff, 2010; Cann *et al.*, 1987; Montinaro *et al.*, 2017; Ramachandran *et al.*, 2005): Africans live in diverse environments (savanna, mountain highlands, deserts, etc.) which have undergone dramatic changes throughout human evolution; they practise different means of subsistence, such as hunting, gathering, pastoralism and agriculture and represent 2000 distinct ethnolinguistic groups, who speak almost a third of the world's languages. The demographic history of Africans has not been linear over time: oscillations in population size and the influence of various migration and admixture events that often correlated with significant environmental changes resulted in intricate patterns of variation across modern populations (Campbell & Tishkoff, 2010). In response to these events, novel genetic and phenotypic adaptations in essential diet genes (lactase persistence, starch hydrolysis, etc.) and exposure to infectious diseases have evolved in the past thousand years. In consequence, present-day patterns of variation in African genomes stems from both demographic and selective occurrences (Campbell & Tishkoff, 2010; Tishkoff *et al.*, 2009).

While Africa, in general, is considered to be a hotspot of scientific research, a particular focus has been on the broader region of southern Africa. For tens of thousands of years, southern Africa was inhabited by hunter-gatherer populations, followed by pastoralists and agriculturalists, who moved into the area from around 2,000 BP (Pickrell *et al.*, 2014). The autochthonous hunter-gatherers of southern Africa, sometimes also referred to as "San" peoples or "Bushmen", are of particular interest to the study of human genetics, as their genomes host some of earliest branching mitochondrial DNA and Y-chromosome lineages present in modern humans. Their extensive genetic, linguistic and cultural diversity is currently

thought to represent the deepest historical population divergence among extant human populations (Bajić *et al.*, 2018; Montinaro *et al.*, 2017; Schlebusch *et al.*, 2012; Skoglund *et al.*, 2017).

Ethnolinguistically, the pre-colonial population landscape of southern Africa can be subdivided into two main groups associated with distinct linguistic profiles: Speakers of "Khoisan" languages presumed to be indigenous to southern Africa (§1.1), and Bantu (Niger-Congo) speakers associated with an Iron Age migratory wave covering major parts of sub-Saharan Africa (§1.2). A third layer linked to a pre-Bantu introduction of Late Stone Age pastoralism into the Kalahari Basin area has tentatively been associated with speakers of the "Khoisan" language phylum Khoe-Kwadi (§1.3). In this work, a multidisciplinary methodology combining new genomic and linguistic data from southern African populations will be used to provide a novel approach to assess the eastern African heritage and dispersal routes of the Khoe-Kwadi (§1.4).

1.1 The Khoisan languages and their speakers

Before being used to designate a presumed language family (Greenberg, 1963), the term "Khoisan" was coined by the physical anthropologist Leonhard Schulze in the 1920s to refer to the shared phenotypical appearance of south African hunter-gatherers and herders. It derives from the Khoe-Kwadi language Khoekhoe, spoken by herders, and combines the autonym [khoe], meaning "person", with the exonym [saa], meaning "gather(er)". At present, the term "Khoisan" is often used to designate both the non-Bantu languages of southern Africa and the people who speak them, without implying genetic or linguistic unity (Güldemann, 2014).

"Khoisan" languages are united by a shared set of typological features, the most prominent one being a heavy reliance on click sounds. While clicks are a common feature in communicative behaviour all over the world, accentuating feelings like disapproval, irritation, and regret, only African "Khoisan" languages make use of clicks as phonemic consonants (Güldemann & Stoneking, 2008). Within Africa, the highest proportion of click sounds (~50%) is found in "southern African Khoisan", which can be grouped into three distinct language families: Kx'a (Northern Khoisan), Tuu (Southern Khoisan) and Khoe-Kwadi (Central Khoisan) (Güldemann, 2014). The people now speaking those languages are thought to descend from some of the earliest humans inhabiting the wider region of southern and eastern Africa who used to be hunter-gatherers and pastoralists.

The existing similarities between the "Khoisan" languages of southern Africa are commonly seen as a reflection of a long history of contact in a restricted area consensually referred to as the 'Kalahari Basin' (**Figure 1**). In this region, archaeological evidence has supported the continuous presence of foragers since the Late Stone Age (30,000 BP), with a

pre-Bantu introduction of domestic livestock around 2,000 BP (Barnard, 1992; Güldemann, 2008). Ethnographical data confirms that immediate ancestors of the modern "Khoisan" speakers lived in small bands of hunter-gatherers and bigger pastoralist groups, both of whom had a partly mobile lifestyle (Güldemann, 2008).

Within the 'Khoisan' languages of southern Africa, there is a major typological split between the Khoe-Kwadi languages and the Kx'a (Northern Khoisan) and Tuu (Southern Khoisan) families (**Table 1**). As Kx'a and Tuu share several structural features (besides clicks), they are usually grouped as "Non-Khoe" vs the more diverse Khoe-Kwadi languages (Güldemann, 2008, 2014). Speakers of the Kx'a language family (also known as Ju–ǂHoan) are mainly found in the north-western Kalahari (north-east Botswana, northern Namibia and southern Angola). They are considered to be the autochthonous people of the region they inhabit. Kx'a, meaning 'earth, ground', is a shared word between the two branches of the family, ǂ'Amkoe and Ju (!Xun) (Heine & Honken, 2010). Tuu languages, although less widespread today than those of the Kx'a family, were once spoken by populations situated in most of South Africa and are still retained in some regions of Botswana, Namibia and northern South Africa. The Tuu language family can be further divided into Taa-Lower Nossob and the !Ui branch (Güldemann, 2014).

While the Non-Khoe languages are exclusively spoken by (former) hunter-gatherers, Khoe-Kwadi is spoken by different ethnolinguistic groups that practise diverse subsistence strategies (Barnard, 1992; Güldemann, 2008). These groups are distributed across a large geographic area which includes not only the Kalahari Basin but also western Namibia, the Okavango River Delta and the vast salt pans of central Botswana (Cashdan, 1986). The Khoe-Kwadi family is assumed to have arrived late to the Kalahari Basin area. After contact with languages of the Kx'a and Tuu families, Khoe-Kwadi supposedly adopted many of their linguistic features while in exchange contributing some of their own, leading to a wide range of shared linguistic features across the entire area (Güldemann & Fehn, 2017). Khoe-Kwadi can be subdivided into two major branches: Kwadi, a now-extinct language spoken exclusively by small-stock keepers in south-western Angola, and Khoe (Central Khoisan), spoken in Namibia, Botswana, South Africa and Zimbabwe (Güldemann, 2008, 2014). Khoe is constituted of the Khoekhoe language complex (Nama-Damara, !Ora, Haillom-ǂAakhoe), spoken by both pastoralists and foragers, and Kalahari Khoe, including an eastern (Shua, Tshwa) and a western branch (Khwe, G|ui-Gǁana, Naro) (Vossen, 1997).

Despite their linguistic proximity, Khoe-Kwadi-speaking people are phenotypically diverse, with north-eastern populations resembling their Bantu-speaking neighbours (average taller stature and darker skin pigmentation), and south-eastern populations resembling Non-Khoe Khoisan groups (average light skin pigmentation and relatively short stature) (**Figure 2**)

(Güldemann, 2008; Güldemann & Stoneking, 2008). The Kwepe, Damara, Khwe, Shua and Tshwa, all from the northern and north-eastern fringes of the Kalahari Basin, are examples of groups that do not match the stereotypical appearance southern African “San” peoples but rather resemble their Bantu-speaking neighbours in both appearance and attire (Güldemann & Stoneking, 2008).

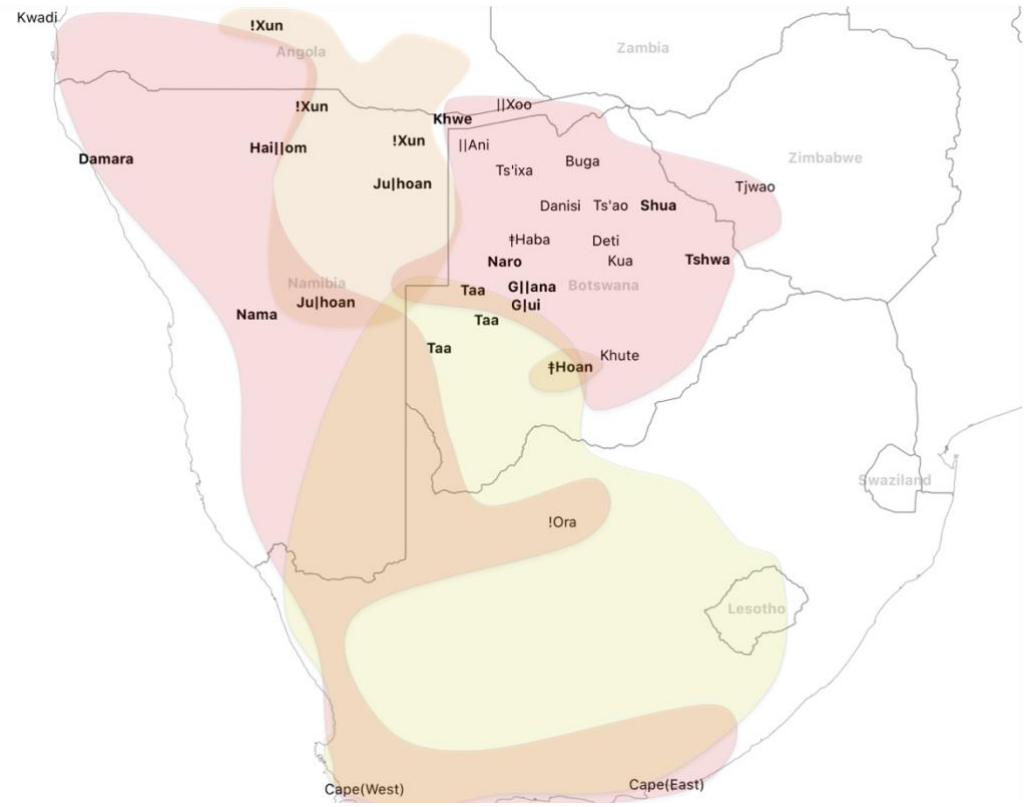

Figure 1 - Distribution of Khoisan Lineages and populations: Khoe-Kwadi (red), Kx'a (orange) and Tuu (yellow). Adapted from Güldemann (2014).

Table 1 - Khoisan lineages and their respective languages and dialects. Adapted from Güldemann (2014). Legend: † = extinct, º = moribund, () = older data sources.

Lineages and *(sub)branches*	Languages (L) or Language Complexes (LC)	Dialects and Dialect Groups
EASTERN AFRICAN KHOISAN		
Hadza	Single L	
Sandawe	Single L	
SOUTHERN AFRICAN KHOISAN		
Khoe-Kwadi		
Kwadi †	Single L	
Khoe		
Kalahari Khoe		
east	Shua	Danisi, Deti, Ts'ao, Ts'ixa, etc.
	Tshwa	Tcire, Tjwao, Tsua, Kua, etc.
west	Khwe	ǁAni, Buga, ǁXoo, etc.
	Gǁana	Gǁana, G|ui, Khute, etc.
	Naro	ǂHaba, Naro, etc.
Khoekhoe	(Cape K.)† LC	
	(!Ora-Xiri)º LC	
	(Eini)† LC	
	Nama-Damara LC	
	Haiǁom	
	ǂAakhoe	
Kx'a		
Ju	Single LC	north: !Xuun (Angola) north-central: !Xuun (Ekoka, Okongo) central: !Xuun (Grootfontein) south-east: Ju|'hoan
ǂ*'Amkoe*	Single LC	west: ǂHoan, N!aqriaxe east: Sasi
Tuu		
Taa-Lower Nossob		
Taa	Single LC	west east
Lower Nossob	(|'Auni)†	
	(|Haasi)†	
!Ui	Nǁng	N|uu=(ǂKhomani)
	(|Xam)†	Strandberg, Katkop, Achterveld, etc.
	(ǂUngkue)†	
	(ǁXegwi)†	

Figure 2 - Modern Khoe-Kwadi speakers show a great variation in appearance and attire (Courtesy of AM Fehn).

1.2 The Bantu migrations

The Bantu languages, whose name derives from the shared root *-ntu – 'person', belong to the Niger-Congo phylum and constitute the most widely distributed language family in sub-Saharan Africa (Bostoen, 2018; Rocha & Fehn, 2016) (**Figure 3**). A recent study estimates that about 310 million people speak Bantu, with one in three Africans speaking one or more of the 556 Bantu languages existing today (Bostoen, 2018). It is believed that the origin of Bantu lies in the Cross River Valley (the borderland between south-eastern Nigeria and western Cameroon), due to the fact that the closest relatives of Bantu within Niger-Congo are spoken there (Greenberg, 1972; Rocha & Fehn, 2016).

Linguistically, Bantu languages can be divided into two major branches: i) East Bantu, a monophyletic subgroup including all Bantu languages of eastern and south-eastern Africa, and ii) the more diverse West Bantu subgroup (Bostoen, 2018; Montinaro *et al.*, 2017; Rocha & Fehn, 2016; Salas *et al.*, 2002), which includes both the north-western Bantu languages of west and west-central Africa, as well as the south-western Bantu languages of Angola, Namibia and Zambia. The current distribution of Bantu languages is usually explained by two competing models (**Figure 3**): according to the so-called "early split" hypothesis, the eastern branch resulted from a movement of Bantu-speakers along the northern fringe of the rainforest towards the Great Lakes (east Africa) and then southwards to south-east Africa. In the same framework, the western branch underwent an initial move towards the south, followed by dispersals across the western part of subequatorial Africa (Rocha & Fehn, 2016). In the alternative model, referred to as the "late split" hypothesis, eastern Bantu diverged from western Bantu only after Bantu-speakers had crossed the rainforest (Rocha & Fehn, 2016). In recent years, the latter hypothesis has received increasing support from various disciplines, including linguistics (Currie *et al.*, 2013; Grollemund *et al.*, 2015) and genetics (Patin *et al.*, 2017; Semo *et al.*, 2020).

Irrespective of the routes their speakers took, the wide distribution and overall similarity of the Bantu languages suggests that their present distribution must be the result of a relatively recent and rapid diversification from a common ancestor, now known as the Bantu expansion (Rocha & Fehn, 2016). As this major demic movement across vast parts of sub-Saharan Africa is strongly associated with the spread of domestic livestock, crop farming and iron working, it has been considered a prime example for the shared expansion of languages, lifestyle and genes (Diamond & Bellwood, 2003). In southern Africa, Bantu languages were introduced by two migratory waves of Iron Age farmers who entered the area from the west and east around 1500 years BP. While Bantu speakers widely replaced the autochthonous inhabitants, both linguistic and genetic data suggest intensive contact with "Khoisan" foragers and herders in

some areas of South Africa, Botswana and Namibia (Barbieri, Güldemann*, et al.*, 2014; Barnard, 1992; Rocha & Fehn, 2016).

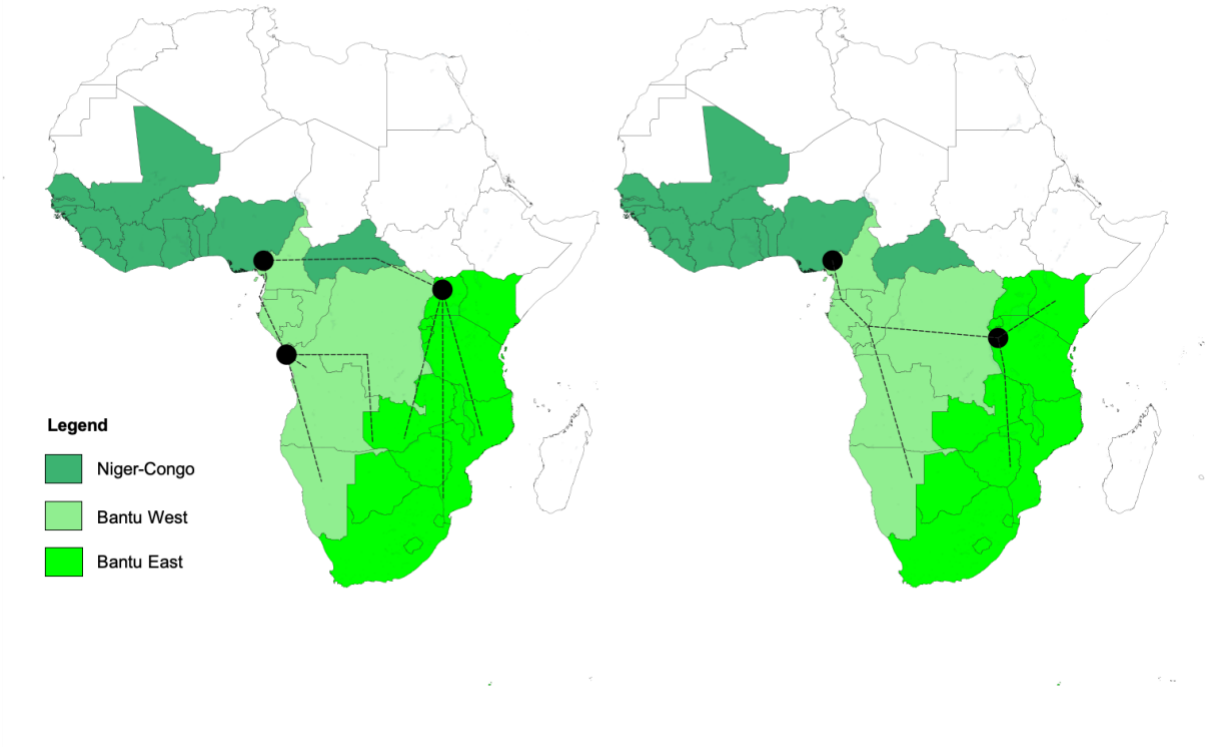

Figure 3 - Distribution of Niger-Congo languages and proposed migration routes of the Bantu speakers. The distribution of non-Bantu Niger-Congo, West Bantu and East Bantu languages are coloured in green and were adapted from Rocha and Fehn (2016). The migration routes according to the early (left) and the late (right) split (dashed lines) models were adapted from Pakendorf *et al.* (2011).

1.3 The Khoe-Kwadi as pre-Bantu pastoralists

Archaeological evidence supports an introduction of pastoralism in southern Africa before the arrival of the first Bantu-speaking farmers. These first food producers are commonly characterized as a Late Stone Age (LSA) culture which practised livestock herding and moulded pottery (Güldemann, 2008; Sadr, 2015), thereby leaving pot shards and domesticated animal remains in the coastal regions of South Africa, Namibia and northern Botswana (Lander & Russell, 2020; Sadr, 2013, 2015; Smith, 2017).

The earliest livestock remains in southern Africa have been dated to the first centuries BC and AD in LSA contexts from northern Botswana and the western and southern coast of South Africa (Sadr, 2013), predating the arrival of Iron Age farmers by two or three centuries (Sadr & Sampson, 2006). The earliest evidence for pottery is found at Leopard's Cave (Namibia) and has been dated to around 2500-2300 BP (Pleurdeau *et al.*, 2012). While pottery also appears at forager sites, there is evidence for its association with livestock: for example, at Bambata Cave (Matobo Hills, Zimbabwe) (Burrett, 2007; Mitchell, 1997), evidence for caprines and decorated pottery can be found, tentatively dated to about 2350-2150BP (Lander & Russell, 2018; Sadr, 2013). Similar findings are also present at the site of Toteng (Kalahari Basin, Botswana), dated to around 2000 BP, which provides directly dated livestock bones and decorated "thin-walled" ceramics (Robbins *et al.*, 2005; Robbins *et al.*, 2008; Sadr, 2013; Smith, 2017).

Apart from direct archaeological evidence, rock art representations of fat-tailed sheep are found across southern Africa (especially Zimbabwe and South Africa), suggesting that local foragers were documenting the arrival of pastoralists to the region (**Figure 4A)** (Cooke, 1965). Another type of rock art, the so-called "geometric art", was found to be distinct from the older fine-line forager rock art and may directly be attributed to the expanding pastoralists (**Figure 4B**) (Güldemann, 2008; Smith *et al.*, 2004).

While the spread of pastoralism is sometimes interpreted as the result of cultural diffusion which introduced livestock to southern African foragers (Sadr, 2013), the more widespread view holds that it is associated with a demic movement that reached the Cape from the eastern part of the continent. Due to the historically attested presence of Stone Age pastoralism among the Khoe-speaking Khoekhoe herders ("Hottentots") of South Africa and Namibia, along with a presumed resemblance of the Khoekhoe language to the Semitic branch of Afro-Asiatic (Bleek, 1851), this migratory movement has been associated with what is now known as the Khoe-Kwadi language family (*e.g.*, Ehret (1982); Westphal (1963)). Most recently, Güldemann (2008) has proposed a testable historical scenario for the migration of Khoe-Kwadi speaking herders into southern Africa. According to his hypothesis, Kx'a and Tuu hunter-gatherers sharing linguistic, cultural and genetic similarities constitute the oldest stratum in the region.

The Khoe-Kwadi, on the other hand, display a linguistic profile different from Kx'a and Tuu, but with a possible areal and genealogical link to languages in eastern Africa. Furthermore, they encompass a cultural diversity unmatched by the Non-Khoe hunter-gatherers, ranging from foragers (Kalahari Khoe), to pastoralists (Khoekhoe), and peripatetic groups (cf. Bollig (2004)) with small stock, which provide various goods and services to their dominant neighbours (Kwadi). The linguistic and cultural diversity of the language family, as well as the substantial sub-branching, implies that the family underwent an expansion, including several divergence and admixture processes (Güldemann, 2008): First, a presumed proto-Khoe-Kwadi population from eastern Africa, with Non-Khoisan genetic profile and a pastoralist subsistence, migrated into north-eastern Botswana, where the pre-Kwadi separated from the pre-Khoe. The pre-Khoe further expanded south into the Kalahari and contacted local foragers speaking non-Khoe languages, some of whom shifted to a Khoe language while maintaining their foraging lifestyle (south-eastern Kalahari Khoe). Other pre-Khoe speakers remained on the northern Kalahari Basin fringe and may have retained their pastoral lifestyle before ultimately reverting to foraging (north-eastern Kalahari Khoe). The expansion of the early Khoe also resulted in an ancestral pre-Khoekhoe group, which reached the southernmost Cape region and admixed with the local foragers while maintaining their original subsistence and language. In a subsequent step, the northern Khoekhoe went north into Namibia and came into contact with groups like the Damara who presumably spoke genealogically related languages and subsequently shifted to Khoekhoe. The modern Nama groups are the direct descendants of these pastoral Khoekhoe, having been subjected to gene flow from either local Khoisan or Bantu groups without a drastic cultural change.

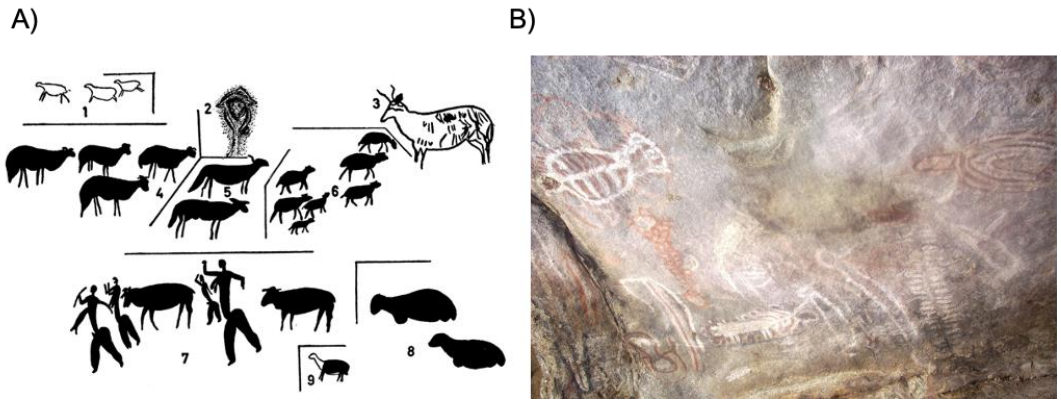

Figure 4 - Rock art created by herders or documenting their arrival. (A) Representations of fat-tailed sheep from several cave paintings across southern Africa. From Cooke, 1965. (B) Geometric forms in Tchitundu Hulu in the Namibe desert of Angola (Courtesy of AM Fehn).

1.4 Goals of the present study

Taking into account the hypothesis outlined above, several inferences can be made about the linguistic and genetic make-up of modern Khoe-Kwadi speakers:

- Khoe-Kwadi languages are expected to display a link to eastern African languages;
- Khoe-Kwadi languages and culture are expected to bear testimony to a pastoral subsistence;
- Khoe-Kwadi speakers are expected to retain part of the genetic profile of the ancestral eastern African population;
- The geographic distribution and inferred branching patterns of the Khoe-Kwadi language family are expected to correlate with archaeological and genetic evidence.

In this project, we aim at testing these assumptions in order to create an integrated model for the diffusion of Khoe-Kwadi languages, pastoralism and genetic material from eastern into southern Africa.

Using linguistic data from fieldwork undertaken by the Human Evolutionary Genetics (HUMANEVOL) group in Angola, Namibia, Botswana and Zimbabwe, in combination with publicly available data, we will reassess the genealogical relations of the Khoe-Kwadi language family. Our approach combines methodologies from historical-comparative linguistics with model-based approaches from bioinformatics in order to arrive at a quantifiable scenario which can be compared with data from human evolutionary genetics and archaeology.

We further use a comprehensive dataset assembling information on genome-wide and uniparental markers from southern Africa to review the distribution of eastern African markers and statistically assess their link to Khoe-Kwadi speakers. The genetic data combines published results with newly collected data from a core-area of the Khoe-Kwadi dispersal: the Angolan Namibe. This area is home to an array of ethnographically and linguistically diverse populations linked to all known population strata of southern Africa (Fehn, 2019b; Oliveira *et al.*, 2018; Oliveira *et al.*, 2019). Among them, the Kwepe are of particular interest to the present study, as two rememberers of Kwadi found among the group in 2014 provide a clear link to previous linguistic studies (see, *e.g.*, Westphal (1971)) and the north-western fringes of the Khoe-Kwadi distribution. In the context of the wider region of southern Africa, the data from the Namibe populations is expected to provide new and valuable insights into the history of the area and help to delimit the routes of the presumed Khoe-Kwadi migration.

By combining empirical evidence from genetics and linguistics, this work follows the Human Evolutionary Genetics Group's aim to develop a new framework for multidisciplinary

population history research (*e.g.*, Oliveira (2019); Semo *et al.* (2019)), and contribute to solving a longstanding question about the population history of sub-Saharan Africa.

Chapter 2: Why are the Khoe-Kwadi good candidates for a pastoral intrusion from eastern Africa?

2.1 Typological split between Khoe-Kwadi and Non-Khoe

Although all “Khoisan“ languages share typological features such as click sounds and a number of structural properties (Güldemann & Fehn, 2017), they cannot be defined as a single genealogical unit with a common ancestor, as proposed by Greenberg (1963). Within southern Africa, there is a basic structural split between the Khoe-Kwadi language family on the one hand, and the Kx’a and Tuu families (Non-Khoe) on the other (Güldemann, 2014). A set of specific features characteristic of Non-Khoe has been identified by Güldemann and Fehn (2017). While some of these features are also shared by a subset of Khoe-Kwadi languages, it is commonly assumed that this is due to contact, rather than genealogical inheritance:

- reduced complexity (lack of bound morphology, a phonological word often the same as lexical root);
- neutral alignment (no case system);
- 1st person inclusive (we, including you) and exclusive (we, excluding you);
- SVO word order;
- lack of ditransitive verbs;
- serial verb constructions;
- head-initial noun phrase with head-final genitive;
- inalienable possessive constructions;
- head of juxtapositional genitive conveys nominal derivation and locative flagging;
- multi-purpose oblique marking;
- a special type of gender system;
- irregular number marking, including nominal and verbal root suppletion (*e.g.*, unrelated singular and plural forms, compare English *person* and *people*).

The structural homogeneity of the Non-Khoe unit is remarkable and certainly portrays a historical relationship between these two families. While this relationship may be explained by inheritance from a common ancestor, it is currently not possible to reconstruct a proto-Kx’a-Tuu language family by using available methodologies from historical-comparative linguistics. Therefore, it is assumed that Kx’a and Tuu became structurally similar through prolonged population and language contact in a contact zone labelled the Kalahari Basin Area ‘Sprachbund’ (Güldemann, 2014; Güldemann & Fehn, 2017). As Khoe-Kwadi is only partially partaking in the intense sharing within this ‘Sprachbund’, it is assumed that they represent a

later arrival to the region (Güldemann, 2014). This hypothesis is further supported by the pronounced differences that exist between and within the two branches of the Khoe group of Khoe-Kwadi, Kalahari Khoe and Khoekhoe, assumingly due to different degrees of contact with Non-Khoe languages after the initial split (Güldemann, 2014).

Although a typological split between Non-Khoe, on the one hand, and Khoe-Kwadi, on the other, does not necessarily imply a link with eastern Africa, two features of Khoe-Kwadi have been identified as indicative of an eastern African origin:

- SVO word-order is shared with many languages of north-eastern and eastern Africa (Heine (1976), his “type D”);
- The sound system of Khoe-Kwadi, including the click inventory, is structurally more similar to that of eastern African click languages - in particular, Sandawe - than it is to Non-Khoe (Elderkin, 2014).

2.2 Lexical indications of a link with eastern Africa

Of the languages presently spoken in eastern Africa, the click language Sandawe has been hypothesized to display a genealogical link with Khoe-Kwadi (Güldemann & Elderkin, 2010). As mentioned above, this assumption is, in part, based on similar sound systems but also the sharing of lexical items. The reconstruction of proto-Khoe-Kwadi words through a careful survey of newly collected and historical data from Kwadi has enabled us to significantly increase the corpus of lexical material possibly shared with Sandawe provided in Güldemann and Elderkin (2010) (**Table 2**).

Apart from the words shared with Sandawe provided in **Table 2** below, the proto-Khoe-Kwadi root *kùdí ‘year’ can also be linked to eastern Africa. While the word is found across the family and has also been borrowed into the Kx’a and Tuu families, it can safely be reconstructed for a great part of the south Cushitic (Afro-Asiatic) language group, thereby adding to the bulk of typological and lexical evidence that hints at an eastern African origin of the Khoe-Kwadi.

Table 2 - Lexical sharing between Khoe-Kwadi languages and Sandawe (Courtesy of AM Fehn; Sources: Kwadi - Westphal (nd-a); Sandawe - ten Raa (2012); Khoe - Vossen (1997), adapted and expanded by AM Fehn).

Abbreviations: p=proto-; KalK=Kalahari Khoe; KK=Khoekhoe).

Gloss (Khoe-Kwadi)	pKhoe-Kwadi	Kwadi	Khoe	Sandawe
'to say, speak, tell'	*buo	(là-là-)bòò-là	*buo-di	bo
'to be there'	*hama	háma-na	*hằắ	hàá-nà-kí ('to sit')
'to stink, smell'	*ha̰m	hḿ	*há̰m̀	himé
'to take'	*sae	see	*sáè	síé
'not'	*tsai (?)	tʃe(la)	te (Ts'ixa, !Ora)	-tsʰi
'to swallow'	*tumu	tumu	*túm̀	tím̀
'night'	*tʰuu	tʰuu	*tʰùú	tʰũũ
'child'	*qx'ae	tʃ'ee	*qx'ae (in *!ũĩ-qx'ae 'other-sex-sibling', Khwe-Ts'ixa+Shua)	k'aré ('young man')
'to cry'	*qx'ae	tʃ'ɛ̀ɛ̀	*qx'àé	k'éé
'mouth'	*qx'ami	k'ámé	*qx'áḿ	k'àmé ('alcoholic drink')
'to give birth, navel'	*ʔaba	ʔawa ('navel')	*ʔábà ('give birth, carry child on back')	haba ('give birth'), habé ('carry child')
'heavy'	*!umu	xúmu-ijo	*!úḿ	!óó-mé ('to fill')
'to sleep' or 'to die'	*ǁʔum 'sleep', ǁʔuo 'die'	ʔmú 'sleep', ʔóò 'die'	*ǁʔúm̀ 'sleep', ǁʔúó 'die'	ǁʔo ('sleep')
'charcoal, ashes'	*ŋǂum	tsũũ ('ashes')	*ŋǂùm̀	kǜű ('red hot charcoals')
'to put'		pee		pèé
'four'		(nɛ́ < Bantu)	*aa-ka	haká
'to draw water'			*a̰rè	hawe ('scoop')
'rainy season'			*bàrà (KalK)	bári ('rainy season')
'to visit'			*dàrá (KalK)	dara ('wait, linger, stay')
'skin'		(kx'óó)	*kʰùó	ʔkʰoo ('house')
'raw'			*k'ù(d)à	k'útsʰè
'bitter'			*k'àú	k'àwàʔé
'that'		(úá)	*ŋǁàá	ná
'this'		(gá)	*nee (?~ŋǀee)	né
'to carry'		(xuu, ʔɲu-ka)	*tánì (KK)	táné ('to pull')
'blue, green'		(só-bɛ̀)	tsãˤã (Naro) (compare also ǂ'Amkoe dzãˤã)	dzaʔǻ
'water'		(k'óó)	*tsʰàá (KalK)	tsʰa ('be cold'), tsʰaá ('tears'), ts'á ('water')
'many'		(kx'áá)	*t(s)ʰíjà (KalK)	tshíya ('all')

Table 2 (cont.) - Lexical sharing between Khoe-Kwadi languages and Sandawe (Courtesy of AM Fehn; Sources: Kwadi - Westphal (nd-a); Sandawe - ten Raa (2012); Khoe - Vossen (1997), adapted and expanded by AM Fehn).

Abbreviations: p=proto-; KalK=Kalahari Khoe; KK=Khoekhoe).

Gloss (Khoe-Kwadi)	pKhoe-Kwadi	Kwadi	Khoe	Sandawe
'faeces, to defecate'		(ʔɲòò)	*tsụ̀ù (KalK)	tsʰo
'to move away, migrate'			*tú̯è	ʔtóri
'to scratch'			*xú(r)é	ʔxadé
'dog'		(ʔáú)	*ʔaba (KalK)	ʔababúa~hababúa ('cheetah')
'buffalo'			*ǀáò	ǀeú
'child'		(kʼoo, tʃʼee)	*ǀűǎ	ǀwã
'leaf'			*gǀàná	ǀàá
'to meet'			*!ʔúá (Nro+KK)	!ʔòó-kí
'to flow, pour'			*ǁuo (KalK)	tɬô-kù ('pour')
'lung'		(púá < Bantu)	*gǁùbà (Tsʼixa+Shua)	ɬuba
'horn'			*ŋǁǎà	tɬana
'to build'		(pʰèdè)	*ŋǁà̰nì (ǁAni+Tsʼixa	tɬine
'to fight'			*ǁʔáà(-kù)	ǁʔáǎ-kí
'to ripen, ripe'			*ǁʔání	ǁʔín-é
'arm'		(lʔìì)	*ǁʔűà	tɬʼũũ
'ear'		(goo)	*ǂáé	kéké
'tree'	*hai	tʃíí	*hàì	tʰèé
'to cook'	*sãĩ	sḗḗ	*sà̰ḭ̀	ʔtʰímé
'blood'		(lʔóó)	taa-ka (Tshwa north)	ʔǃʔaa-ka
'lightning'			*tárì (KalK)	ʔǃnare
'day'		(ʔúdí)	*tʰáè	ʔǃné
'hair'	*lʔũũ	lʔoo~lʔom	*lʔṵ̊ṵ̊	ʔǃʔũ ('body hair')
'mucus, snot, cold'			*lqxʼṵ̋à̰ (KalK; compare also Ju *(g)lqxʼũã)	ʔlʔwaa ('ulcer, sore')
'top'	*ǂʔama	ʔɲama	*ǂʔáṁ	ʔlʔǎǎ-kì ('above')
'to think'			*ǂʔàní	ʔlʔèé ('to see')

2.3 Lexical indications of a pastoral subsistence

In Khoe-Kwadi, words that relate to subsistence practices other than hunting and gathering have been said to be more numerous and semantically diversified than in the Non-Khoe languages, especially in the family's Khoe branch (Güldemann, 2008). These food production-related lexical proto-forms (a putative word reconstructed in a specific proto-language, *e.g.* proto-Khoe) are atypical for an ancient stone-age foraging culture and cannot convincingly be traced to borrowing from modern Bantu languages. Therefore, they may be seen as indicators for a pastoral subsistence imported by proto-Khoe-Kwadi speakers into southern Africa.

It should, however, be noted that the number of subsistence related lexical items for proto-Khoe-Kwadi is relatively small, compared to those found in Khoe and especially Khoekhoe. In addition, there is evidence that the food production-related lexicon found in the family is at least partly derived from generic terms that adopted a more specialised meaning over time. These cases are discussed in the column "comments" in **Table 3** below. In general, the lexical evidence for a pastoral lifestyle of the proto-Khoe-Kwadi is rather slim, except for the item *gùù 'sheep' discussed further below.

Table 3 - Lexical evidence for a non-foraging subsistence (courtesy of AM Fehn, see also Güldemann (2008)).

Form	Meaning	Comment
proto-Khoe-Kwadi		
*gùù	'sheep'	
*gúè	'cattle'	possibly from Bantu *gombe via deletion of mb
*ʔáú(-gu)	'dog'	
proto-Khoe		
*ʔàdì(-gu)	'dog'	
*guu-de	'to herd'	
*lʰùí	'to drive cattle'	
*dV́bì	'to castrate' (> ox)	
*ŋǁùbù~ ŋǁụ̀bù	'to churn, (shake)'	from Non-Khoe *ŋǁụbu 'shake'
*lqx'àó	'to milk in container'	
*tsxòm̀	'to milk into mouth'	from a meaning 'to squeeze out, wring out'
proto-Khoekhoe		
*ǁxàò	'lamb'	from Non-Khoe *ǁxao 'leave child behind'
*gúmà	'cattle'	from pBantu *gombe via mb>m and e>a
*gǁóò	'bull'	prob. from pKhoe *gǁúò 'huge, big'
*ǂxàdà	'bull, stallion'	from a meaning 'scrotum, testicles'
*ŋlűa̋	'calf'	
*!ʔűá	'to herd, rear livestock'	
*lʔùnà	'to separate young from mother'	
*ʔa̋ű	'to curdle (milk)'	
*!hàdè	'to thicken (of curdled milk)' (> sour milk)	

While not entirely conclusive, the reconstructions provided above suggest that prior to the emergence of the Khoekhoe pastoral culture, speakers of proto-Khoe and probably also of proto-Khoe-Kwadi were familiar with domesticated animals and food production (Güldemann, 2008). From all the evidence available, the form *gùù ‘sheep’ has been repeatedly cited as a significant indication for the Khoe-Kwadi as earliest livestock raisers in southern Africa. The root not only occurs in Khoe-Kwadi but also throughout the Kx’a family, in the northernmost Taa varieties (Tuu), as well as in the great majority of south-western and south-eastern Bantu languages spoken in the wider region of southern Africa.

From a comparative analysis of terms for ‘sheep’ in Bantu languages across sub-Saharan Africa (AM Fehn, pers.comm.), it becomes evident that *gùù is indeed restricted to southern Africa and therefore does not have a Bantu origin (**Figure 5A**). It seems likely that the word was borrowed by Bantu speakers entering the region when they encountered the resident sheepherders. From a contemporary perspective, it is interesting to note that the distribution of *gùù in Bantu does exceed the modern Khoe-Kwadi and even “Khoisan”-speaking area by several hundred kilometres (**Figure 5B**).

In summary, it can be said with some certainty that languages of the Khoe-Kwadi family reflect affinities to both a pastoral subsistence pattern and an eastern African origin. Even after careful scrutiny and a considerable expansion of the languages and areas surveyed, the data clearly supports an origin outside the Kalahari Basin, as suggested previously by Westphal (1963), Ehret (1967), Güldemann (2008) and Güldemann (2020).

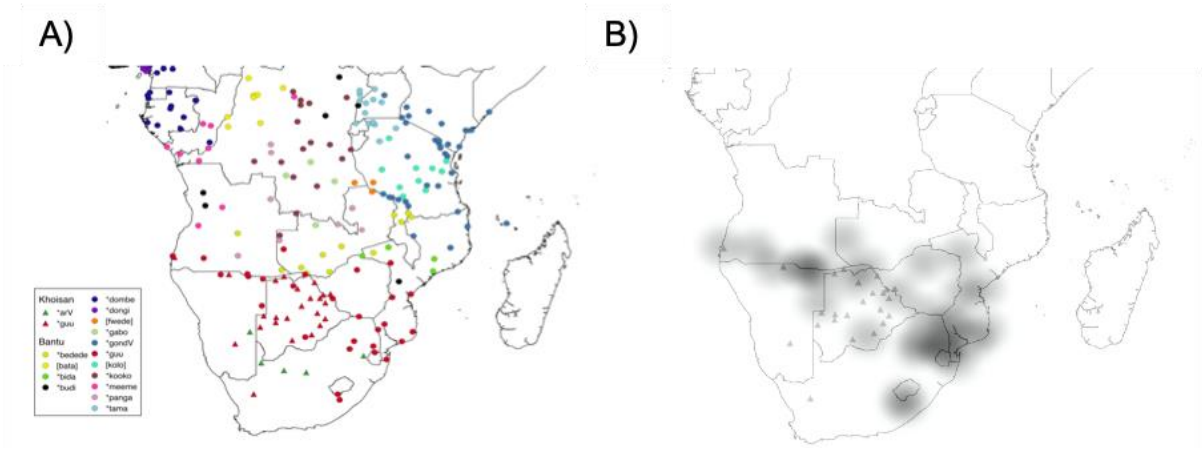

Figure 5 - Distribution of Bantu and “Khoisan” lexical roots for ‘sheep’ (A) and specifically the root *gùù (B) in the wider region of southern Africa. Red circles indicate Bantu languages with *gùù (A) and the grey shades indicate the areal distribution of *gùù in Bantu languages (B). The Khoisan distribution of *gùù is marked by triangles (B). (Courtesy of AM Fehn)

2.4 Ethnographic evidence

Apart from the linguistic evidence, ethnographical data is often cited to support an eastern African origin of the Khoe-Kwadi. The discussion thereby focuses on the Khoekhoe herders, who are the only documented Khoe-Kwadi speaking pastoralists about whom a sufficient amount of information is available.[1] As modern Khoekhoe mostly live in urban contexts and no longer practise their original subsistence, evidence for Khoe-Kwadi pastoral traditions has to be obtained from historical sources. The best-described group is probably constituted by the Cape Khoekhoe, who was first observed by Portuguese sailors at the end of the 15th century (Raven-Hart, 1967).

The Cape Khoekhoe were not only sheep herders but also herded cattle, which conditioned many aspects of their culture. They were a pastoral-foraging society, where men were responsible for the herding while the women were foraging (Fauvelle-Aymar, 2008). They occasionally hunted, but only to obtain materials or as a way to get meat by not slaughtering their domestic stock (Fauvelle-Aymar, 2008). The most particular husbandry practices that are defining for this group are the raising and use of war bulls, and the insufflation of air into the vagina of cows to induce milk production (**Figure 6**) (Fauvelle-Aymar, 2008; Le Quellec, 2011). The latter practise is especially interesting since it is not only documented in the Khoekhoe, including the Nama from Namibia, but also in several populations from eastern Africa, like the Wagogo and Warimi in Tanzania, the Maasai in Kenya and the Afar, Somali and Oromo in the Horn of Africa (Fauvelle-Aymar, 2008). Furthermore, both customs are only present in eastern and southern Africa and the Sahel region, where intensive cattle farming occurs, and are practised only by populations who are specialised cattle herders (Fauvelle-Aymar, 2008).

The livestock associated with the Khoekhoe herders is not indicative of a Bantu origin either: first, the typical cattle of the expanding Bantu pastoralists were of the Sanga breed (a cross between zebu and hump-less taurines found across central Africa), which differ from the longhorn taurines herded by the Khoekhoe. The latter is still found among pastoral groups from southern Angola; second, the breeds of sheep found among the Khoekhoe are common in north-east Africa and Arabia, but entirely absent in the Bantu homeland in west-central Africa (Blench, 1993, 2007; Boonzaier, 1997; Epstein, 1971; MAUENSTEIN, 1980). Furthermore, a defining characteristic of Bantu-speaking pastoral cultures in south-western Africa is the production of butter using a leather bag suspended from two poles and swung from side to side by a seated producer. Evidence of this practise is not present in historical

[1] The formerly Kwadi-speaking Kwepe of south-western Angola are herders and may have a history of sheep and cattle pastoralism. However, their present-day cultural practices resemble those of their Bantu-speaking Himba and Kuvale neighbours and are therefore characteristic of the south-western Bantu pastoral complex, rather than of an ancient Khoe-Kwadi herding tradition.

accounts of the Khoekhoe, nor practised in regions encompassed by the possible distribution of Khoe speakers in the region (Blench, 2007).

Figure 6 - Engraving from the 17th century depicting pastoral practices, *e.g.* cow insufflation in the Khoekhoe. From Le Quellec (2011)

Although a major branch of the Khoe family, the Kalahari Khoe populations, have not practised pastoralism in historical times, it has been proposed that the shared ancestor of Khoekhoe and Kalahari Khoe, the proto-Khoe, may have practised a pastoral-foraging culture. These proto-Khoe would have raised sufficient numbers of sheep to form viable herds, but still practised the kind of foraging lifestyle found among Kalahari Khoe speakers (Ehret, 2008).

Apart from pastoral techniques, the Khoekhoe have cultural practices which possibly link them to an eastern African origin. For example, their death rituals bear similarities to those of the Meru and Maasai from central Kenya: Rituals like abandoning the old and dead people in the bush or carrying dying men through the back door of the house, are all practices that have been linked to an eastern African origin (Fauvelle-Aymar, 2008).

Architectural features have been cited as well: A distinctive marker of Khoekhoe culture is the ‘mat house’, a hut made from a semi-circular frame covered in layers of mats (**Figure 7**). These houses are characteristic of the Khoekhoe but have not been found among the pastoralist Bantu from the region (Blench, 2007). At the same time, the ‘mat house’ is typical of pastoral peoples from Upper Egypt to north-east Kenya, in particular, Cushitic-speakers (*e.g.* Beja), and no similar houses have been found among other pastoral nomads in Africa (Blench, 2007; Prussin, 1995). Other items of material culture are also connected to east Africa: the skin sandals manufactured by the Khoekhoe are identical to the ones still used by eastern African herders today (Blench, 2007; Boonzaier, 1997). While not attested among the Khoekhoe as such, children’s dolls with beads and leather skirts (as still seen among the Zulu from South Africa) are stylistically identical to those made by Nilotic herders in northern Kenya and may be indicative of areal influence from eastern Africa in the south (Blench, 2007; Frobenius, 1933).

The ethnographic features discussed in this section are distributed across southern Africa, from south-western Angola to the Cape, and therefore cover exactly the region occupied by Khoe-Kwadi speakers. Along with the linguistic data, they provide strong evidence that eastern African cultural practices, including pastoralism, were introduced into southern Africa before the region became dominated by Bantu-speaking agro-pastoralists (Blench, 2007; Fauvelle-Aymar, 2008).

Figure 7 - African mat houses. (A) Mat house belonging to the Bishari tribe (Beja people), Sudan (Blench, 2007); (B) Reconstruction of a typical Khoekhoe mat house (https://www.genadendal.info/the-khoikhoi/); (C) Contemporary mat house near Namaqualand National Park, South Africa (http://www.agaves.nl/fieldtrips/ZuidAfrica_2006/NL_Dag_6.htm).

2.5 Genetic evidence: east African genetic markers

Adding to the ethnographic and linguistic evidence mentioned previously, there have been continuous advances in human evolutionary genetics that support a link between southern African populations to eastern Africa (Bajić *et al.*, 2018; Barbieri, Güldemann*, et al.*, 2014; Breton *et al.*, 2014; Lin *et al.*, 2018; Macholdt *et al.*, 2014; Oliveira *et al.*, 2019). This support not only comes from the distribution and diversity patterns of uniparental markers such as mitochondrial DNA (mtDNA) and Y-chromosome (NRY) haplotypes but also from specific regions of the autosome linked to physiologically relevant phenotypes, such as lactase persistence and skin colour.

2.5.1 mtDNA

The majority of mtDNA haplogroups found in modern Khoe-Kwadi speakers have a Khoisan origin; in particular, haplogroups L0d and L0k may even be present at 100% within a given population (*e.g.* Naro).

However, most Khoe-Kwadi populations display a diverse haplogroup composition, including a small number of haplogroups whose origin can be traced to east Africa.

Among those, L5 is the most widespread haplogroup shared between the Kalahari Basin groups (Barbieri, Güldemann*, et al.*, 2014; Barbieri *et al.*, 2013; de Filippo *et al.*, 2010) and populations from the eastern part of the continent, stretching from south-east Africa to Kenya, Ethiopia and north towards the Arab Peninsula (Abu-Amero *et al.*, 2008; Kivisild *et al.*, 2004). Before data from southern Africa became available, L5 was found at its highest frequency in the Mbuti Pygmies in eastern central Africa (~15%). In our dataset, the Khoe-Kwadi-speaking Tshwa show the highest percentage of L5 (~18%), followed by the east African Sandawe who have a comparably low percentage (~5%) (**Figure 8**) (Barbieri, Güldemann*, et al.*, 2014; Tishkoff *et al.*, 2007). Other Khoe-Kwadi populations with this haplogroup are the Shua (~5%) and the Nama (~4%) (Barbieri, Güldemann*, et al.*, 2014; Schlebusch, 2010). Within southern Africa, this haplogroup only occurs in the Khoe-Kwadi and eastern Bantu populations but is not found in either the Kx'a or Tuu. In total, the Khoe-Kwadi have in average five times more L5 in their gene pool (~18%) than Non-Khoe Kwadi populations (3%), albeit this is being triggered by the high frequency found in the Tshwa sample since there is no statistically significant difference (Mann-Whitney U test; p=0.6071).

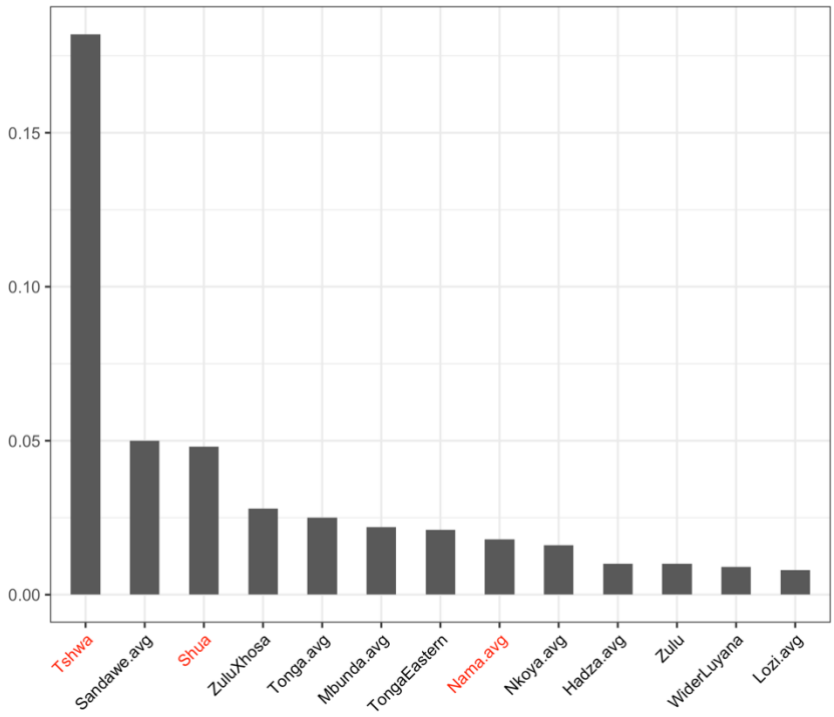

Figure 8 - Distribution of haplogroup L5 in our studied populations, ordered from highest to lowest frequencies. Populations coloured in red belong to the Khoe-Kwadi group.

Another haplogroup with eastern African links, L4, is less widespread in the data (**Figure 9A**) (Barbieri, Güldemann*, et al.*, 2014; Barbieri *et al.*, 2013; de Filippo *et al.*, 2010; Fernandes *et al.*, 2015; Marks *et al.*, 2015; Oliveira, 2019; Salas *et al.*, 2002; Schlebusch, 2010; Soares *et al.*, 2016; Tishkoff *et al.*, 2007). Although our data shows Kx'a-speaking populations to have the highest frequencies of this haplogroup, it is also attested among Khoe-Kwadi speakers like the Hai||om (Khoekhoe) and Naro (Kalahari Khoe). In our dataset, the average frequency of L4 in the Khoe-Kwadi (~0.2%) does not significantly differ from the frequency found in Kx'a and Tuu-speaking populations (~0.3%) (Mann-Whitney U test; p=0.782) (**Figure 9B**) (Barbieri, Güldemann*, et al.*, 2014; Barbieri *et al.*, 2013). The low frequencies found in our sampled populations are not necessarily surprising, considering L4 only reaches ~17% in the original eastern African mtDNA pool (Barbieri, Güldemann*, et al.*, 2014; Barbieri *et al.*, 2013; Fernandes *et al.*, 2015).

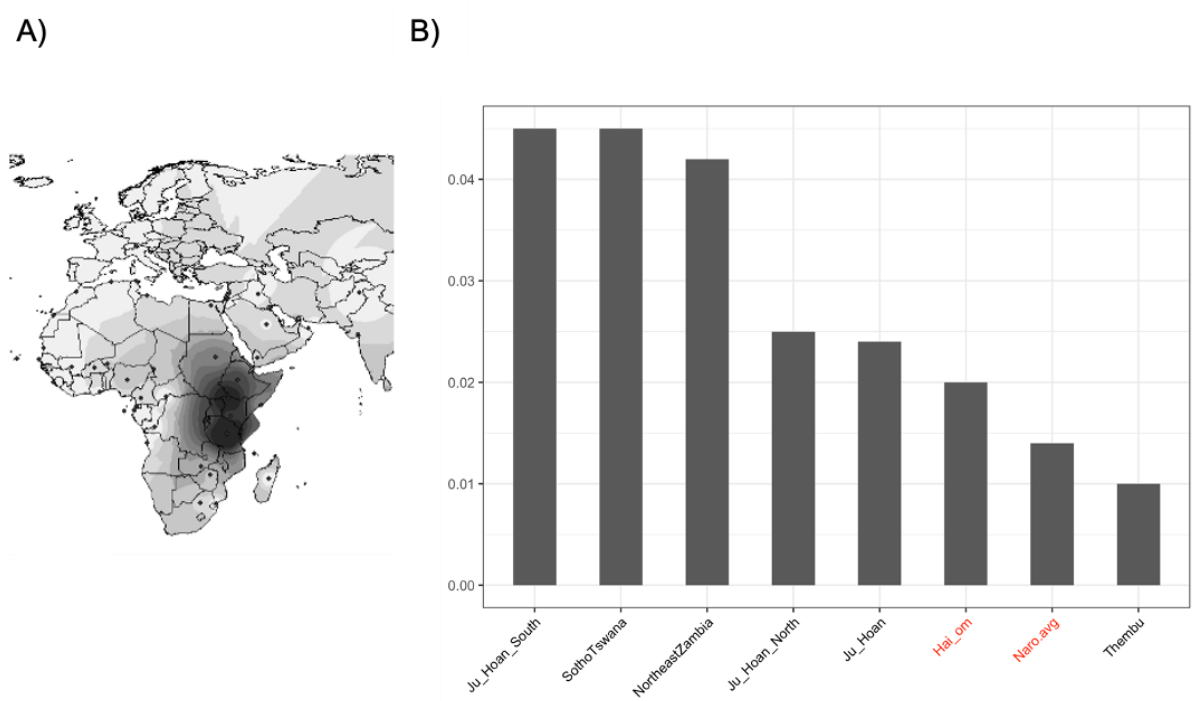

Figure 9 - Distribution of haplogroup L4. (A) Worldwide distribution of L4 from Fernandes *et al.* (2015). (B) Distribution of L4 frequencies in southern African populations, ordered from highest to lowest. Populations coloured in red belong to the Khoe-Kwadi group.

2.5.2 NRY

Another major support for an eastern African influence in southern Africa comes from the Y-chromosome haplogroup E1b1b (also called E-M293), which is a variant of the broadly distributed E3b1-M35* clade common in east and central Africa (Henn *et al.*, 2008). This variant has its geographic origin in present-day Tanzania, with the Datog (Nilo-Saharan) displaying the highest frequency in the region (~43%) (**Figure 10A**) (Henn *et al.*, 2008). Its distribution across eastern African populations, including the click-speaking Sandawe (~23%), strongly suggests that E1b1b spread through Tanzania to southern-central Africa around 2000 years ago, therefore providing evidence for human migrations between the two regions that possibly accompanied the introduction of pastoralism to southern Africa (Henn 2008).

In our sample, the Kalahari Khoe-speaking Khwe have the highest frequency of this haplogroup (~46%). Many Khoe-Kwadi populations display frequencies of E1b1b which are significantly higher than those found in other groups (Mann-Whitney U test; p=0.037): average frequencies number to around ~14.7%, while the average found among Non-Khoe-Kwadi populations is ~5.5%. Nevertheless, the haplogroup is also widely attested among Kx'a, Tuu and eastern Bantu speakers, indicating a broad areal influence (**Figure 10B**) (Bajić *et al.*, 2018; Schlebusch, 2010).

A preliminary study of the NRY composition of the Tjwao, a Tshwa-speaking population from north-western Zimbabwe undertaken by the Human Evolutionary Genetics group yielded ~24% of E1b1b, as well as the presence of haplogroup J1 (Mineiro, 2016). J1 is thought to have originated in the Middle east before entering eastern and northern Africa (Semino *et al.*, 2004; Tofanelli *et al.*, 2009). Due to its geographic trajectory and the discovery of high frequencies in the Arab Peninsula and Cushitic-speakers from Ethiopia, it has been proposed that J1 accompanied the expansion of pastoralists into arid habitats (Chiaroni *et al.*, 2010). In Zimbabwe, J1 was only found among Tjwao speakers while being absent in Bantu speakers sampled in the same region. This finding may support the presence of this haplogroup in southern Africa before the Bantu expansions reached this part of the continent.

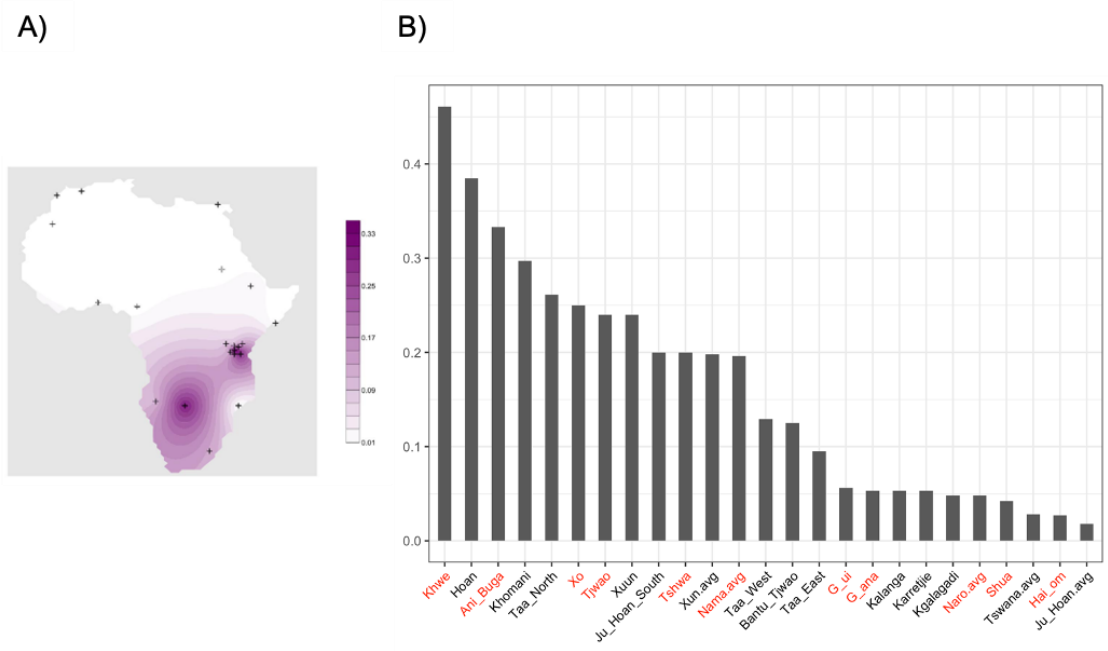

Figure 10 - Distribution of haplogroup E1b1b. (A) Distribution of frequencies across the African continent (Henn *et al.*, 2008). (B) Distribution of frequencies in our study populations, ordered from highest to lowest. Populations coloured in red belong to the Khoe-Kwadi group.

2.5.3 Lactase Persistence (-14010ºC variant)

Lactase persistence (LP) variants that confer the ability to digest milk into adulthood, are likely targets of selection when populations convert from a hunter-gatherer to a pastoralist/farming lifestyle (Rocha, 2012). Lactase persistence arose multiple times through convergent evolution, and different variants can be associated with particular geographic regions and populations, like the SNP 14010G>C, present in the LP-regulatory region (**Figure 11A**) (Tishkoff *et al.*, 2007). Since this allele predominates in pastoralists from Kenya (~28%) and Tanzania (~32%) (Tishkoff *et al.*, 2007) while being rare or absent in other regions of the continent, we can assume that its presence in southern Africa was brought by the migration of pastoralists from eastern Africa who interacted with the ancestors of the Khoe-Kwadi (Coelho *et al.*, 2009).

Macholdt et al. (2014) found that this allele occurs at a significantly higher frequency in pastoralists (20.2%) than in foragers (6.7%) and agropastoral Bantu speakers (3.9%); they further discovered significantly higher frequencies in Khoe-Kwadi speakers than in other populations from southern Africa (Tuu, Kx'a and Bantu) (Figure 11B) (Macholdt et al., 2014). In our dataset, the average frequency of -14010C in the Khoe-Kwadi reaches ~9.8%, almost twice as high than the Non-Khoe-Kwadi average (~5.1%) (Mann-Whitney U test; p=0.016). Within Khoe-Kwadi, this allele appears in populations speaking languages of all major branches (Kwadi, Kalahari Khoe, Khoekhoe): it reaches exceptionally high frequencies in the pastoralist Nama (~28%), followed by the G||ana (~20%) and the Tshwa (~17%) (**Figure 12A**) (Breton *et al.*, 2014; Macholdt *et al.*, 2014; Schlebusch, 2010). These groups, along with Herero/Himba speaking pastoralists from Namibia, show positive selection for the allele (Macholdt *et al.*, 2014). As the G||ana and Tshwa are commonly classified as hunter-gatherer populations, it has been suggested that their high frequencies and selection signals are indicative of a recent shift from pastoralism to foraging, possibly triggered by environmental reasons or displacement associated with the arrival of food-producing Bantu speakers in the area (Güldemann, 2008; Macholdt *et al.*, 2014).

In the Namibe region, the formerly Kwadi-speaking Kwepe only display a comparatively low frequency of the variant (~4%) (**Figure 12B**) (Pinto *et al.*, 2016). However, other populations from the region – in particular, the peripatetic Tjimba, Kwisi and Twa – all display frequencies >15% and thereby above the Khoe-Kwadi average. When all populations from the area (**Figure 25**), including the pastoral Himba and Kuvale, are taken into account, the average frequency (~13,2%) matches the Khoe-Kwadi profile (Mann-Whitney U test; p=0.430) and displays a statistical difference when compared to the remaining Non-Khoe-Kwadi populations (Mann-Whitney U test; p=0.017). In general, our population sample suggests an areal influence likely linked to the presence of Khoe-Kwadi speakers in south-western Angola,

whereby the eastern African herders may have preferably interacted with the peripatetic groups.

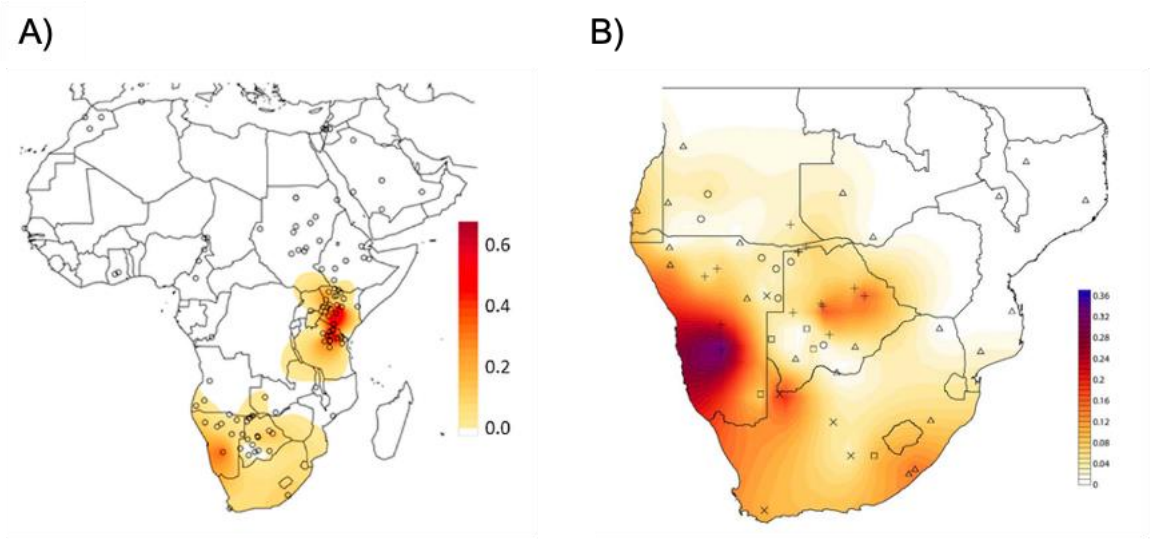

Figure 11 - Distribution of the -14010C allele in African (A) and southern African (B) populations. (Macholdt *et al.*, 2014; Pinto *et al.*, 2016).

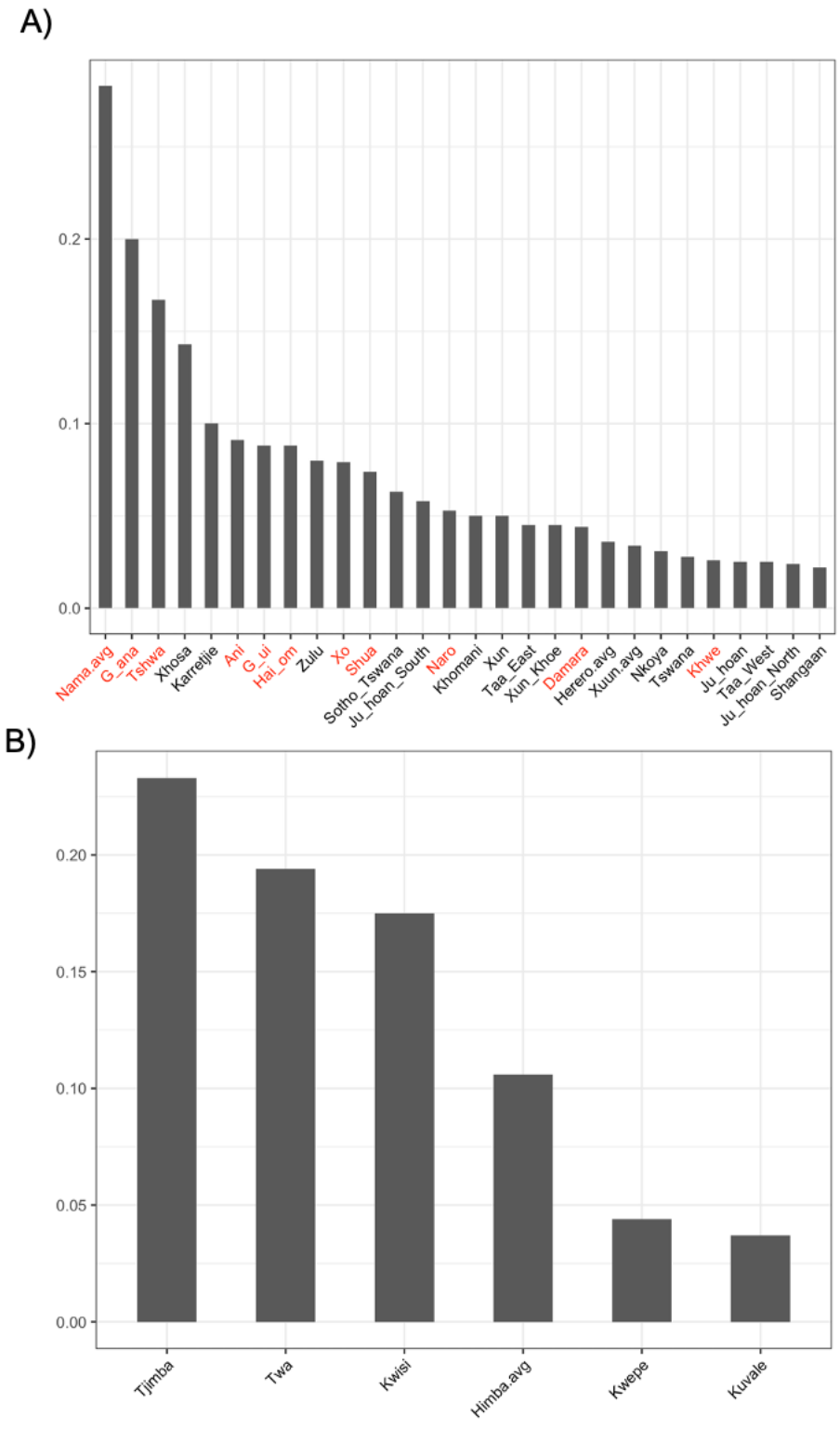

Figure 12 - Frequencies of the -14010C allele in southern African populations (A) and Angolan Namibe populations (B), ordered from highest to lowest. Populations coloured in red belong to the Khoe-Kwadi group.

2.5.4 "Light" skin colour gene (SLC24A5)

A feature of many autochthonous hunter-gatherers of southern Africa is their characteristic "light" skin that reflects a possible local adaptation to a region of Africa with reduced UV radiation (Lin *et al.*, 2018).

A rs1426654*A allele SLC24A5 with likely Eurasian origin (**Figure 13A**) was found with high frequencies in some Khoisan populations. This includes Khoe-Kwadi speakers, although only a subset is generally perceived as "light-skinned" (see **Figure 2**) (Lin *et al.*, 2018).

While northern African populations have frequencies of almost 100% of SLC24A5, southern African groups like the Nama have about ~48% of this variant (**Figure 13B**) (Lin *et al.*, 2018). The Khwe have lower frequencies (~12%), in line with their generally darker skin colour (see **Figure 2**) (Lin *et al.*, 2018). Overall, the average frequency attested in Khoe-Kwadi populations (~30%) is very similar to that of the studied Non-Khoe-Kwadi groups from southern Africa (30.6%) (Mann-Whitney U test; p=0.531).

Although SLC24A5 is associated with Eurasians, there is a strong resemblance between the Khoisan and the European haplotypes of this gene, suggesting a common origin and subsequent distribution through a migration event (Lin *et al.*, 2018). While it cannot be excluded that the variant entered autochthonous southern African populations quite recently, in colonial times, it seems more likely that it was brought to the area by the same population movement that introduced other markers from eastern Africa and the Near east. In this scenario, the migrating population from eastern Africa would have been in contact with light-skinned west Eurasians before entering southern Africa (Lin *et al.*, 2018).

An eastern African source for the southern African variant of SLC24A5 is not only supported by significant frequencies in the Ethiopian Somali (~56%) and the pastoralist Maasai (40%), but also by an Approximate Bayesian Computation (ABC) approach that favours an east African source model: this model assumes that the allele derives from eastern African pastoralists within 1000 to 2000 years ago, with a second pulse of recent European admixture during the last 300 years (Lin *et al.*, 2018).

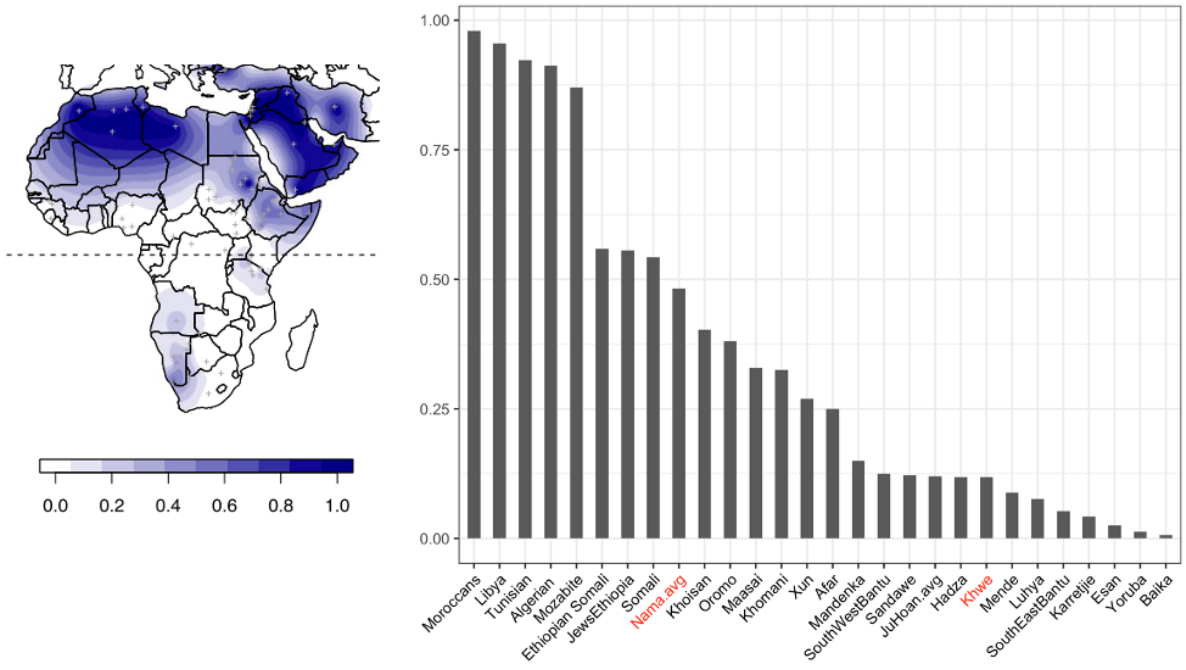

Figure 13 - Distribution of the rs1426654*A allele variant in Africa (A) and in southern African populations (B) (Lin *et al.*, 2018). Histogram is ordered from highest to lowest. Populations coloured in red belong to the Khoe-Kwadi group.

Chapter 3: Diversity of the Khoe-Kwadi

In the previous chapter, established evidence for a link between eastern and southern Africa with a likely connection to people speaking languages of the Khoe-Kwadi family has been provided. In the following sections, I will present novel analyses of newly collected and publicly available data from linguistics (§3.1) and human genetics (§3.2), which contribute to the creation of an integrated model for the diffusion of Khoe-Kwadi languages, pastoralism and genetic material from eastern Africa into southern Africa. The focus will be on the internal diversity of the Khoe-Kwadi, and their relations with neighbouring people speaking languages of the Kx'a, Tuu and Bantu families.

3.1 Linguistic diversity

In this section, we analyse the internal diversity and sub-structuring of the Khoe-Kwadi language family using methodologies from historical-comparative linguistics and bioinformatics. Our analysis is based on a comprehensive dataset of lexical data from 35 Khoe-Kwadi languages, including data from the now-extinct Kwadi language of south-western Angola, and historical data from two Khoekhoe varieties formerly spoken in the Cape region of South Africa (**Figure 14**). Our dataset covers all main lineages of the family (Kwadi, Kalahari Khoe, Khoekhoe) (see **Table 1**) and provides a maximally complete areal coverage of these lineages (**Figure S 1**).

The lexical data used in the study was in part collected by members of the HUMANEVOL group from Khoe-Kwadi-speaking communities in Angola (Kwadi), Botswana (Ts'ixa, Khwe, Shua), Namibia (Khwe) and Zimbabwe (Tjwao); supplementary data were retrieved from publicly available sources (**Table S 1**) and re-transcribed into a working orthography to facilitate comparison.

While linguistic classification commonly relies on multiple types of data (lexical, phonological, grammatical), we opted for lexical data because of its ready availability for all Khoe-Kwadi varieties, and also for its suitability to be used in computational analyses. We sorted the words available from our wordlists into sets of related roots (cognate sets) and created a binary-coded dataset (<1> for the presence of a given root, <0> for absence, <?> for missing data).

To learn more about the structure of Khoe-Kwadi, we used the binary-coded dataset to compute a dissimilarity matrix and applied the NeighborNet algorithm (Bryant & Moulton, 2004) to detect conflicting signals indicative of contact and reticulation between individual varieties. Our NeighborNet analysis (**Figure 15**) identifies the three main lineages - Kwadi, Kalahari Khoe and Khoekhoe - and shows that they have a star-like structure, where all three main lineages appear to be equidistant. Within the Kalahari Khoe subgroup (purple cluster),

we can observe a multitude of “box-like” structures in the centre: these indicate reticulation, meaning ongoing contact and admixture between individual languages and language clusters.

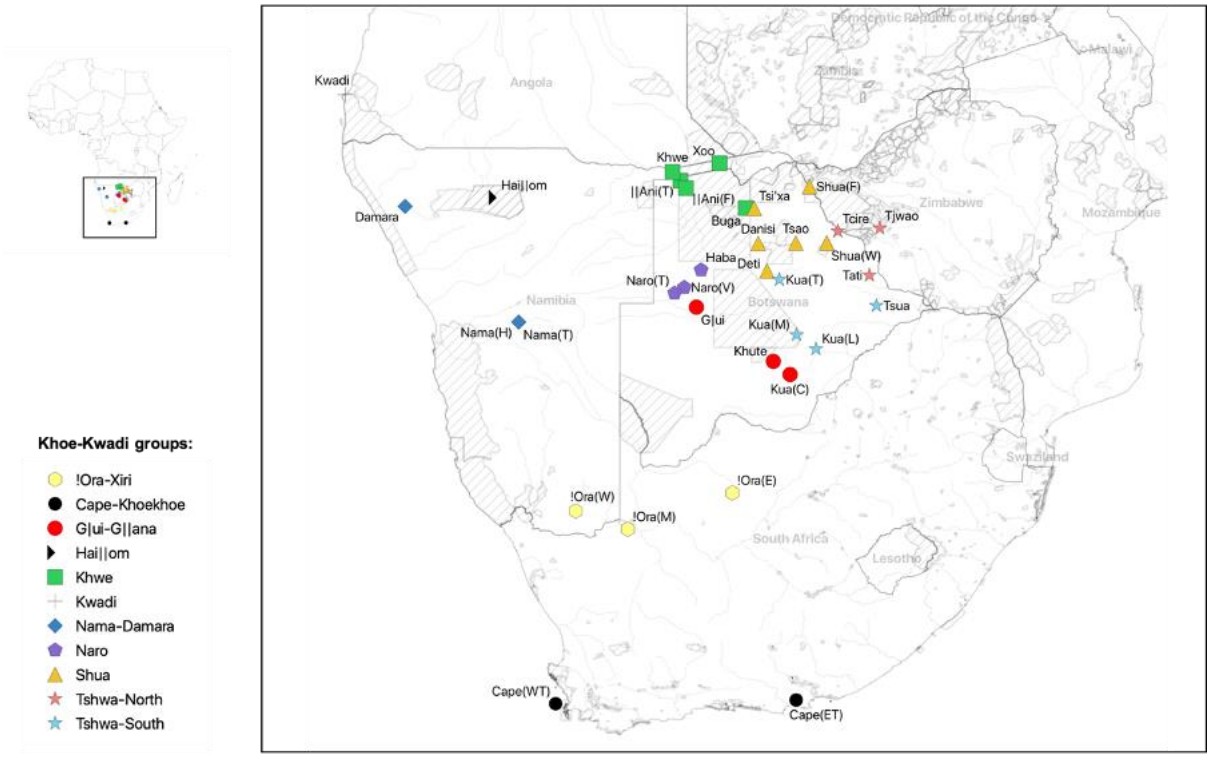

Figure 14 - Geographic distribution of the 35 Khoe-Kwadi doculects. Black lines indicate rivers and dashed polygons refer to established conservation areas.

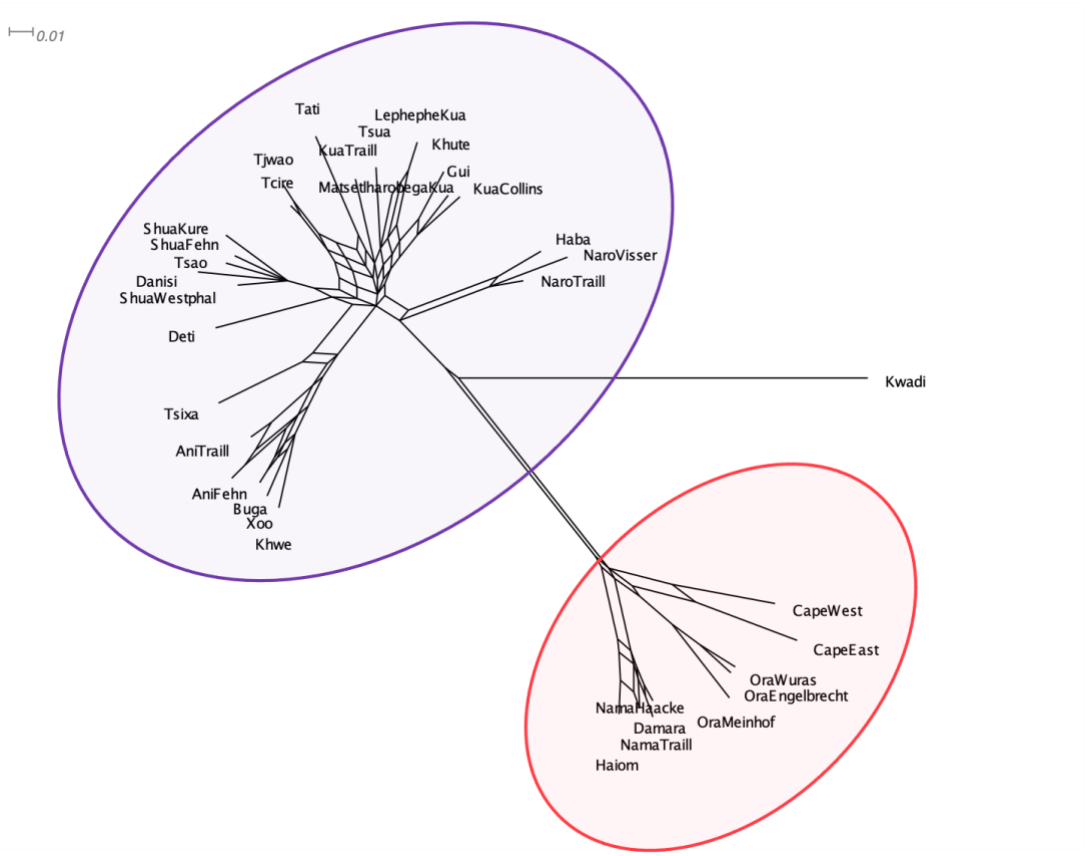

Figure 15 - Split-graph of the NeighborNet analysis of 35 varieties of Khoe-Kwadi. Kalahari-Khoe and KhoeKhoe varieties are circled in purple and red, respectively. Scale bar indicates distance. Mean weights <0.005 were filtered out.

To quantify the tree-likeness of our dataset, we further computed the Delta and Q-residual scores (**Table 4**). The average scores for the Khoe-Kwadi family as a whole suggest that the data is moderately tree-like (Delta: 0.26; Q-residual: 0.01) (**Table 4**). The lineages Kwadi (Delta: 0.22; Q-residual: 0.01) (**Table S 2**) and Khoekhoe (Delta: 0.24; Q-residual: 0.03) are most fitting to a tree-like pattern, while the Kalahari-Khoe display higher scores (Delta: 0.29; Q-residual: 0.02) (**Table 4**). For comparison, we looked at the Indo-European tree from Gray et al. (2010), whose tree-like net was accompanied by scores slightly lower than what can be observed for Khoe-Kwadi (Delta: 0.22; Q-residual: 0.02) (Gray *et al.*, 2010). The threshold for non-tree like networks can be compared to a tree of Polynesian languages, displaying scores which are considerably higher than what can be observed for Khoe-Kwadi or any of its sub-lineages (Delta: 0.41; Q-residual: 0.02) (Gray *et al.*, 2010). Although the Kalahari-Khoe scores are not as high as those of the Polynesian tree, they still confirm the conflicting signal in its internal structure which was also detected in the SplitsGraph (**Figure 15**).

As languages of the Kalahari Khoe subgroup are spoken in a rather small area mostly encompassing the Kalahari Basin and its northern and eastern fringes (**Figure 14**), it is conceivable that geography may be the primary driver of linguistic differentiation, with

neighbouring varieties sharing a considerable part of their lexicon. Clusters of more closely linked languages may be interpreted as dialect chains, which is the explanation adopted for the network-type structure of the Polynesian language tree (Gray *et al.*, 2010). In the case of the Kalahari Khoe, it may be hypothesized that areally restricted borrowing from Non-Khoe languages of the Kx'a and Tuu families also influenced the lexical differentiation of this subgroup: while languages spoken in the central Kalahari may borrow from both Kx'a and Tuu (Traill & Nakagawa, 2000), Tuu influence is likely to lessen on the northern fringe which shows a predominance of loanwords from Kx'a (!Xun) (Köhler, 1973/74).

To further quantify the influence of geography on the structure of Khoe-Kwadi and its lineages, we computed Mantel tests for the full dataset and the Kalahari and Khoekhoe subgroups, using Pearson and Spearman correlations (**Table 5**). For all groups, we obtained positive values (r>0, **Table 5**) with statistical significance (p-value<0.05, **Table 5**), as well as positive linear models for distance plots (**Figure S 2**). Therefore, we may assume that linguistic distances within Khoe-Kwadi increase with geographic distance, *i.e.* geography does play a role in the internal organization of the Khoe-Kwadi language family.

Table 4 - Delta and Q-residual scores computed for Khoe-Kwadi and its sub-branches.

Group	Delta score	Q-residual score
Khoe-Kwadi	0.2566	0.0119
Kalahari-Khoe	0.2896	0.0230
Khoekhoe	0.2408	0.0299

Table 5 - Pearson and Spearman correlation tests for Khoe-Kwadi and its sub-branches. "*" indicates p-value<0.05.

Group	Pearson Correlation		Spearman Correlation	
	r	p-value	r	p-value
Khoe-Kwadi	0.777	1.00E-06*	0.744	1.00E-06*
Kalahari-Khoe	0.658	1.00E-06*	0.640	1.00E-06*
Khoekhoe	0.845	5.35E-04*	0.876	4.24E-04*

In a subsequent step, we assessed the phylogenetic relationships between the Khoe-Kwadi languages using a Bayesian phylogenetic approach. Unlike distance-based clustering methods (see **Figure S 3**, **Figure S 4**, **Figure S 5**), this character-based approach allows us to establish priors, define clock rates, specify rates of model evolution, and eventually model the best-fitting evolutionary scenario for our dataset. The model best-fitting our dataset (log ML: -4131.2144, **Table S 3**) is the Covarion model, which is included in the babel package providing models suited for cases of language diversification (Atkinson, 2011; Bouckaert, 2015; Penny *et al.*, 2001). In addition, the relaxed clock model allowing variation of clock rates (Bouckaert, 2015; Bouckaert *et al.*, 2012) was found to be the best fit. If the tree is computed with no priors, Kalahari Khoe and Khoekhoe are both clearly identified, while the position of Kwadi remains unresolved (**Figure S 6**). In the consensus tree, Kwadi is shown as sharing a common ancestor with languages of the Kalahari Khoe subgroup, but the posterior probability is low (65%). A similar problem arises with distance-based clustering methods (**Figure S 3**, **Figure S 4**, **Figure S 5**), none of which is decisive on the position of Kwadi. This is in stark contrast to evidence from linguistic domains other than lexicon, which clearly show a close link between Kalahari Khoe and Khoekhoe, leading to the establishment of a Khoe language family (Vossen, 1997). Güldemann (2004) shows Kwadi to be a higher-order relative of this Khoe family by focusing on relations between the languages' pronoun systems – a methodology necessitated by the stark overall differences between the branches, which make a link between Kwadi and Kalahari Khoe as seen in our tree rather unlikely. We, therefore, opted to consider the position of Kwadi the result of lack of resolution in the lexical data and computed a Bayesian evolutionary tree model with priors, enforcing Kwadi as an outgroup.

In the resulting consensus tree (**Figure 16**), both the Khoekhoe (100%) and Kalahari Khoe (95%) subgroups are identified with high posterior probabilities. The Cape varieties cluster with data from modern Standard Namibian Khoekhoe (Nama-Damara and Haillom), as well as with the now extinct !Ora varieties from the Orange River Valley of South Africa and southern Namibia. Although the tree shows them as outliers within Khoekhoe, this is probably due to a high amount of missing data, as phonological comparisons indicate that the Cape varieties were closer to !Ora (Haacke, 2018). The relatively short branch length estimated for the full cluster indicates a high similarity of Khoekhoe varieties spoken from the Cape to northern Namibia, possibly in support of a rather rapid northward expansion following the arrival of the Nguni-speaking Bantu and European colonists on the southern tip of the continent.

Within Kalahari Khoe, we see a major split between a clearly distinct northern cluster (82%), corresponding to Khwe-Ts'ixa (100%) and Shua (91%), and a less well-structured southern cluster (57%) which consists of Glui-Gǁana+Naro (100%) and two separate clusters

roughly corresponding to what we tentatively term southern (50%) and northern Tshwa (97%). While southern Tshwa firmly clusters with Glui-Gǁana+Naro (91%), the position of northern Tshwa is unresolved. Considering data from linguistic domains other than lexicon, in particular pronoun systems (Fehn & Phiri, 2017; Pratchett, 2020), northern Tshwa indeed shows proximity to southern Tshwa. Therefore, its uncertain position in the tree may be due to a northward migration and subsequent contact influence from Shua, as speakers of northern Tshwa and Shua are known to interact and intermarry along the north-eastern border of Botswana.

Interestingly, our results are at odds with the subclassification proposed by Vossen (1997), who distinguishes western (Glui-Gǁana, Naro, Khwe) and eastern (Shua+Ts'ixa, Tshwa) Kalahari Khoe subgroups. The core lexicon instead suggests a split between northern (Khwe+Ts'ixa, Shua) and southern Kalahari Khoe (Glui-Gǁana, Naro, Tshwa). In contrast, the clear proximity between Shua and Tshwa as detected by Vossen (1997) may well be the result of contact following a northward migration of some Tshwa speakers, rather than genealogical inheritance. Although our analysis is based on lexicon only, it is in line with some more recent proposals based on other types of linguistic data which link Tshwa to Glui-Gǁana (Pratchett, 2020) and Ts'ixa to Khwe (Fehn, 2016).

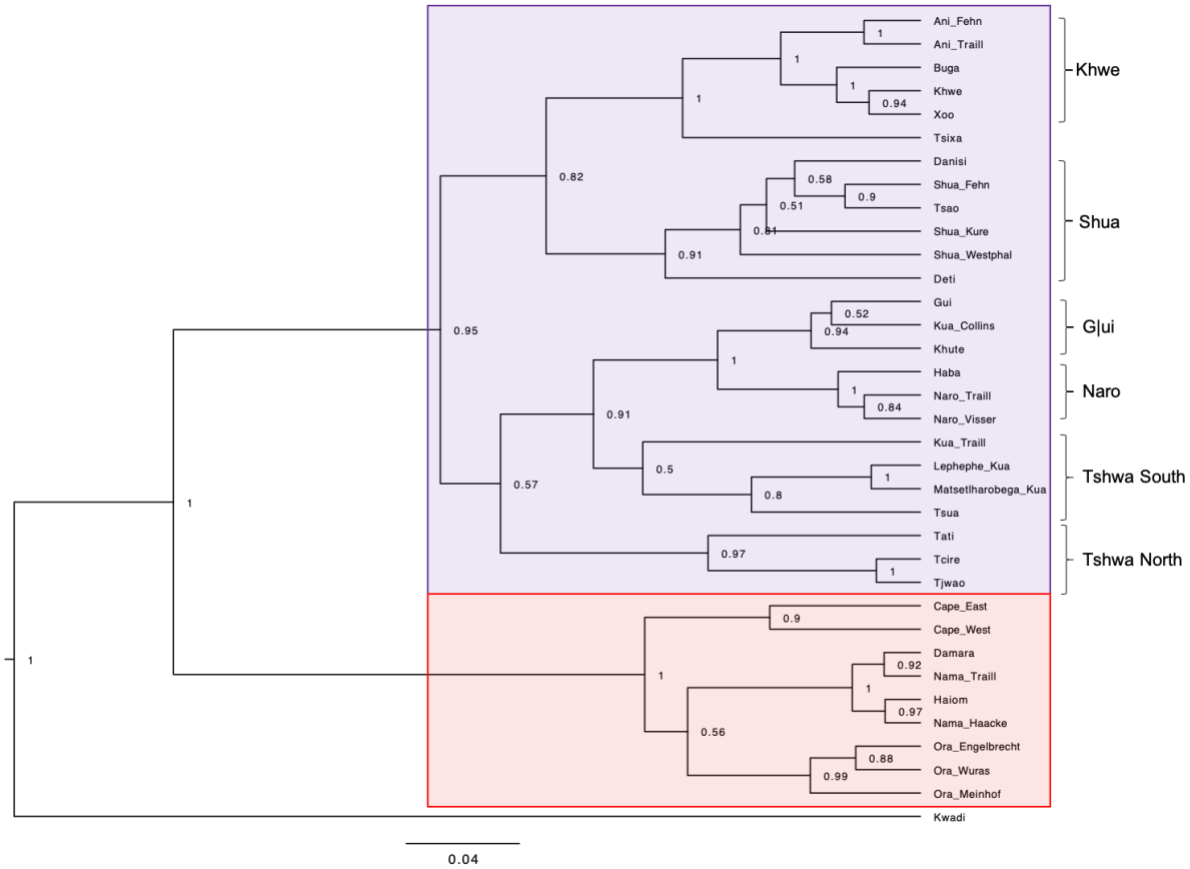

Figure 16 - Bayesian consensus tree with posterior probabilities in each internal node, with Kwadi as an outgroup. Red and purple squares indicate, respectively, the Khoekhoe and Kalahari-Khoe sub-branches.

In order to assess the migration route and geographical diversification pattern of the proto-Khoe-Kwadi, we followed previous works on Bantu (Grollemund *et al.*, 2015), Pama-Nyungan (Bouckaert *et al.*, 2018) and Indo-European (Bouckaert *et al.*, 2012) language families and produced a Bayesian tree that considers geographic location as priors. This method infers the ancestral latitude and longitude for each tree node and allows for the reconstruction of the ancestral location from which the family diverged.

The best tree computed from our analysis (log ML: -4492.8205, **Table S 3**) shows that the highest posterior probabilities (in red) for the location range of the ancestral language of the Khoe-Kwadi falls in the border area of Namibia and Botswana (**Figure 17**). In combination with a presumed eastern African origin of the ancestral population, the data supports an arrival of the proto-Khoe-Kwadi in south-western Africa, from where they split according to an initially tree-like pattern (see also **Figure 15**): The ancestors of the Kwadi went northwards, the ancestors of the Khoekhoe moved south, and the ancestors of the Kalahari Khoe moved east and diversified while entering such diverse ecosystems as the Kalahari dry savannah and the Okavango swamps.

The proposed migration contradicts some of the major hypotheses on how the present-day Khoe-Kwadi speakers reached their current whereabouts, most of which locate the proto-Khoe-Kwadi language in north-eastern Botswana: Cooke (1965) suggested that the first herders spread westward from eastern Botswana and then south through Namibia into the Cape of South Africa (**Figure 18** (left)). More recently, it was proposed that the ancestor of the Khoe-Kwadi moved along the eastern fringes of the Kalahari desert, then south toward the centre of South Africa, towards Namibia, and further south into the Cape (**Figure 18** (right)) (Ehret, 2008; Elphick, 1985).

Interestingly, our result concurs with Heine and Konig (2008), who reconstructed the homeland of proto-Khoekhoe within the methodological framework of “linguistic geography”. Their results suggest that the split of the proto-Khoekhoe from their closest relatives, the proto-Kalahari Khoe, would have occurred in the area most immediate to the present-day location of both sub-branches: the western fringes of the Kalahari along the border between Botswana and Namibia.

A western route as backed by our analysis is also in line with the archaeological record, which places the first evidence for pastoralism in southern Africa (~2000BP) in northern Botswana and along the western coast of Namibia and South Africa (**Figure 19**).

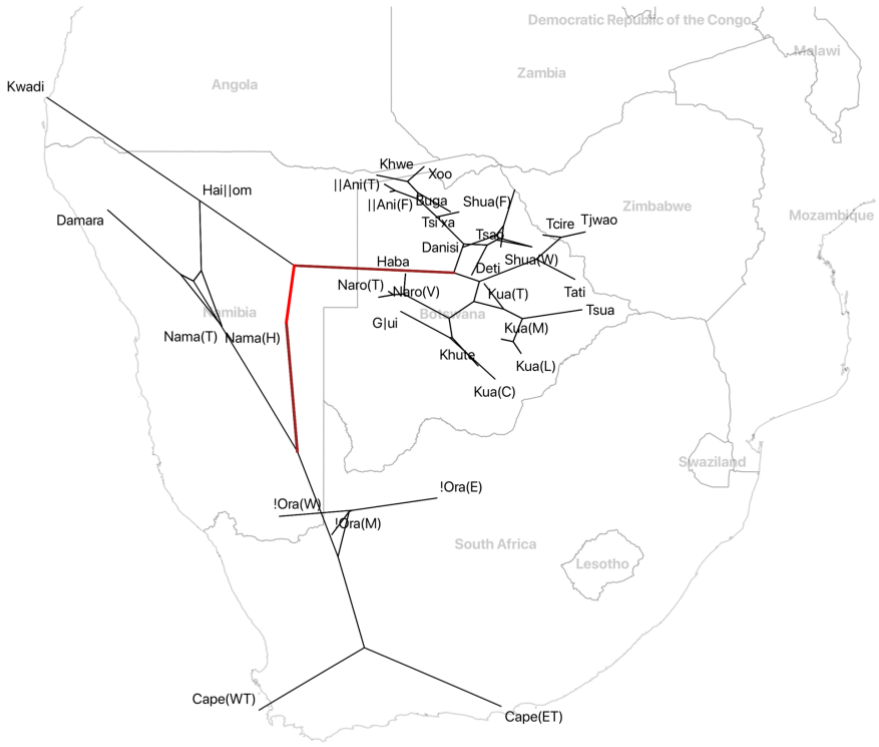

Figure 17 - Inferred geographic origin of the Khoe-Kwadi family. Bright red branches indicate the 95% highest posterior density regions of the location of the internal nodes.

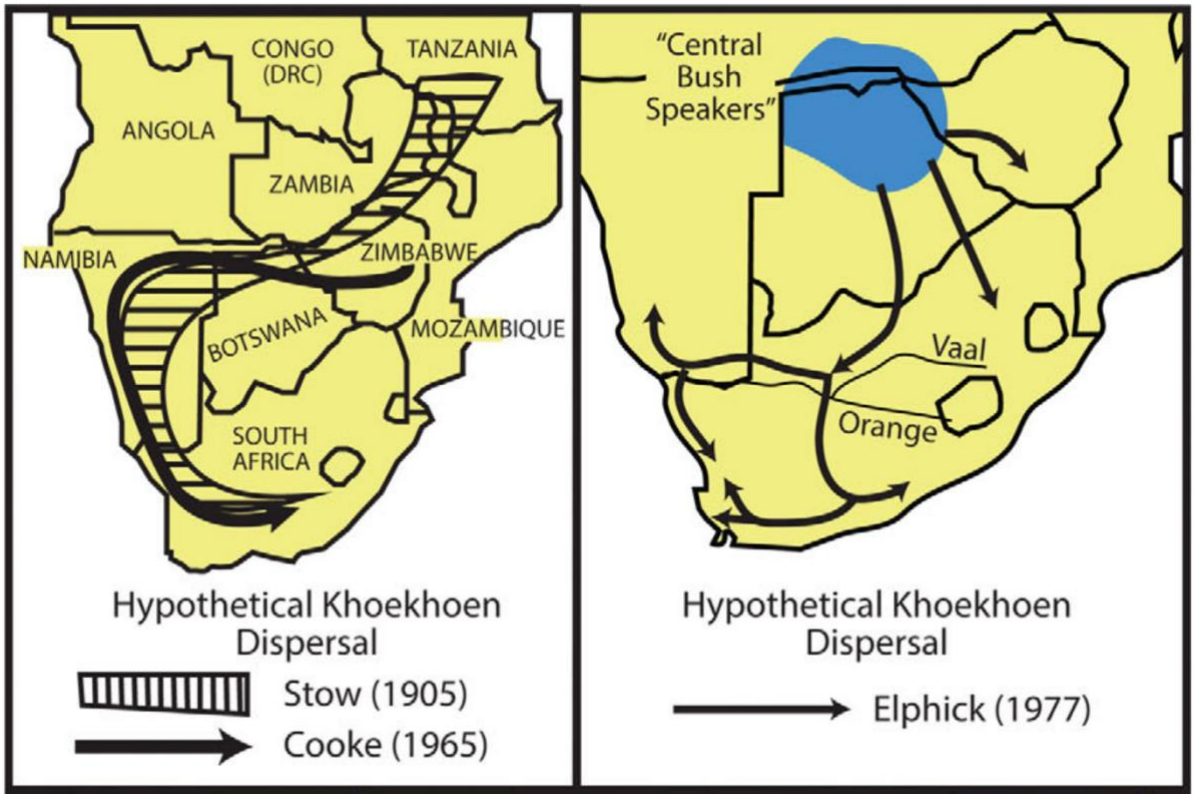

Figure 18 – Hypothetical routes of the Khoe-Kwadi ("Khoekhoen") dispersal in southern Africa. From Orton *et al.* (2013).

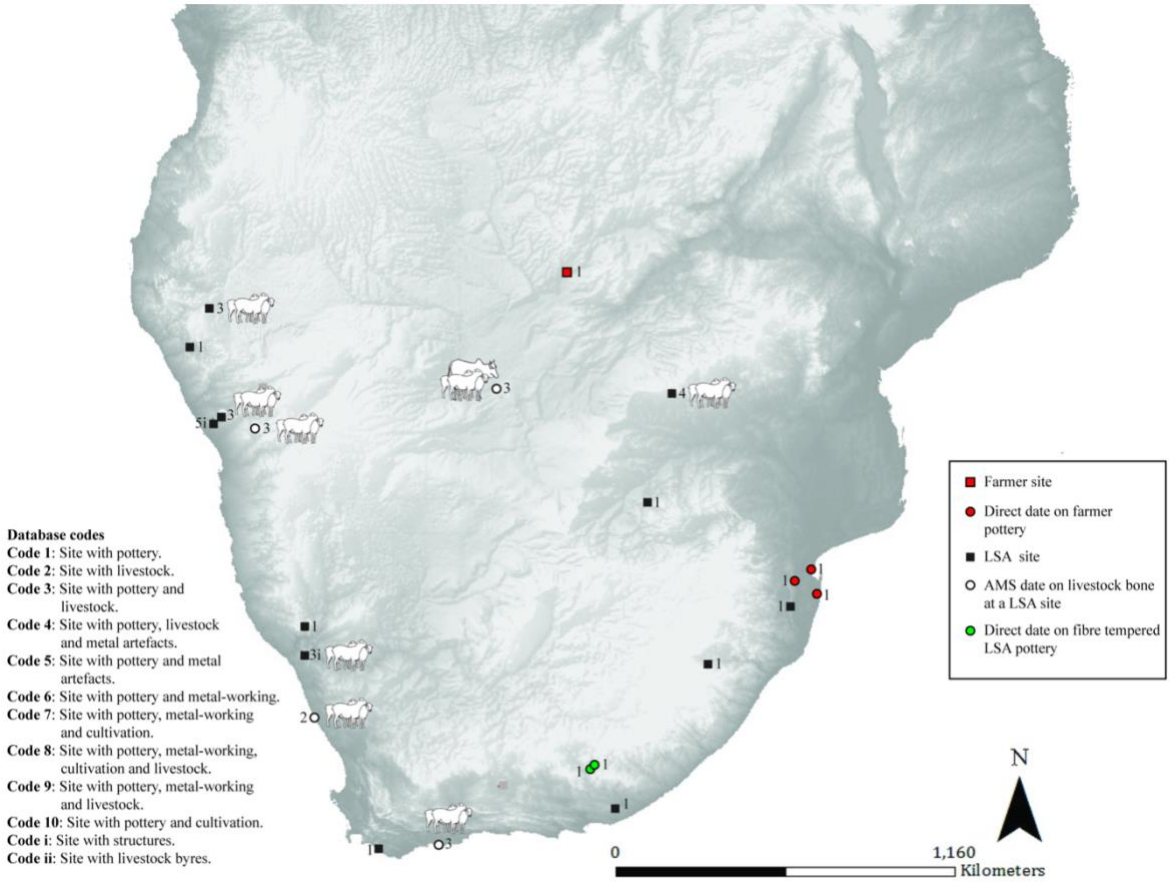

Figure 19 - Archaeological evidence for pastoralism and farming during the period 149BC – 51AD. From Lander and Russell (2018).

3.2 Genetic diversity

3.2.1 Overview of population structure

Considering the broad evidence from linguistics and archaeology (§2.1-§2.4), along with the presence of various genetic markers with an eastern African origin (§2.5), Khoe-Kwadi-speaking populations may be expected to have a common heritage. To test whether modern-day Khoe-Kwadi speakers share a genetic identity, we investigated autosomal data from 65 African populations typed on the Affymetrix Human Origins Array (**Table S 4**).

In a first step, we used Principal Component Analysis (PCA) to assess where the Khoe-Kwadi fit and how they relate to other groups from the region in their overall genomic diversity (**Figure 20**). PC1 clearly separates the autochthonous Non-Khoe Khoisan populations (Kx'a and Tuu) from the Bantu-speakers. PC2 further divides a Northern Khoisan component predominantly associated with Ju-speakers from a southern component shared between Taa and ǂ'Amkoe (ǂHoan) speakers. Even if further PCs are taken into account, our dataset lacks the power to distinguish between East and West Bantu speakers (see Semo *et al.* (2020) for an overview of Bantu genetic diversity). The Kgalagadi and Tswana (East Bantu) fall in between Khoisan and Bantu populations, due to their significant admixture with autochthonous groups (Pickrell *et al.*, 2012).

The PCA was further unable to identify a genetic component associated with Khoe-Kwadi speakers (green symbols): they appear scattered across the plot with no cluster definition. In general, Khoe-Kwadi speakers genetically resemble their closest geographic neighbours, irrespective of which language they speak. For example, speakers of the Botswanan Kalahari Khoe language G||ana are in close geographical and linguistic contact with speakers of Taa (Traill & Nakagawa, 2000) (see **Figure 1**), with whom they also share a major part of their genetic make-up. Likewise, the Khoekhoe-speaking Hai||om approach the Northern Khoisan cluster dominated by !Xun-speakers who are their immediate geographic neighbours in northern Namibia. The Naro, who are in contact with both Northern (Jul'hoan) and Southern Khoisan (Taa) fall in between both genetic clusters, in line with their geographic position.

Of particular interest is the genetic heritage of the Damara. They fall almost entirely within the Bantu cluster, indicating close contact or even language shift of a formerly Bantu-speaking population. Such a scenario has also been proposed based on mtDNA data, using ABC, which suggests a common ancestor for Herero (West Bantu) and Damara (Oliveira *et al.*, 2018). A similar proximity to Bantu-speaking populations is displayed by the formerly Kwadi-speaking Kwepe from south-western Angola, who were also shown to derive most of their uniparental lineages from the south-western Bantu gene pool (Oliveira *et al.*, 2017; Oliveira *et al.*, 2018). Other Kalahari Khoe speakers like the Khwe and Shua appear almost midway between Bantu

and Khoisan, indicating similar admixture patterns as those found with the Bantu-speaking Kgalagadi and Tswana.

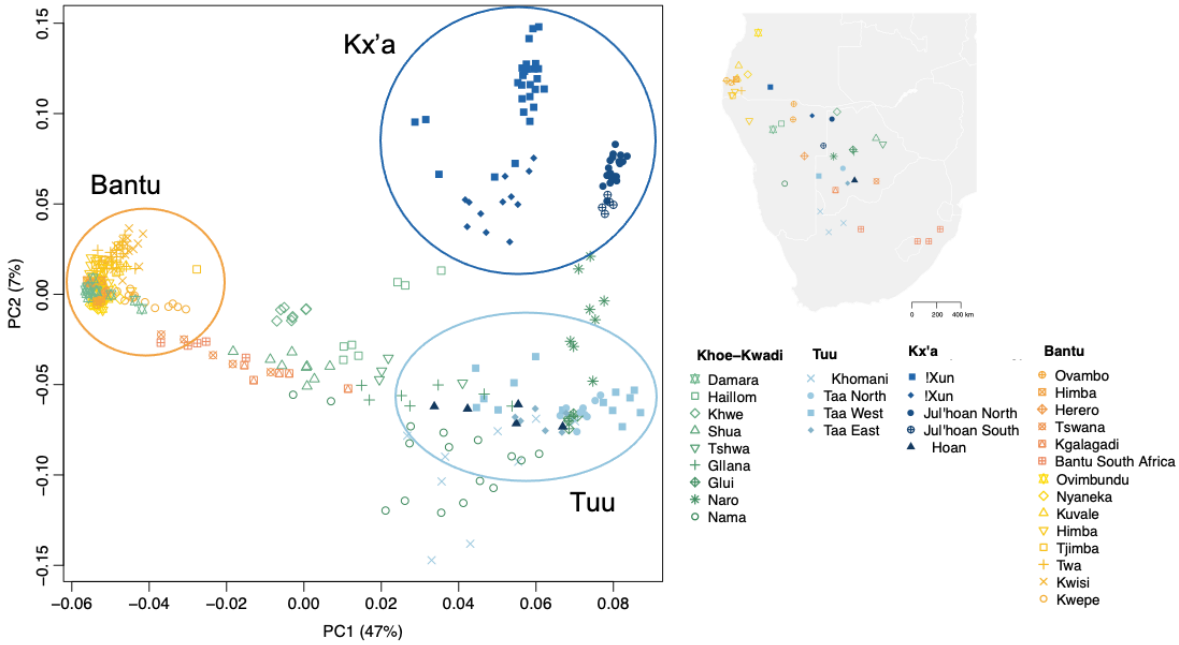

Figure 20 - Principal Component Analysis computed for the southern African Affymetrix Human Origins Array dataset. The map displays the position of the projected populations.

3.2.2 Admixture

Since PCA only provides general patterns of SNP diversity, we further computed admixture proportions for each Khoe-Kwadi-speaking group in our dataset, in order to discriminate contributions from pre-identified African populations. Since a genetic link to eastern Africa can be shown by using individual markers (§2.5), we expect that modern Khoe-Kwadi speakers are the result of admixture between present-day residents of southern Africa (Khoisan and Bantu) and an incoming population with an eastern African genetic heritage. We therefore quantified the respective contributions of autochthonous southern African foragers (Khoisan), west Africans (Bantu) and eastern African pastoralists as represented by Ju|'hoan South, Yoruba and Somali, respectively.

At first glance, the distribution of the three ancestries across Khoe-Kwadi populations is as diverse as suggested by the PCA results (**Figure 21**): the G|ui and Naro are mostly Khoisan (88% and 92%, respectively) and display only a limited amount of west African ancestry. In contrast, the Damara conform to their position in the PC-plot (**Figure 20**) and display a predominantly west African ancestry (94%), hinting at a Bantu origin. In line with their high percentages of LP (**Figure 12A**) and the skin colour allele SLC24A5 (**Figure 13B**), the Nama have the highest proportion of eastern African ancestry found in the dataset (21%). The Khwe and Shua appear highly admixed between Khoisan and Bantu but also display comparatively high percentages of eastern African ancestry (9% and 11%, respectively).

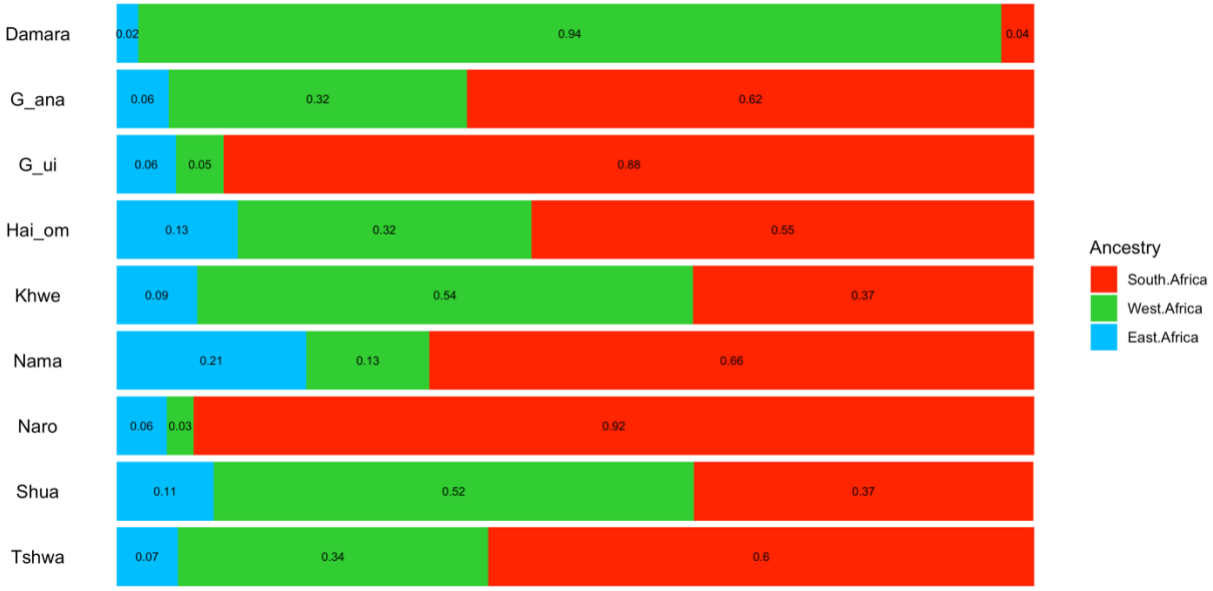

Figure 21 – Proportions of pre-identified ancestry components found in Khoe-Kwadi populations typed on the Affymetrix Human Origins Array.

Considering the tripartite genetic composition of the Khoe-Kwadi, we further assessed the best-fitting source population for each component by tracing present-day African populations to ancestral populations represented by ancient DNA samples published in Skoglund *et al.* (2017). We found that the west African component also carried by modern-day Bantu speakers best corresponds to an ancient sample from the Mende population from Sierra Leone. Interestingly, Mende and Bantu both belong to the Niger-Congo language family and therefore not only share a genetic but also a linguistic ancestor. The “Khoisan” component associated with pre-agriculture southern African foragers is best represented by DNA from an individual who lived in South Africa 2000 years BP. Finally, the eastern African component present in southern Africa can be linked to an early pastoralist burial site from eastern Africa found in Luxmanda, Tanzania (3100BP).

When ancient DNA data is used as a proxy for the ancestral components found in Khoe-Kwadi populations (**Figure 22**), no significant difference from the overall frequencies obtained with modern proxies is observable: east Africa (Mann-Whitney U test; p=0.112), south Africa (Mann-Whitney U test; p=0.4797) and west Africa (Mann-Whitney U test; p=0.8598). Nevertheless, some minor differences arise: the low frequencies of eastern African ancestry detected in the Damara using modern proxies disappear entirely when ancient proxies are used; the same observation applies to western African ancestry in the G|ui and Naro.

While confirming general patterns observed in the PCA, both admixture analyses reveal that all Khoe-Kwadi speakers share a genetic component with a clear link to eastern Africa. However, different populations show varying degrees of eastern ancestry, which in general appears to have been strongly diluted by the regional southern African components associated with Non-Khoe Khoisan and Bantu speakers.

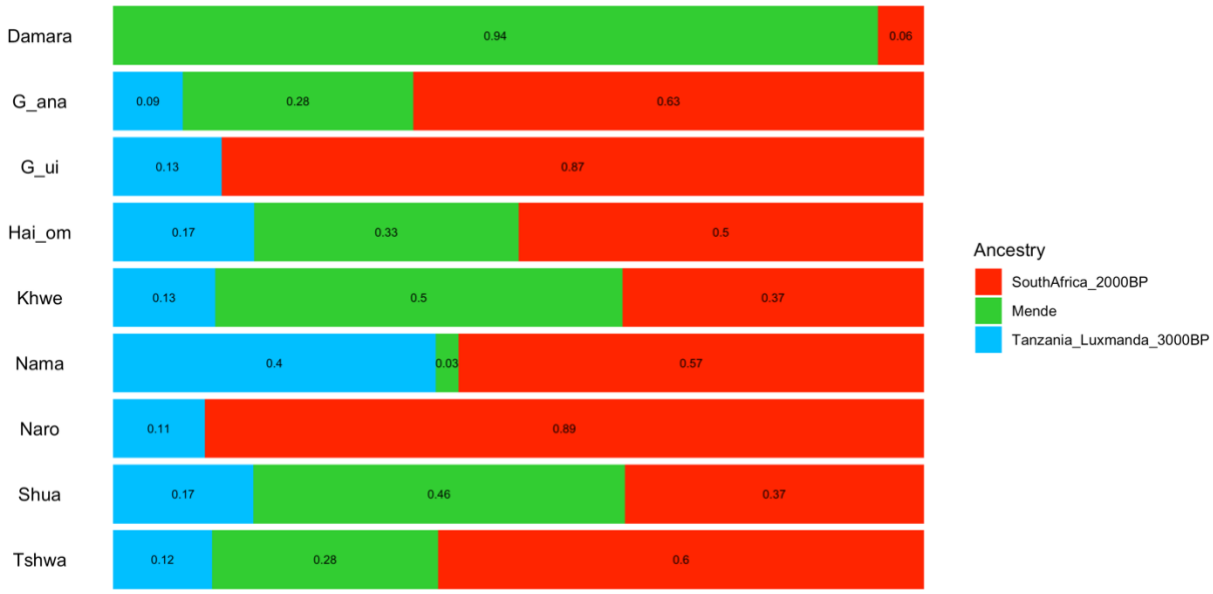

Figure 22 – Proportions of ancestry components with ancient DNA proxies found in Khoe-Kwadi populations typed on the Affymetrix Human Origins Array.

3.2.3 Structure of the Khoisan component

As previous analyses have shown, all Khoe-Kwadi populations show major contributions from southern African "Khoisan" and/or Bantu populations. To see whether those contributions can be assigned to subclusters, *i.e.*, Northern vs Southern Khoisan (**Figure 20**) or Eastern vs Western Bantu, we performed a masking-analysis for both "Khoisan" and Bantu components: we assigned the SNP data to either a "Khoisan" or Bantu ancestry and masked the remaining data. We then projected the masked data to compare it with other non-masked populations from our dataset. The analysis proved to be non-informative for the Bantu component, as the Affymetrix SNP data is not significantly informative to separate the two Bantu clusters (see above). The assigned "Khoisan" ancestry of our Khoe-Kwadi speaking groups was compared to two populations maximally representative of the Northern (Ju|'hoan North) and Southern Khoisan (Taa west) component.

Results (**Figure 23**) show that the majority of populations belonging to the Kalahari Khoe subgroup of Khoe-Kwadi show an affinity to the Taa west and thereby to the Southern Khoisan cluster. Exceptions are constituted by the Khwe who drift towards the northern cluster, and the Shua and Naro who are placed midway in between both subgroups. This observation is in line with both the PCA results and the geographic location of the studied groups, showing the Naro lodged between northern and southern cluster, while Khwe and Shua not only approach the Bantu speakers but also display less of a tendency to approach the southern cluster, compared to other groups (**Figure 20**, PC2). While the Naro are in contact with populations from both the northern and the southern cluster, Khwe and Shua are spread along the northern Kalahari Basin fringe (**Figure 20**). Therefore, they may have been in more intensive contact with !Xun-speakers conforming to a Northern Khoisan genetic component. While the Khwe stayed in the north-west and probably had little contact with Tuu speakers, the Shua penetrated further towards the south-east and may have interacted with Tuu speakers after having acquired a substantial portion of Northern Khoisan ancestry from !Xun-speakers they encountered along the way.

Among Khoekhoe-speaking populations, only the Nama display a clear affinity to the cluster and, thereby, to the Tuu. This result is further reinforced by the close similarities between the Nama and the formerly !Ui-speaking ǂKhomani (Uren *et al.*, 2016) which are also visible in the PCA (**Figure 20**). The Hai||om, Khoekhoe-speaking foragers from northern Namibia, are slightly inclined towards the northern (!Xun-speaking) cluster. Again, this reflects the geographical proximity between Hai||om and !Xun, and may also be seen as support for the hypothesis that the Hai||om were previously !Xun-speakers who interacted with Khoekhoe pastoralists and subsequently changed their language (Güldemann, 2014; Pickrell *et al.*, 2012). The little "Khoisan" contribution found in the Damara is located in between the northern

and the southern cluster. Whether this reflects influence from north and south, from a third "Khoisan" population equidistant from both, or insufficient resolution to differentiate either, this cannot be answered with the data at hand.

In general, the "Khoisan" contributions found in Khoe-Kwadi populations from our sample are in line with geography: Kalahari-Khoe and Khoekhoe speakers, living in the south-east of the Kalahari Basin, are at close range with Tuu and ǂ'Amkoe speakers with whom they share a substantial part of their genetic profiles. In contrast, populations from the northern Kalahari Basin fringe, *i.e.* the Khwe, Shua and Haillom, display increased contributions from the northern cluster dominated by !Xun-speaking populations.

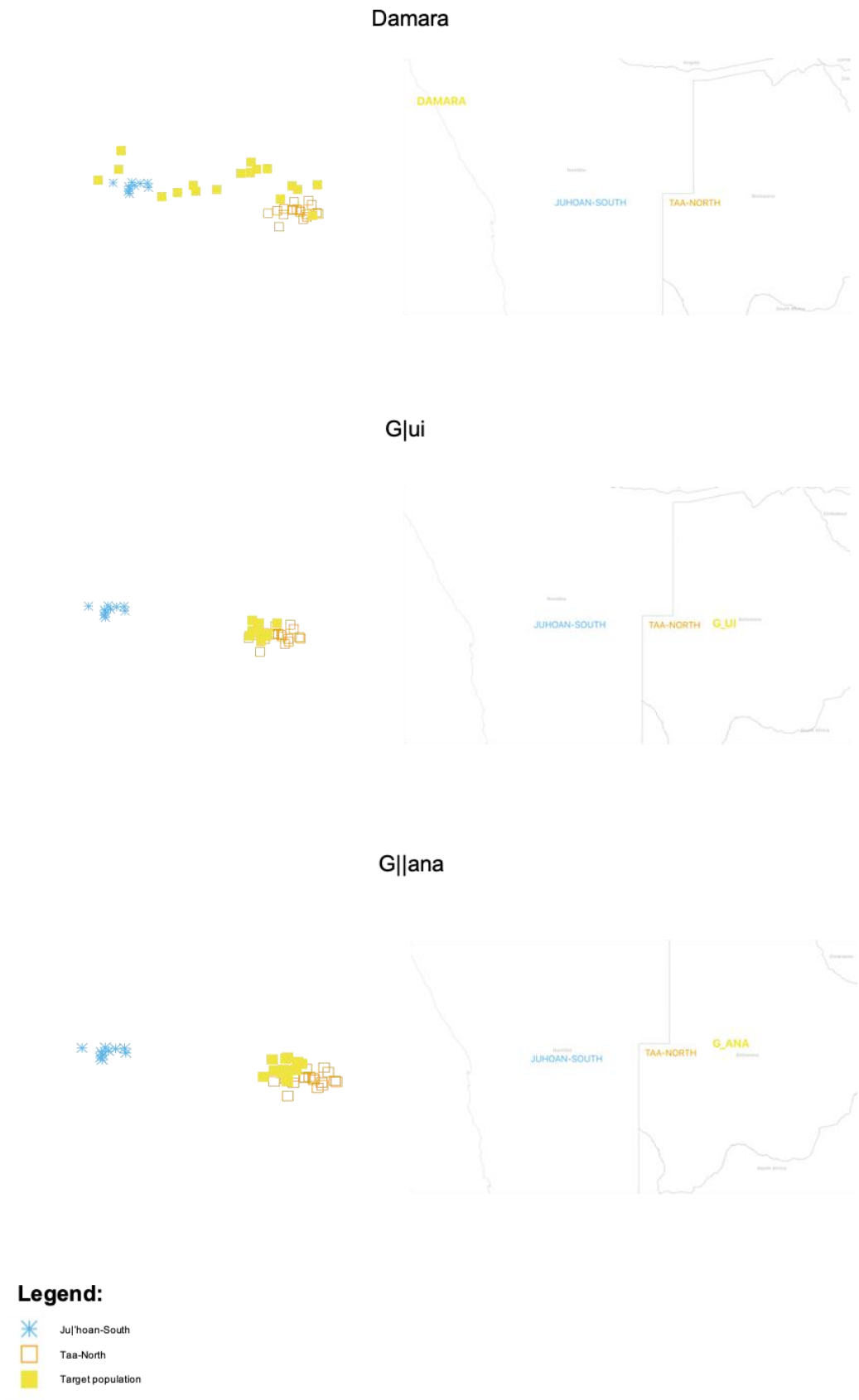

Figure 23 - Principal Component Analysis computed for the masked “Khoisan” genomes of each Khoe-Kwadi population (left) with accompanying maps (right).

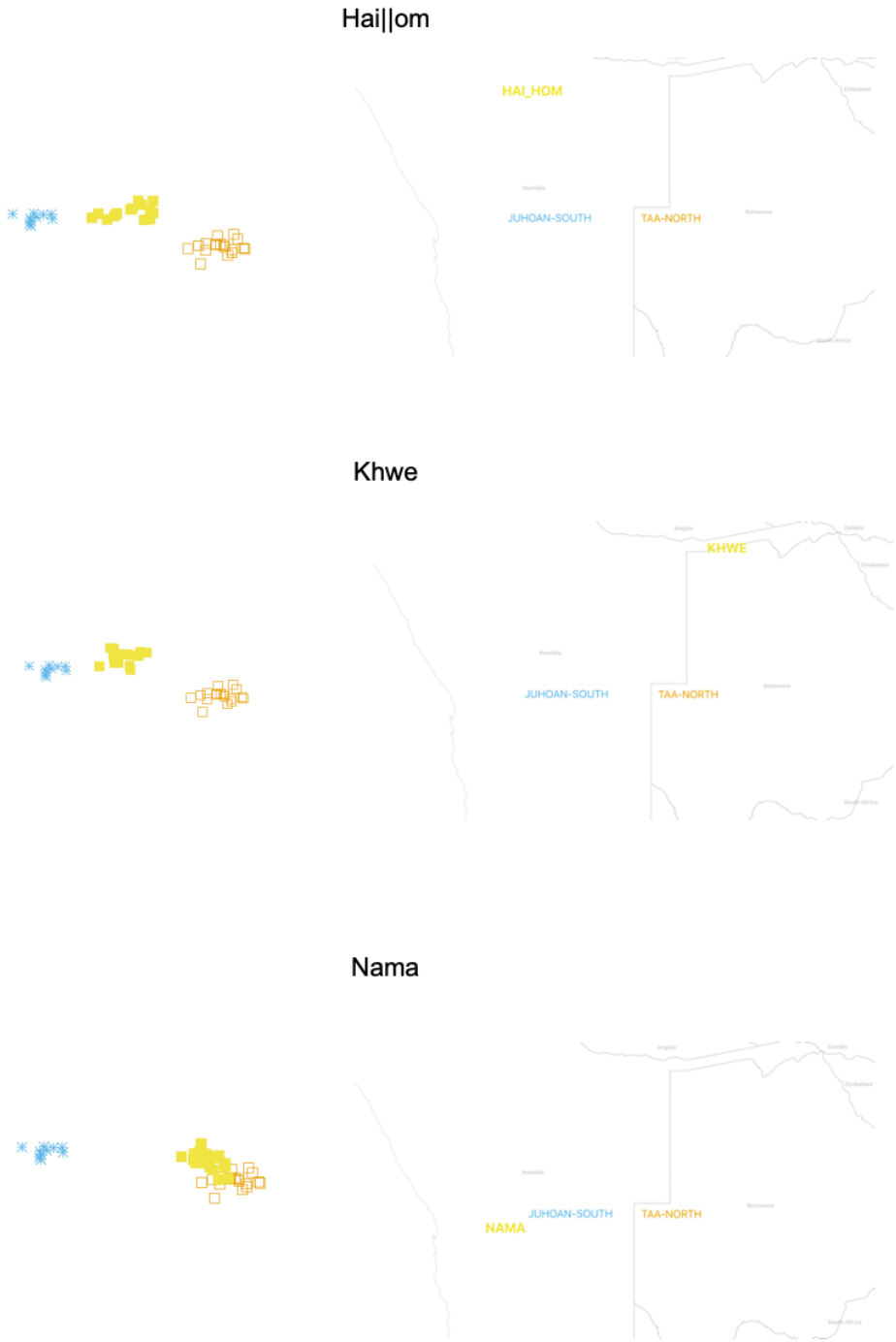

Figure 23 (cont.) - Principal Component Analysis computed for the masked "Khoisan" genomes of each Khoe-Kwadi population (left) with accompanying maps (right).

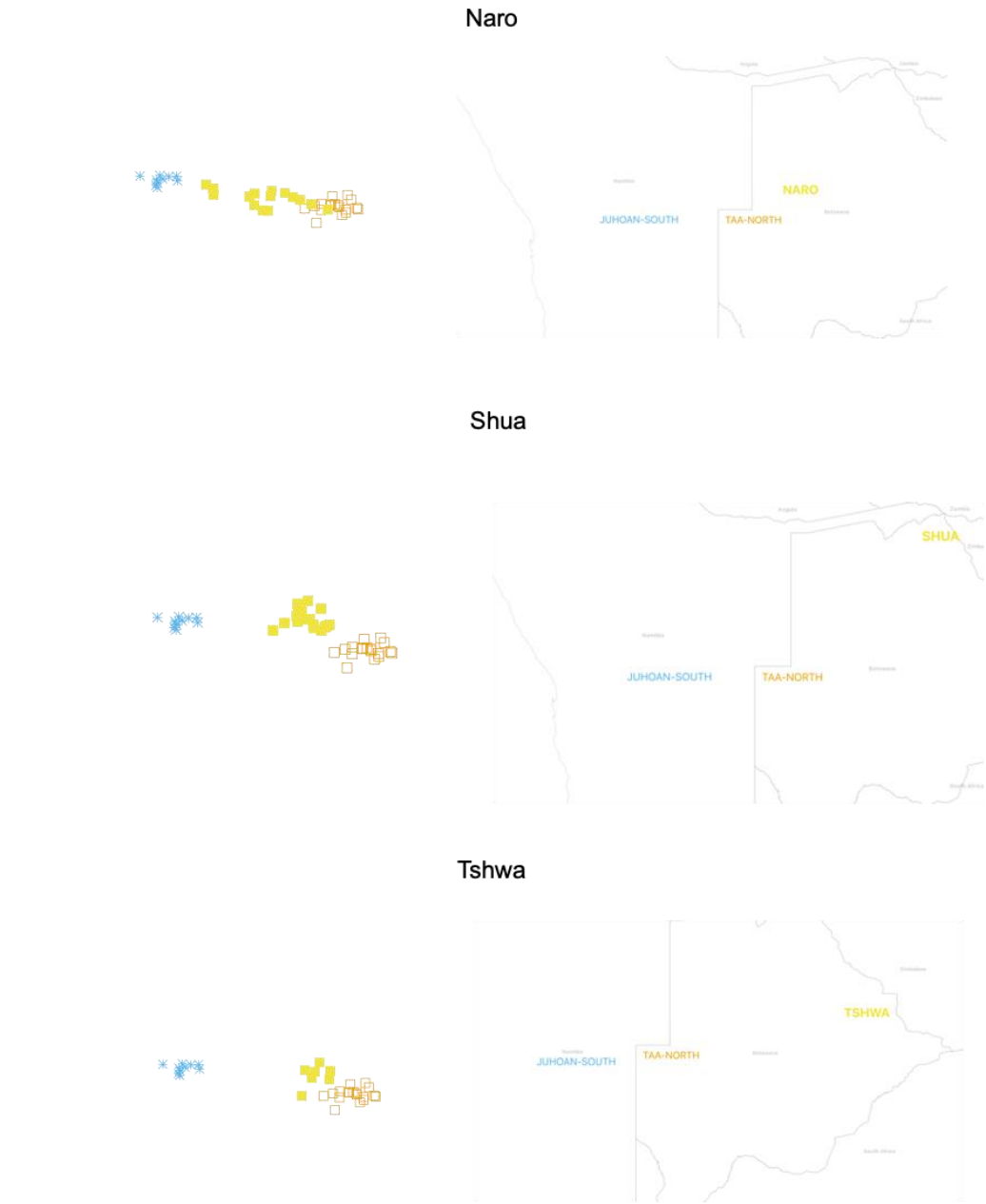

Figure 23 (cont.): Principal Component Analysis computed for the masked "Khoisan" genomes of each Khoe-Kwadi population (left) with accompanying maps (right).

3.2.4 IBD sharing

To arrive at more in-depth contact profiles for the groups sharing genetic material, we opted for an "identity by descent" (IBD) approach which allows us to characterize interactions within the last ~3000 years, according to the number and length of IBD segments shared between populations (Al-Asadi *et al.*, 2019). In a subsequent step, we merged the IBD analysis with the ancestry-assignment analysis (§3.2.2) in order to infer the proportion of shared segments belonging to a west African, south African and east African ancestry, respectively.

Results indicate that most Khoe-Kwadi share more IBD segments with Khoisan populations (Kx'a and Tuu) than with any Non-Khoisan group[2] (**Figure 24**). Exceptions are constituted by the Damara (**Figure 24.1**), Khwe (**Figure 24.5**) and Shua (**Figure 24.8**). They share a substantial part of their ancestry with western African populations, in accordance with their elevated Bantu contributions discussed in previous sections (**Figure 20**). Different patterns of ancestry-driven IBD sharing for lengths ranging from 1 to 5 cM are in the following, exemplified by a closer examination of Kalahari Khoe speakers from the G|ui, Khwe and Shua populations:

Previous analyses have established the G|ui as a population sharing close genetic similarities with Non-Khoe speakers from the central Kalahari (**Figure 21**, **Figure 22**, **Figure 23**). The IBD analysis (**Figure 24.3**) confirms this trend, identifying Taa and ǂHoan[3] speakers as major contributors to their genetic make-up. While they also share a considerable number of IBDs with the Bantu-speaking Kgalagadi (**Figure 20**) (Pickrell *et al.*, 2012), these reflect the Khoisan-admixture found in this Sotho-Tswana-speaking group, rather than a substantial amount of Bantu ancestry in the G|ui.

As expected from the masking analysis, the Khwe display predominant IBD-sharing with Ju populations (**Figure 24.5**), especially with the geographically close !Xun from Angola and northern Namibia (**Figure 1**). While the Khwe share approximately half of their assigned ancestry with southern African Khoisan populations, they also display considerable IBD sharing with populations from western Africa. In pace with their geographic location, the bigger part of their western African (Bantu) IBD fragments (1 to 5cM) derives from south-western Bantu speakers, like the Herero. The link to south-western Bantu speakers is even more evident when IBD fragments ranging from 5 to 10 cM are used (**Figure S 8**).

[2] We attributed the Non-Khoisan category to populations that speak Afro-Asiatic, Niger-Congo and Nilo-Saharan languages.

3 Although the ǂHoan belong to the Kx'a language family, this population resides among Tuu and Kalahari-Khoe speakers in the central Kalahari (**Figure 1**); their genetic profile is near-identical with Taa-speakers from the region (Pickrell et al., 2012).

The Shua, like the Khwe, share almost the same number of IBDs with Khoisan and Non-Khoisan (**Figure 24.8**) populations. However, within their Khoisan component, they share most IBDs with the central Kalahari populations (ǂHoan[3] and Taa); this association becomes even stronger when shared IBDs of 5 to 10cM are considered (**Figure S 8.8**). Concerning their Non-Khoisan contributions, the Shua are also different from the Khwe in sharing the greatest number of western African IBDs with eastern Bantu populations, here represented by BantuSA, Tswana and Kgalagadi in our dataset.

Like established in previous analyses, the variation in patterns of IBD sharing between Khoe-Kwadi speakers and other African populations mostly depends on the groups' geographic location and their association with neighbouring Non-Khoe and Bantu-speaking groups: While groups from the central Kalahari display a higher amount of shared IBDs with Taa, ǂHoan and East Bantu-speakers, populations from the northern fringe, especially the Khwe, display a stronger link to Ju and south-West Bantu.

As our Affymetrix Array data is not well-suited to separate intra-Bantu genetic diversity, it is sometimes hard to discriminate the IBD contributions of individual Bantu populations, especially in the smaller IBD segments (1 to 5 cM) that may represent older sharing, prior to the separation of different Bantu groups (cf. **Figures S8**).

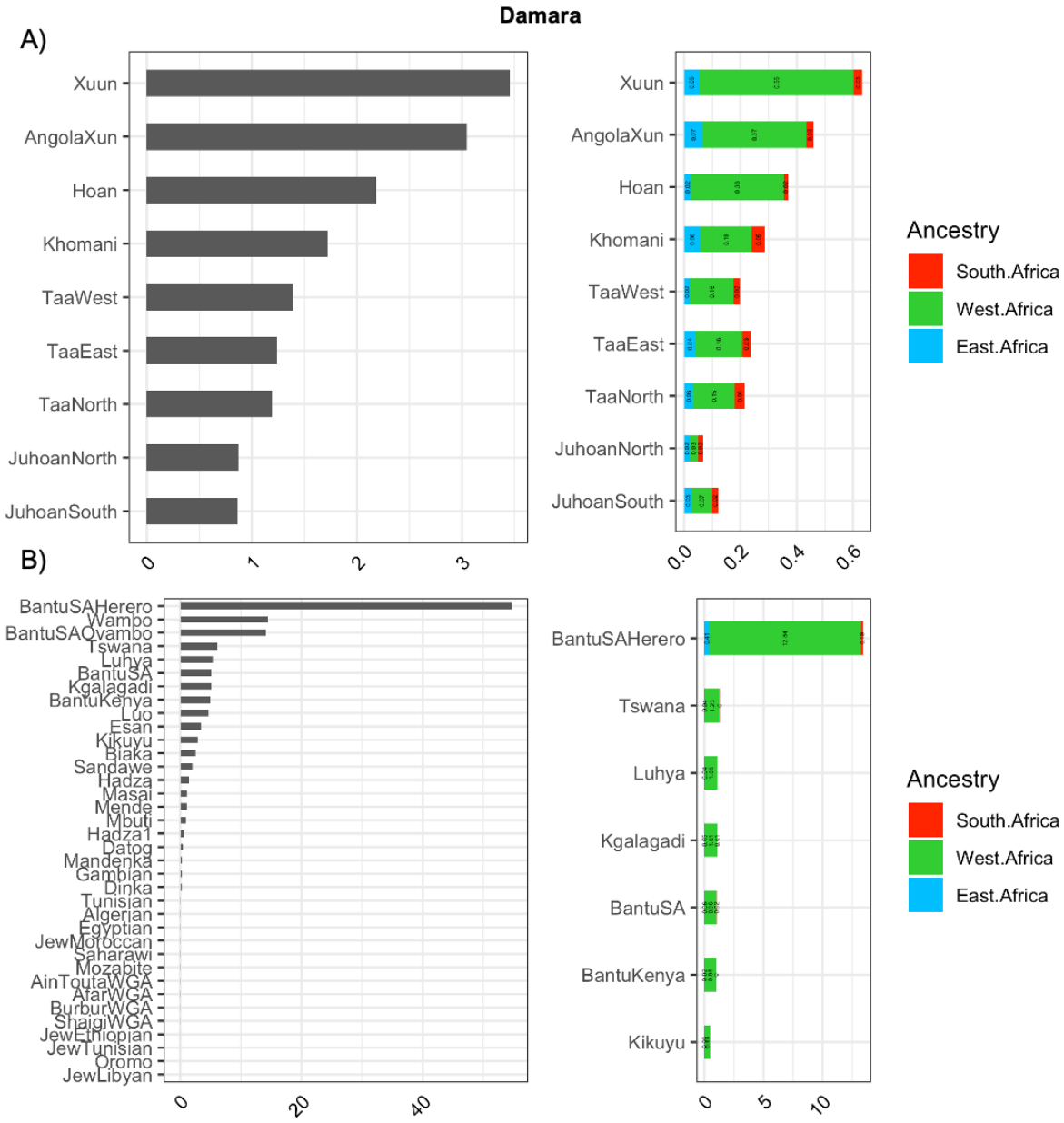

Figure 24.1 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Damara and Khoisan (A) and Non-Khoisan (B) populations.

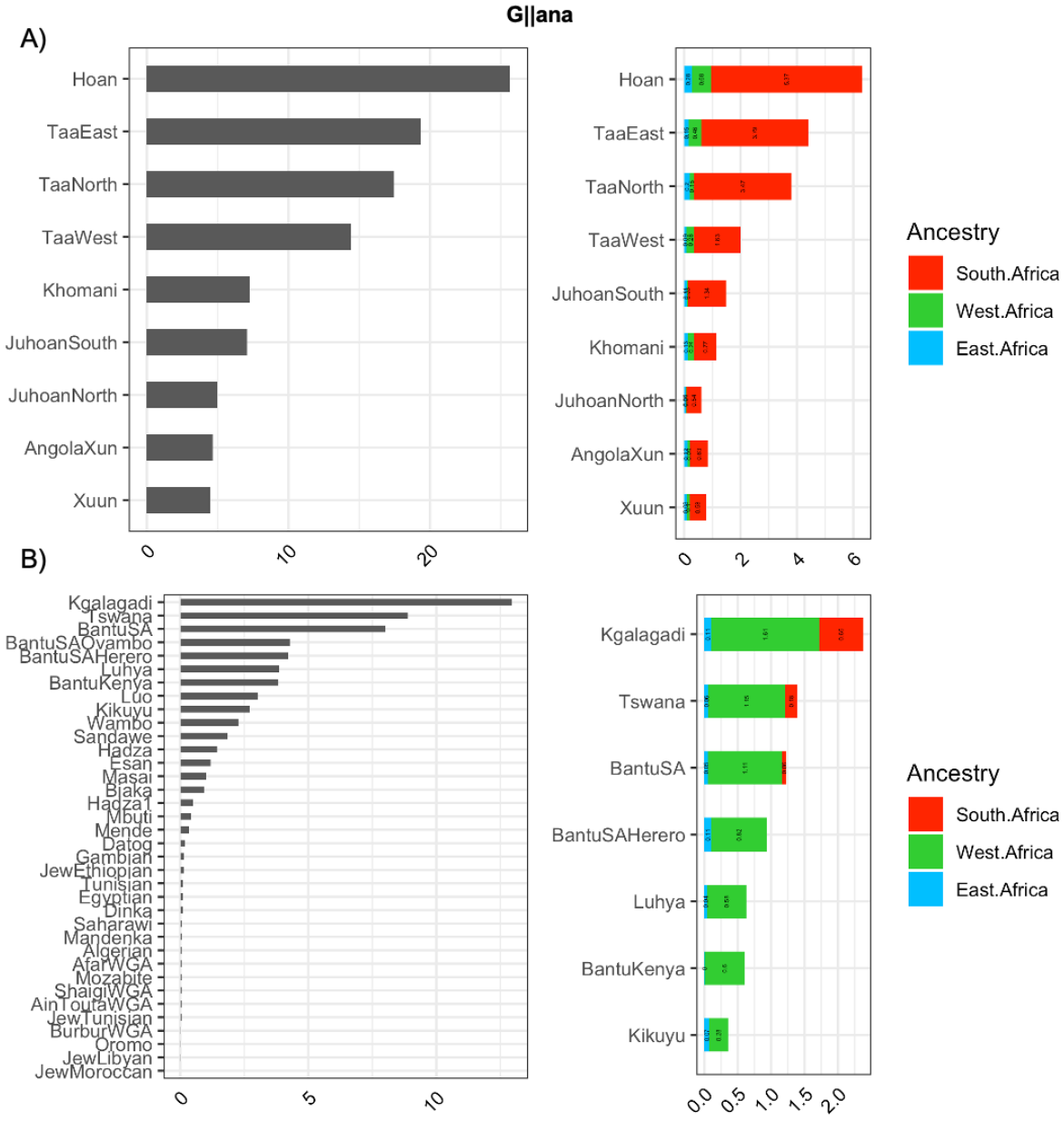

Figure 24.2 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the G||ana and Khoisan (A) and Non-Khoisan (B) populations.

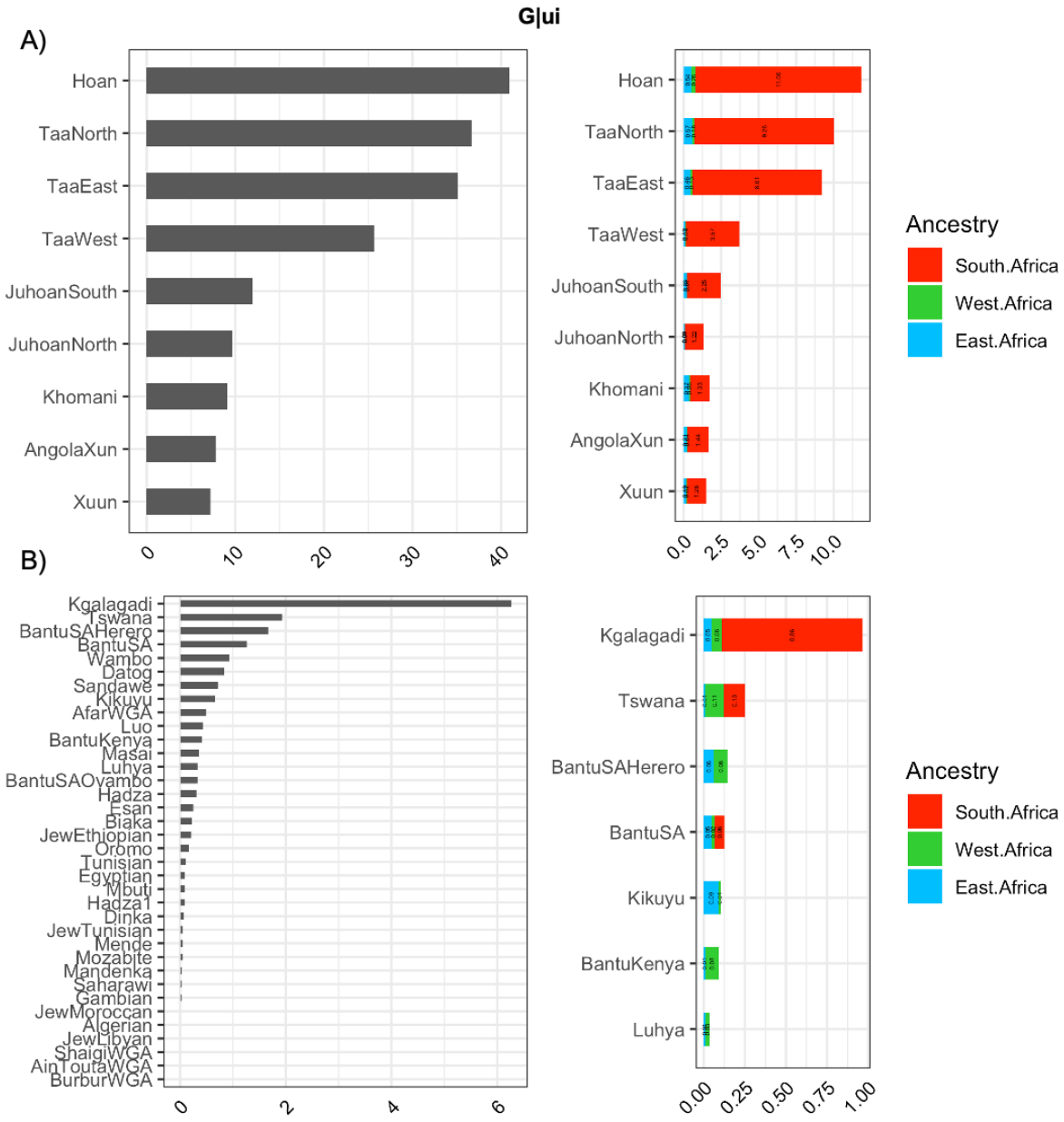

Figure 24.3 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the G|ui and Khoisan (A) and Non-Khoisan (B) populations.

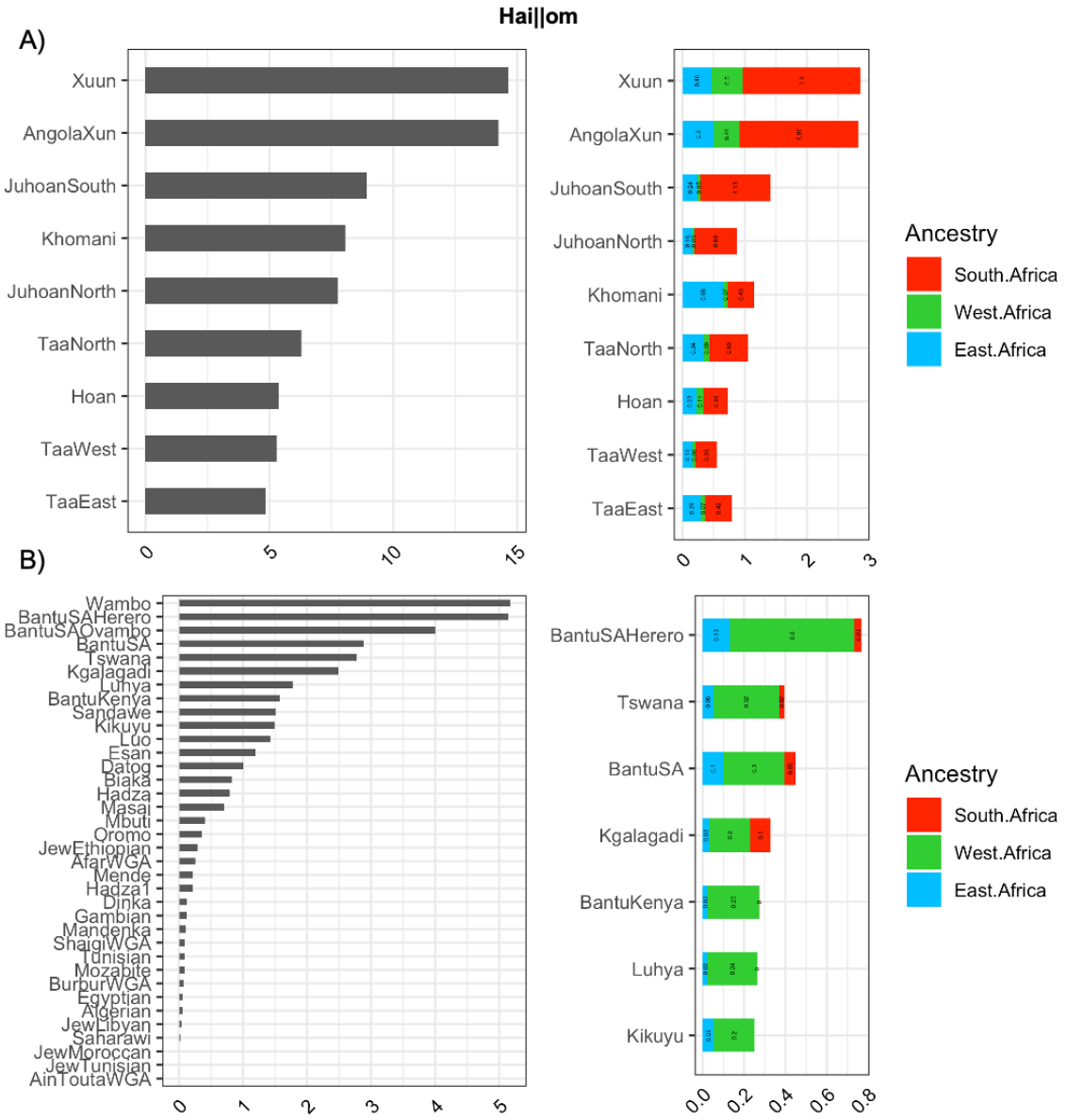

Figure 24.4 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Hai||om and Khoisan (A) and Non-Khoisan (B) populations**.**

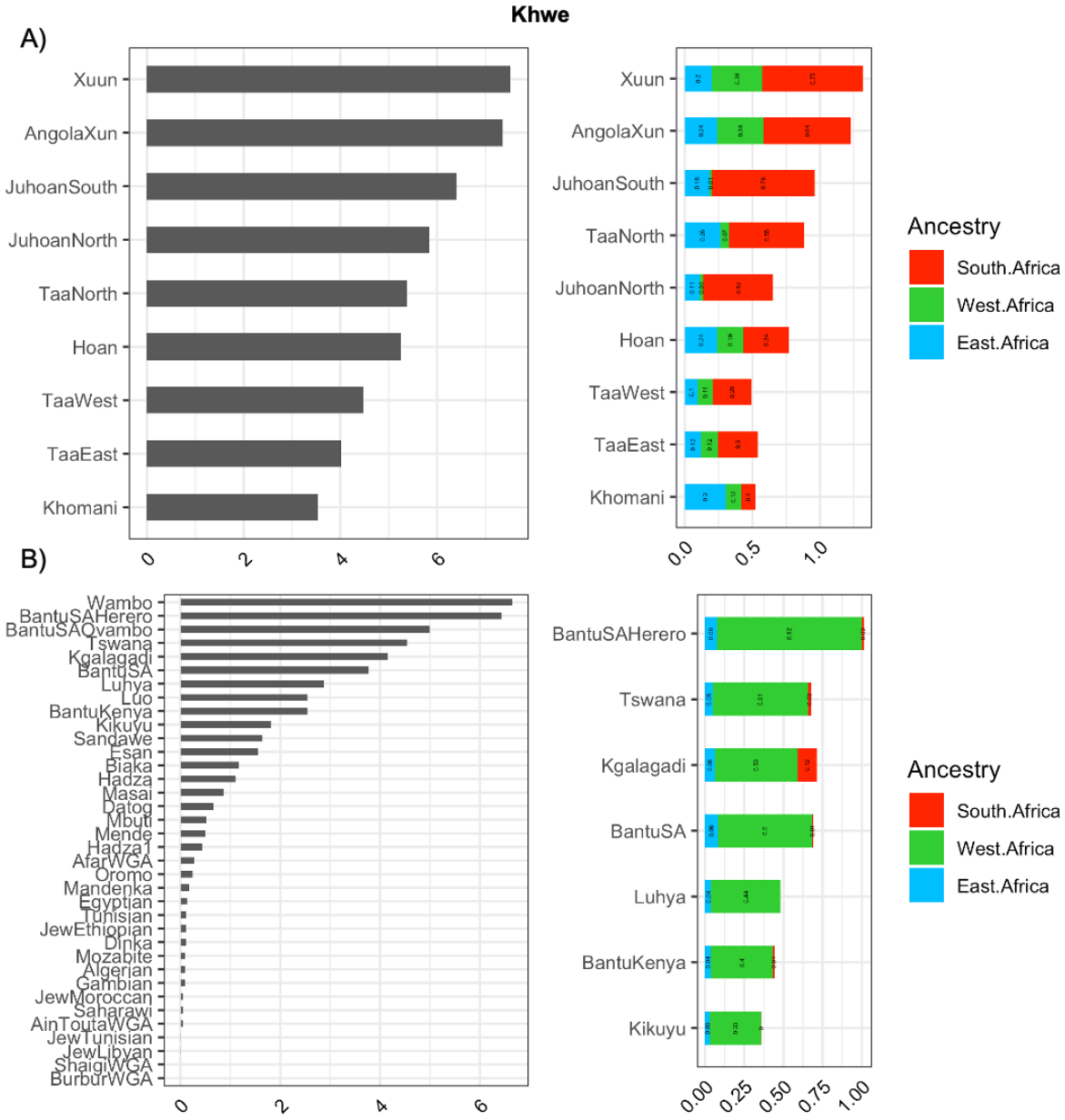

Figure 24.5 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Khwe and Khoisan (A) and Non-Khoisan (B) populations.

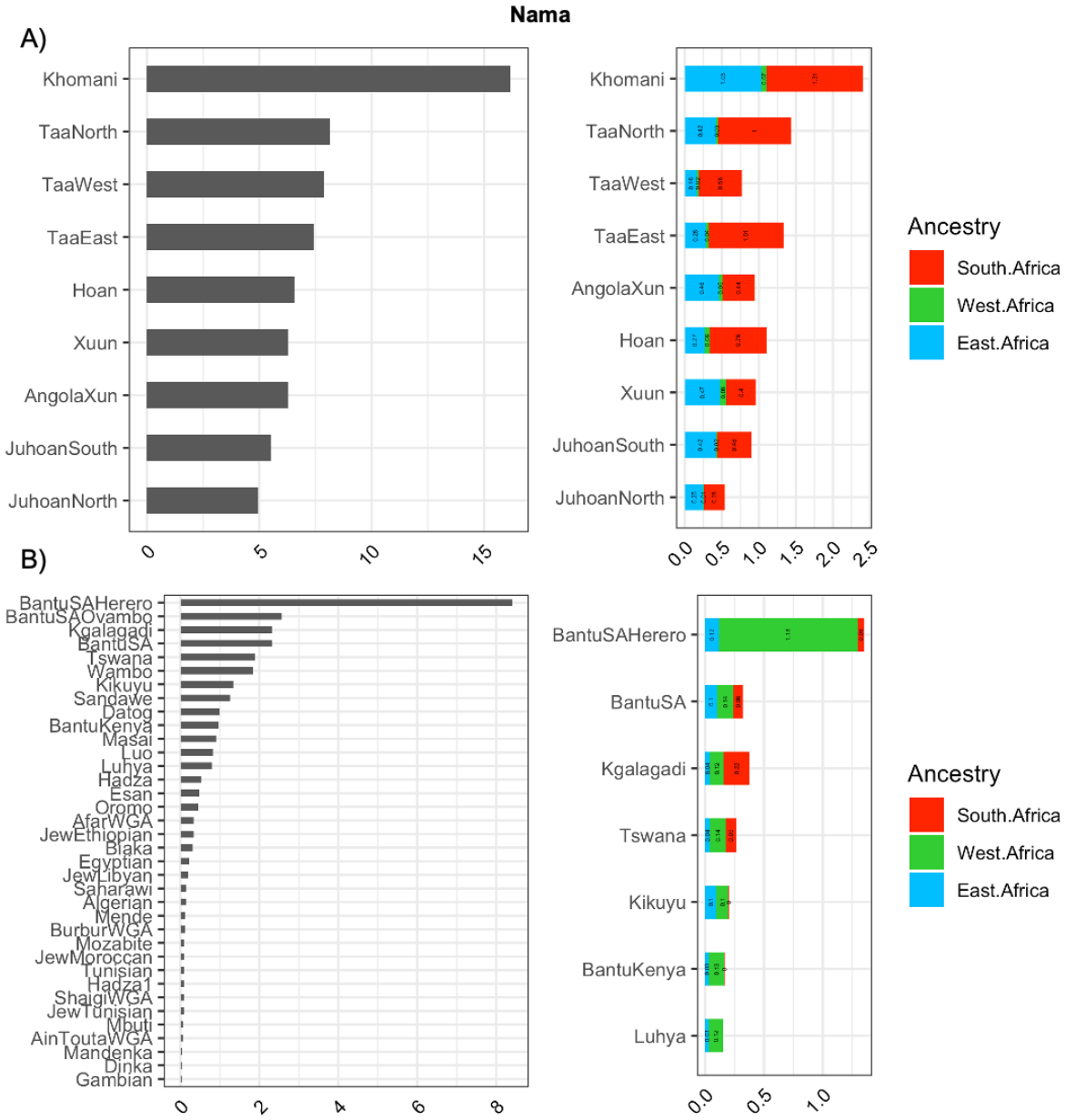

Figure 24.6 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Nama and Khoisan (A) and Non-Khoisan (B) populations.

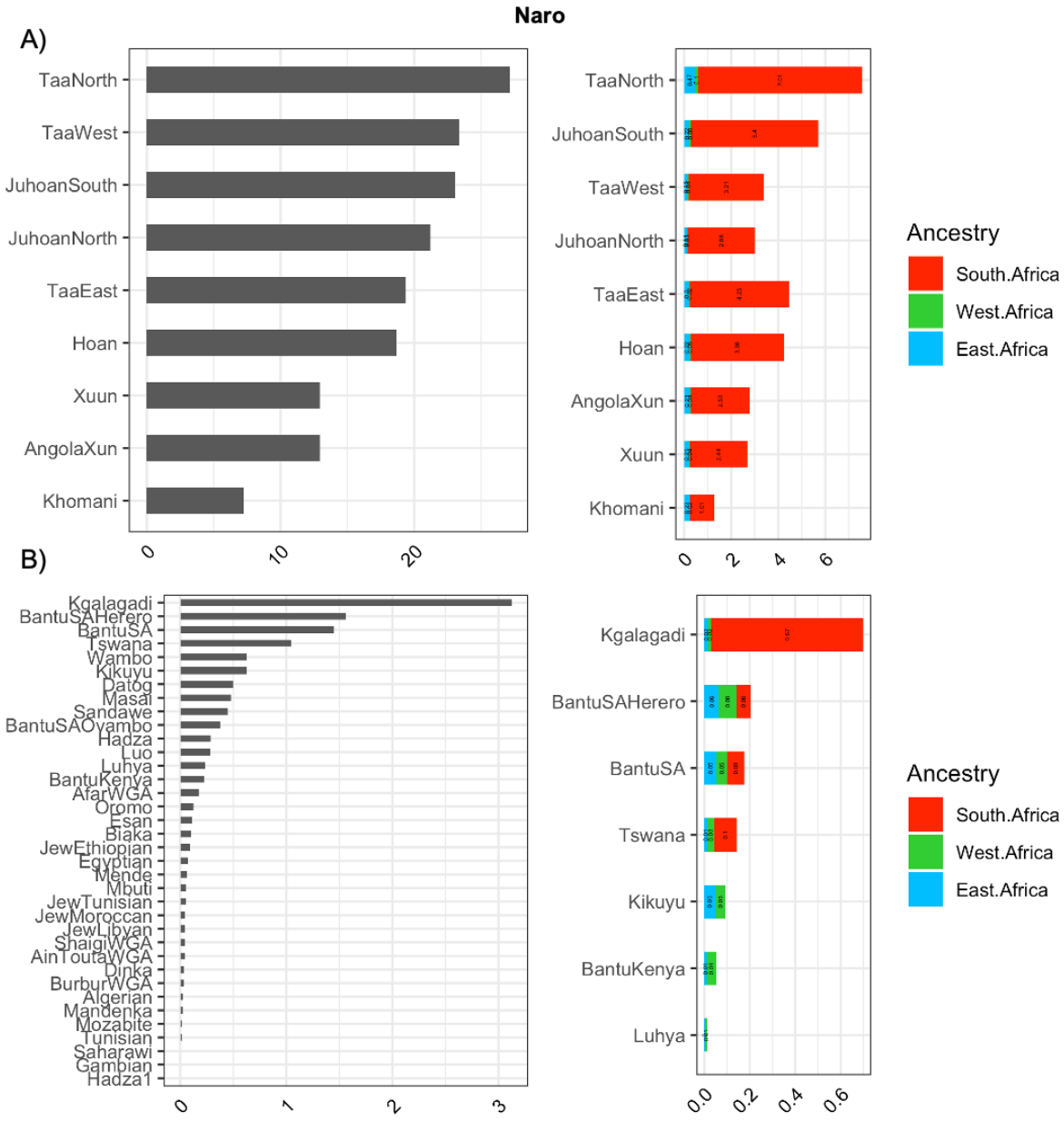

Figure 24.7 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Naro and Khoisan (A) and Non-Khoisan (B) populations.

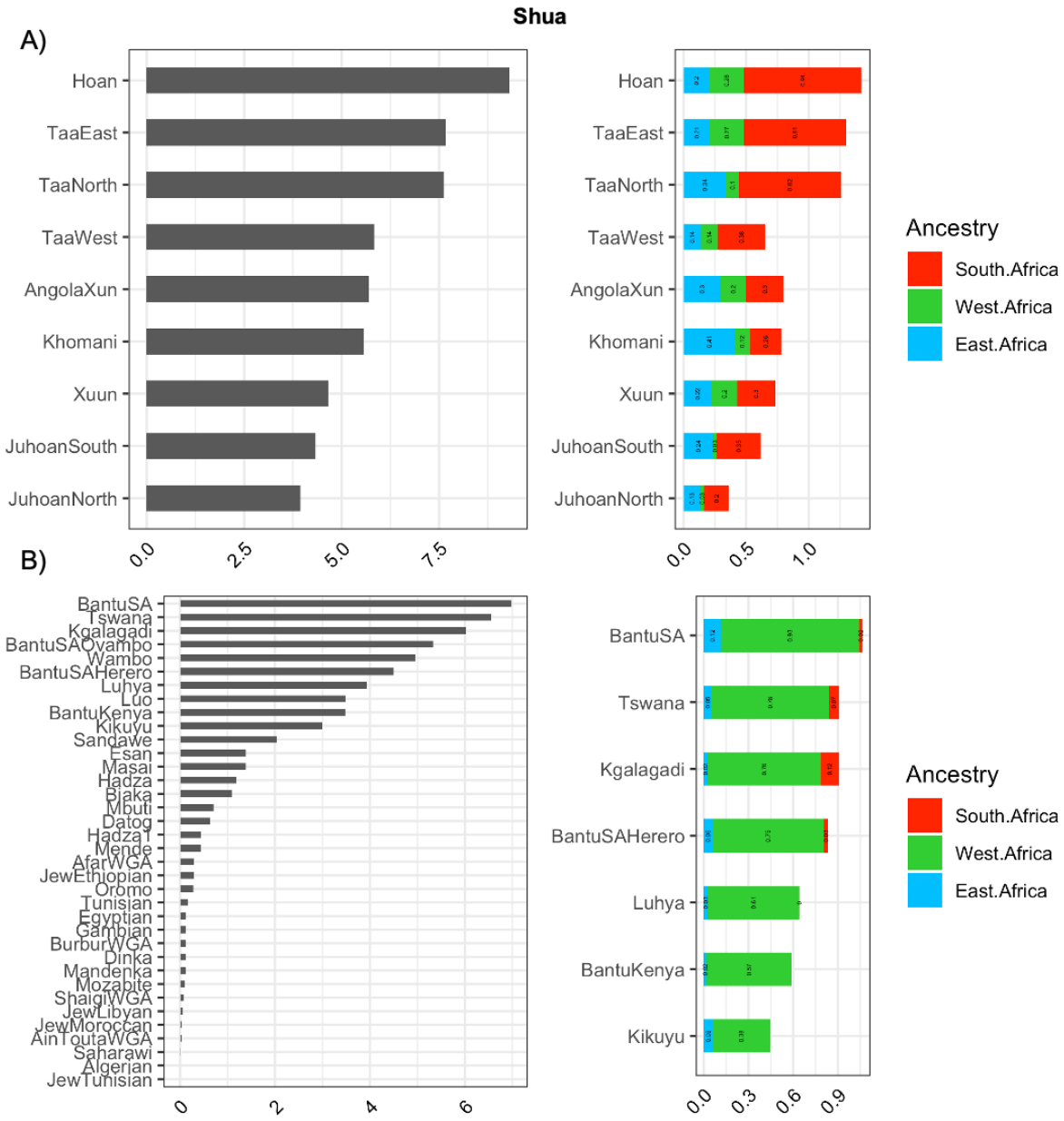

Figure 24.8 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Shua and Khoisan (A) and Non-Khoisan (B) populations.

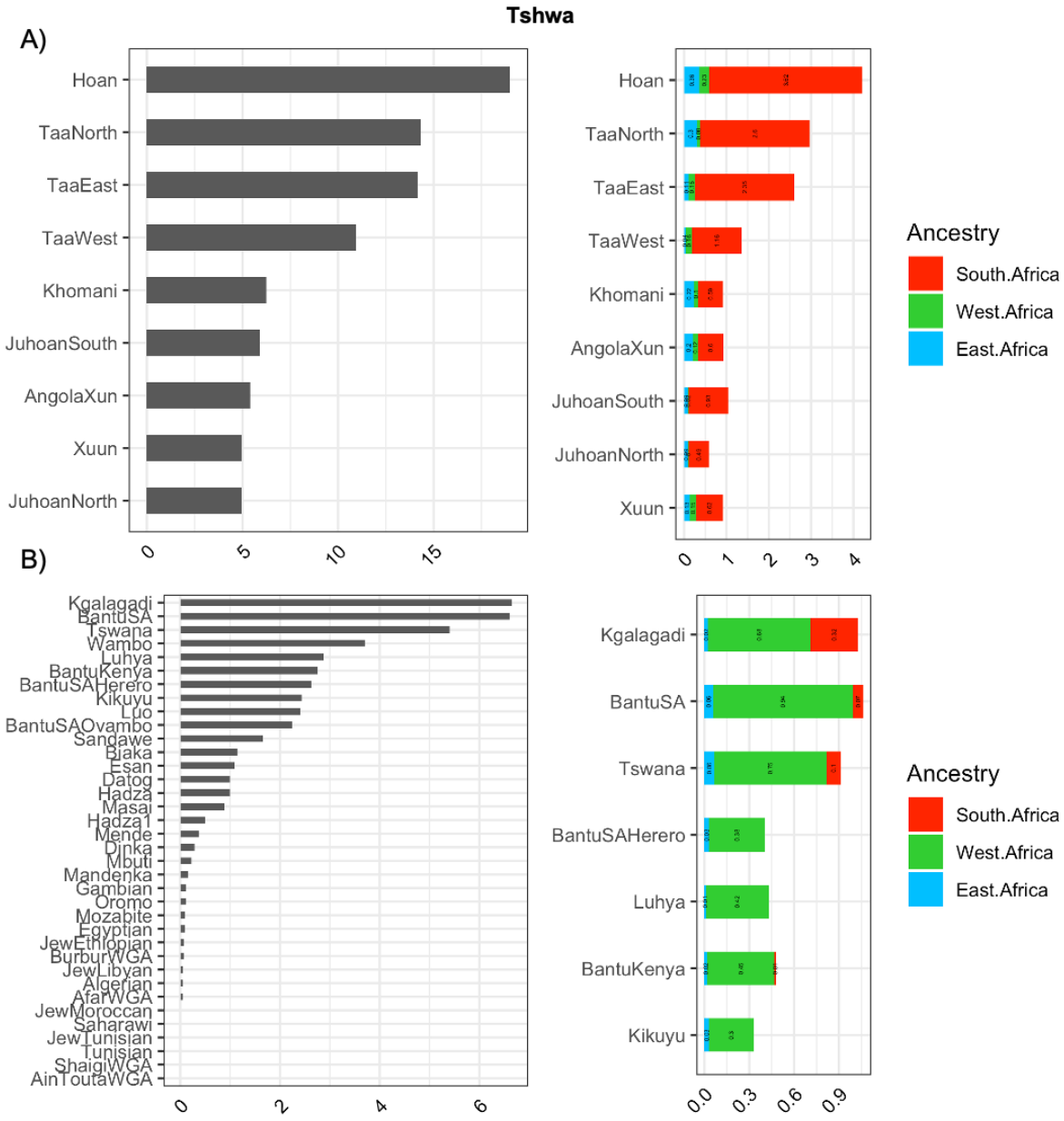

Figure 24.9 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Tshwa and Khoisan (A) and Non-Khoisan (B) populations.

3.2.5 The Namibe populations

From 2013 to 2014, the HUMANEVOL group conducted several field missions to the Namibe and Kunene provinces in Angola, in order to contact ethnic groups representing the various population layers found in the region (**Figure 25**). While a small number of ethnographic studies had been conducted prior to the study, linguistic and genetic data from south-western Angola had previously been near-absent.

A representative of the Bantu layer are the Kuvale and Himba, who form part of the broader Herero ethnolinguistic division and inhabit the semi-arid areas of the Angolan Namib Desert. Their reliance on a strong pastoral tradition clearly separates them from other south-West Bantu speakers from the area, in particular, the agropastoral Nyaneka-Nkhumbi and Ovambo, with whom they share close linguistic ties (Mitchell, 2010; Vansina & Fansina, 2004; Vivelo, 1977). The Kuvale and Himba exert dominance over ethnic minorities in the Namibe desert as exemplified by the dissemination of their languages and cultural attire (Estermann, 1962, 1981). These minorities are best defined as peripatetic peoples, *i.e.* small-scale groups providing specialized goods and services to dominant populations who do not consider them equals (Bollig, 2004). In the Namibe desert, they are represented by the Kwepe, as well as by the Kwisi, Twa and Tjimba, who have previously been linked to an enigmatic pre-Bantu layer of Non-Khoisan foragers (MacCalman & Grobbelaar, 1965).

The Kwepe are the only group known to have spoken the now extinct Kwadi branch of the Khoe-Kwadi language family (Güldemann, 2004). They were first described in the 19th century by Portuguese explorers as a group of click-speaking impoverished shepherds that physically resemble their Bantu-speaking neighbours (Oliveira, 2019). By the middle of the 20th century, the Kwepe were on the verge of extinction (De Almeida, 1965), and nowadays only a few hundred individuals live on the margins of the Kuroka River in the Namibe province (Oliveira, 2019). They keep a small stock of sheep and provide services to their dominant Kuvale neighbours with whom they share their current language (Fehn, 2019b; Oliveira *et al.*, 2018). Following the Portuguese anthropologist António de Almeida and the South African linguist Ernst O.J. Westphal who collected Kwadi linguistic data from four elders, our group was also able to collect linguistic recordings from two remembers of the Kwadi language (Almeida, nd; Oliveira *et al.*, 2018; Westphal, 1963).

The Kwisi and Twa are described in ethnographic studies as foragers that belong to one of the oldest population layers of southern Africa (De Almeida, 1965; Estermann, 1976). However, due to their physical appearance which differs from the Khoisan foragers of the Kalahari, they are thought to be descendants of a distinct dark-skinned forager stratum which would also include the Damara from Namibia (Barnard, 1992; De Almeida, 1965; Estermann, 1976). As they nowadays speak Bantu, it has been suggested that the Kwisi and Twa lost

their original language and shifted to Kuvale, due to pressure from the dominant group (Estermann, 1976; Redinha, 1975). A similar scenario has been proposed for the Damara, who presently speak the Khoekhoe variety Nama (Barnard, 1992). While the Twa are proud of their foraging tradition, the Kwisi consider their name derogatory and prefer to hide their identity which signifies low social status among their Bantu neighbours (Estermann, 1976; Oliveira *et al.*, 2018). At present, the Twa and Kwisi are no longer hunter-gatherers, but keep small stock and are involved in patron-client relationships with the dominant pastoral groups, providing services like blacksmithing, healing and sorcery (Bollig, 2004; Oliveira *et al.*, 2018). A further group, the Tjimba, live among the Himba and speak their language (Oliveira *et al.*, 2018). Although it has been hypothesized that they belong to the same pre-Bantu layer as the Kwisi, Twa, and Damara, they are often described as impoverished Himba who lost their cattle (MacCalman & Grobbelaar, 1965; Vivelo, 1977).

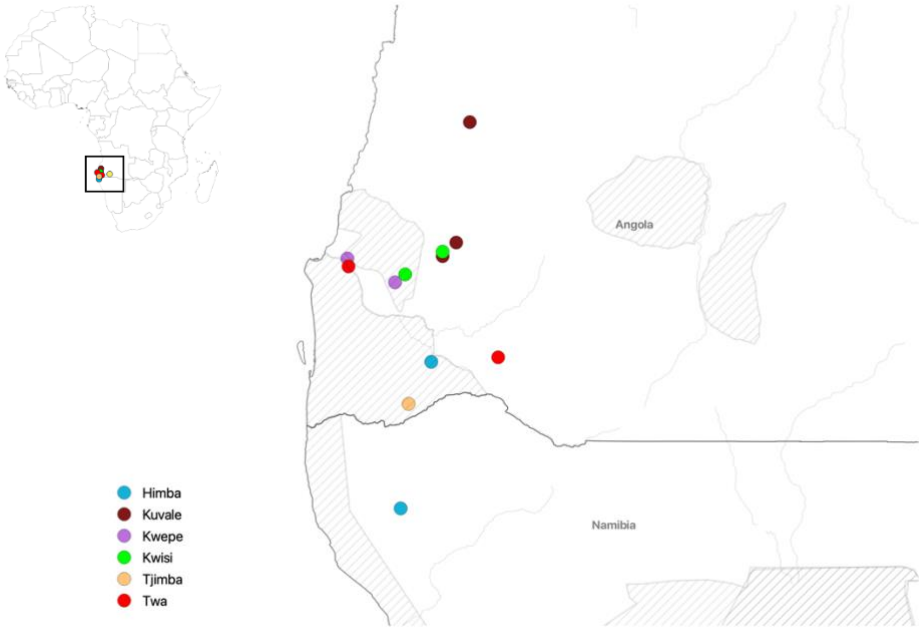

Figure 25 – Sampling locations for populations from the Namibe and Kunene provinces in Angola. Black lines indicate rivers and dashed polygons refer to established conservation areas.

Intriguingly, all peripatetics claim to be the first inhabitants of the region they dwell in. However, despite their different oral traditions and means of subsistence, our PCA analysis shows the ethnic minorities from the Angolan Namibe as closely resembling their dominant Bantu neighbours (**Figure 20**). None of them displays a clearly visible "Khoisan" contribution, with the possible exception of the Kwepe and some Kwisi and Twa individuals who are slightly pulled towards the "Khoisan" pool.

This impression is further reinforced by the admixture analysis (**Figure 26**), which confirms that all groups from the Angolan Namibe display a clear prevalence of the west Africa component associated with the Bantu migrations. However, the peripatetic populations have an elevated amount of south Africa ancestry, which may be indicative of contact. While the overall pattern remains constant in the ancestry analysis using ancient DNA proxies (**Figure 27**), some changes occur: while the Kuvale lose their negligible amount of eastern African ancestry, the southern African component becomes even more prominent in the peripatetics. This increase may indicate that the modern proxy population for the southern African forager component (Ju|'hoan North) does not approach the Khoisan component found in the peripatetics. In turn, the ancient DNA sample provides a better representative.

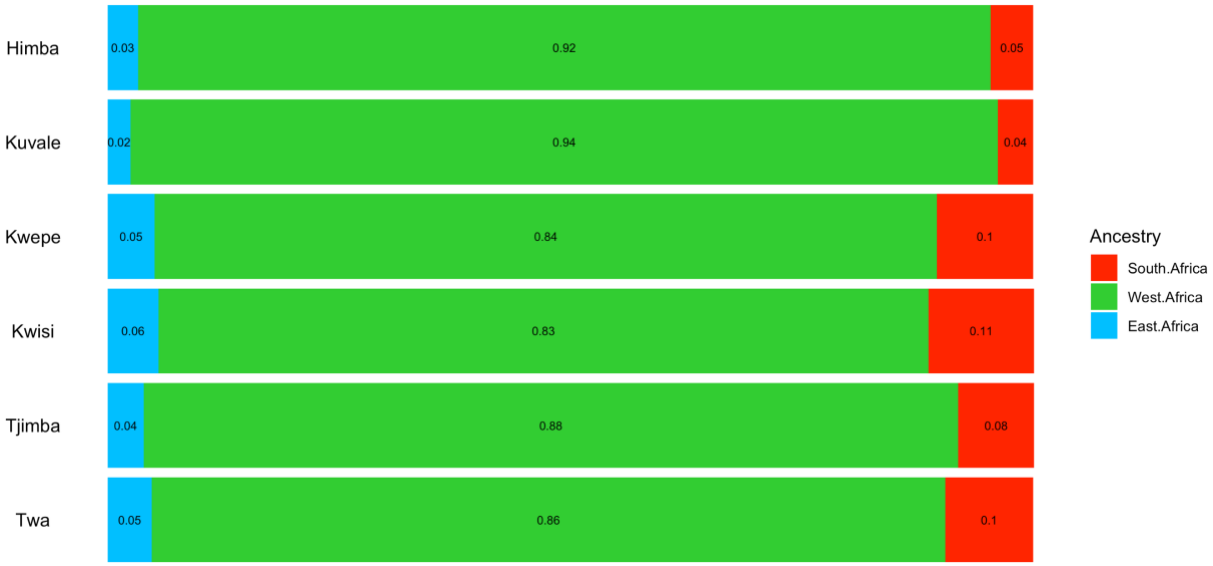

Figure 26 – Proportions of pre-identified ancestry components found in populations from the Angolan Namibe typed on the Affymetrix Human Origins Array.

Diving deeper into their Khoisan genetic heritage, we found it rather difficult to differentiate between influence from northern or Southern Khoisan on the Angolan Namibe (**Figure 28**). This difficulty may have been brought by influence from a population that had already merged both components, by differential contact with both !Xun (northern) and Nama-like (southern) populations (**Figure 25**), or by the presence of a Khoisan component equally distant from subgroups identified in our dataset.

To further pursue the differing contact influences on the populations from the Angolan Namibe, we included them into our IBD analysis. As expected, all populations share considerably more western African derived IBD segments with Bantu populations than with Non-Bantu groups[4] (**Figure 29**). Most of the Bantu-speaking populations they share IBDs with reside in the same area (*e.g.* Kwepe, Kuvale, Himba, Nyaneka, Damara[5]), and the amounts of IBD sharing are evenly distributed for all IBD lengths, suggesting ongoing gene flow within the region.

IBD sharing with Non-Bantu populations mostly targets northern Kalahari Khoe (Khwe and Shua) and !Xun. While most of the shared IBDs maintain a western African ancestry and are therefore linked to the Bantu migrations, there is still a significant contribution of east African ancestry with special concentration in the peripatetics. This pattern mirrors results based on single markers, such as eastern African NRY (Oliveira *et al.*, 2019) and LP (Pinto *et al.*, 2016). The eastern African component found in the Angolan Namibe is most likely identical with the component found in Khoe-Kwadi speaking populations, seeing it is shared with the Nama and the Hai||om. While IBD sharing with “Khoisan” population occurs, an ancestry analysis of the shared IBDs reveals their Bantu origin. A sharing of autochthonous southern African IBDs, on the other hand, is almost absent, despite detected Khoisan ancestries of ~19% and ~17% in the Kwisi and Kwepe, respectively. The lack of direct sharing between the Namibe populations and existing Khoisan populations from southern Africa may be seen as evidence for influence from a yet undescribed Khoisan component which is non-identical with the Northern (Ju) and Southern (Tuu+ǂ’Amkoe) components identified based on contemporary data from Namibia, Botswana and South Africa.

[4] We attributed the Non-Bantu category to populations that speak Khoisan, Afro-Asiatic and Nilo-Saharan languages.
[5] We included the Damara in the Bantu category since most of their genetic profile is of Bantu origin.

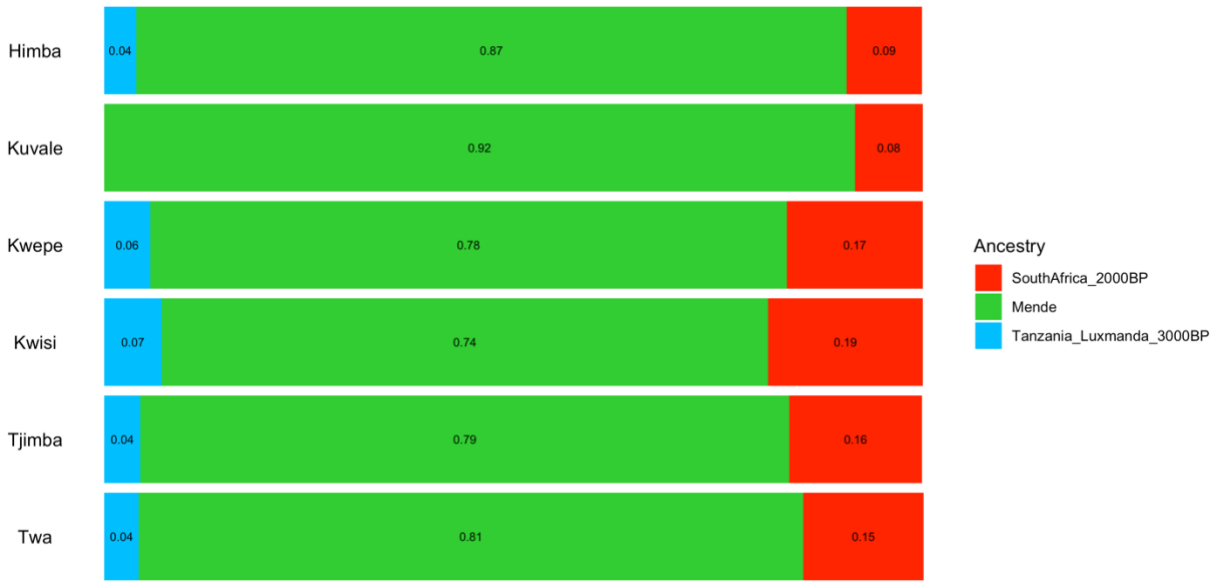

Figure 27 – Proportions of ancestry components with ancient DNA proxies found in populations from the Angolan Namibe typed on the Affymetrix Human Origins Array.

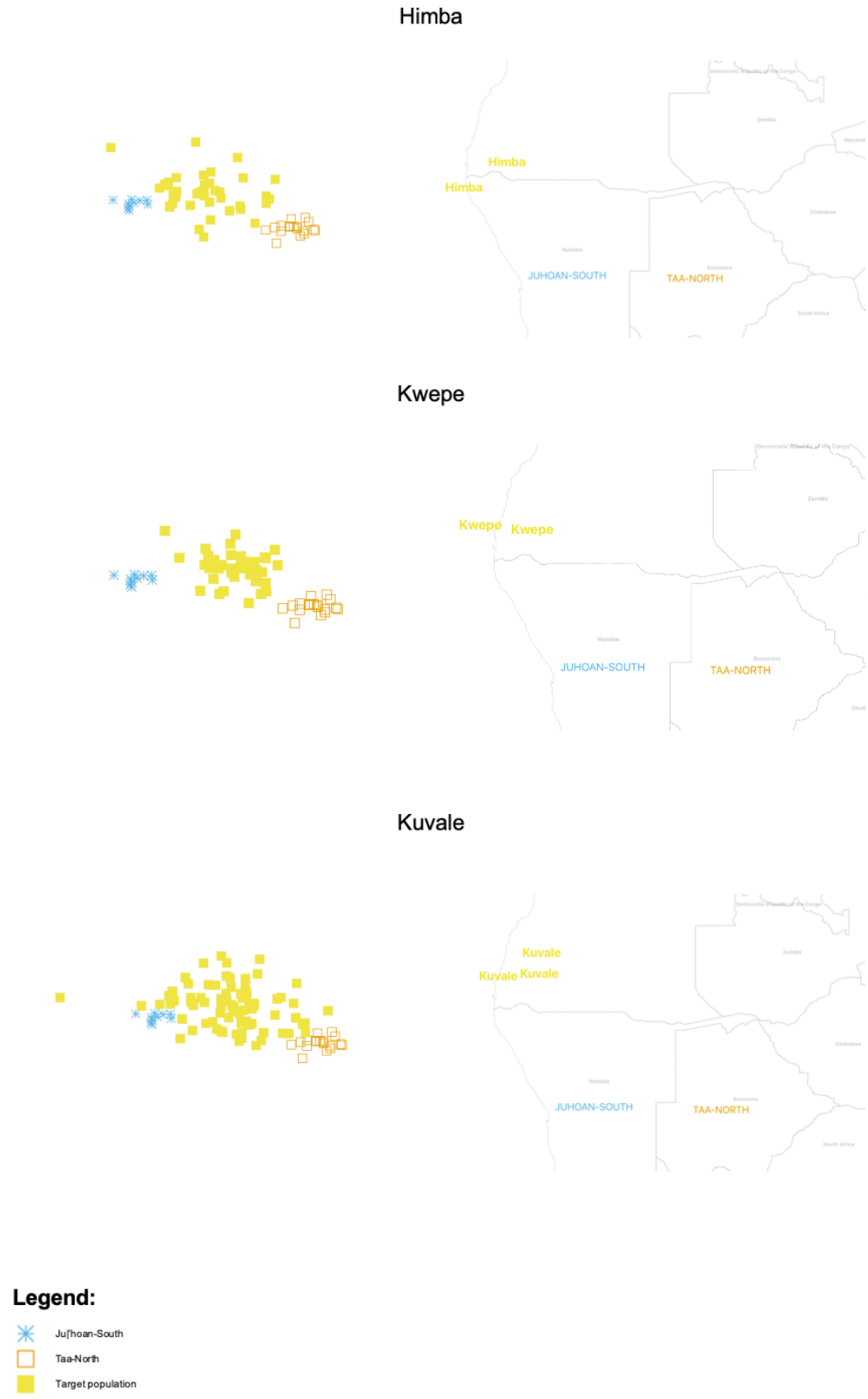

Figure 28 - Principal Component Analysis computed for the masked "Khoisan" genomes of each Namibe population (left) with accompanying maps (right).

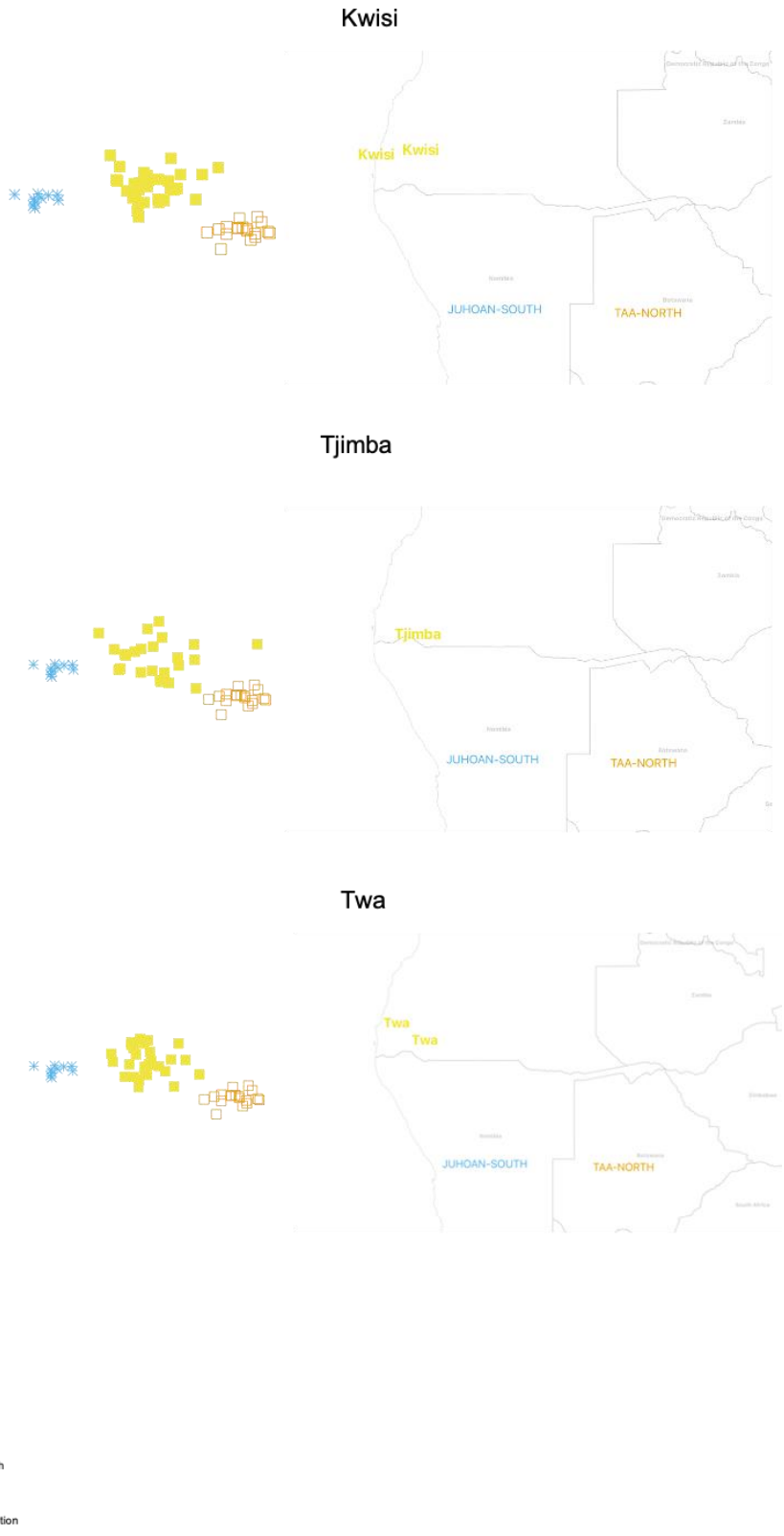

Figure 28 (cont.) - Principal Component Analysis computed for the masked "Khoisan" genomes of each Namibe population (left) with accompanying maps (right).

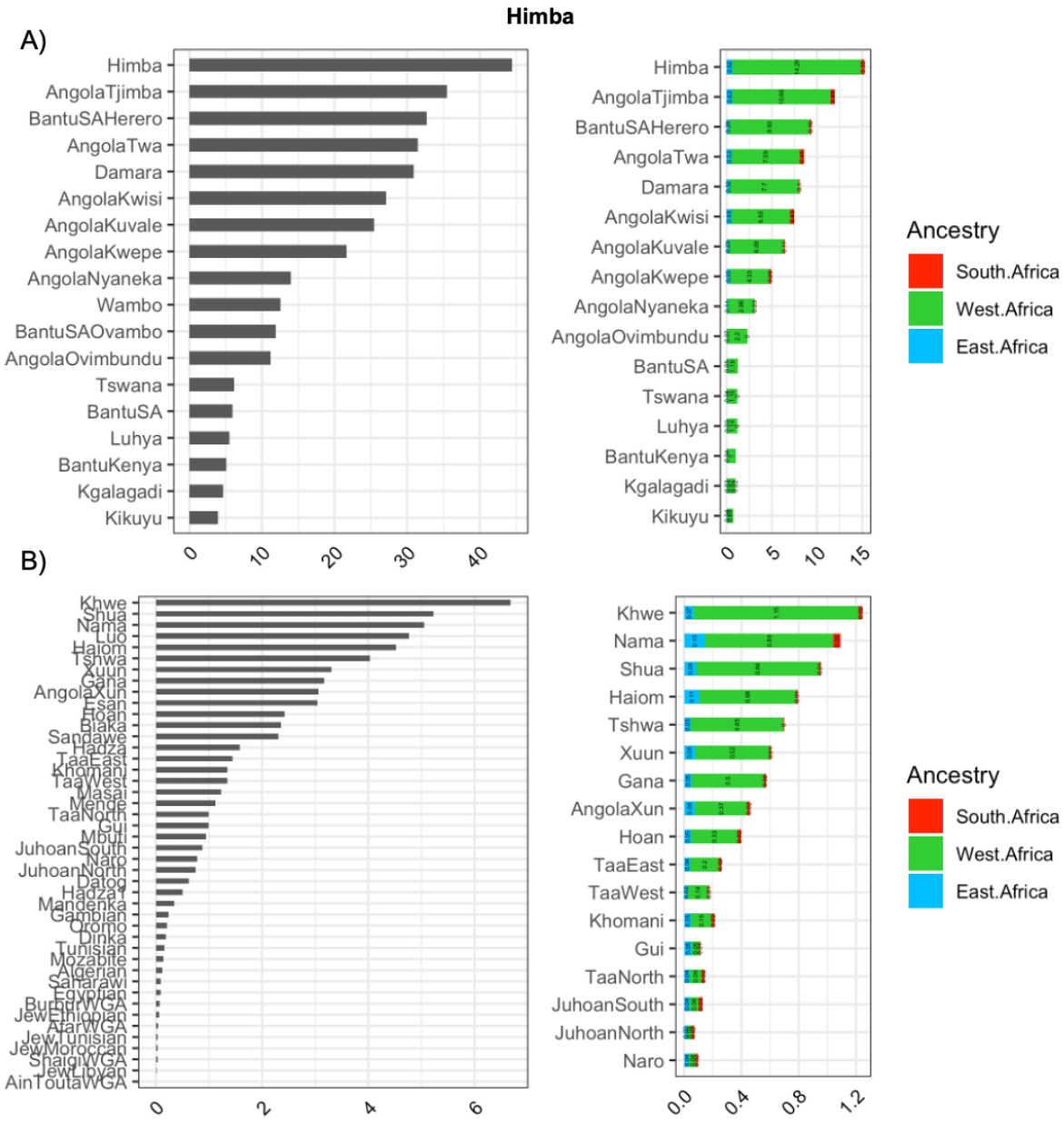

Figure 29.1 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Himba and Bantu (A) and Non-Bantu (B) populations.

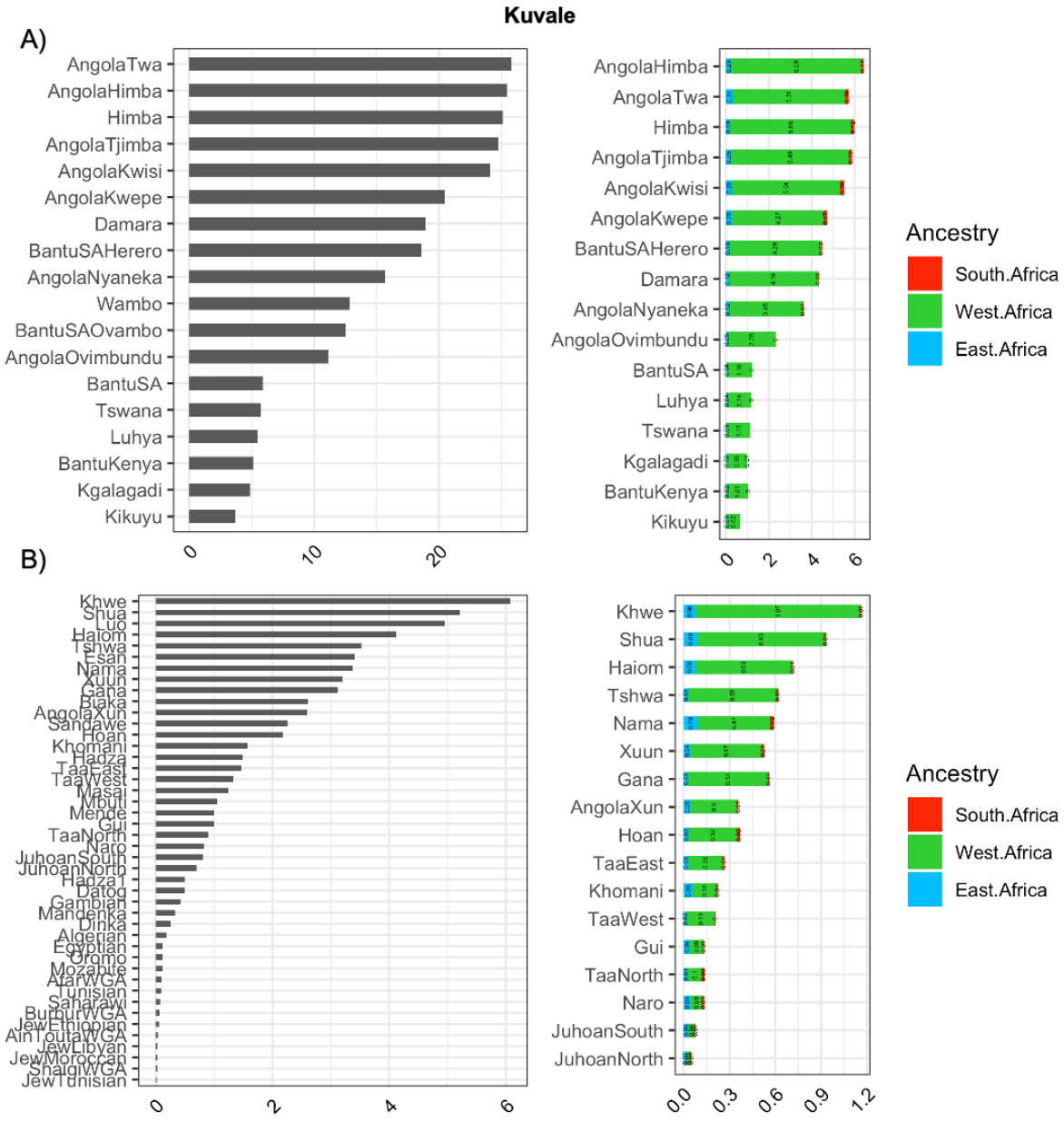

Figure 29.2 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Kuvale and Bantu (A) and Non-Bantu (B) populations.

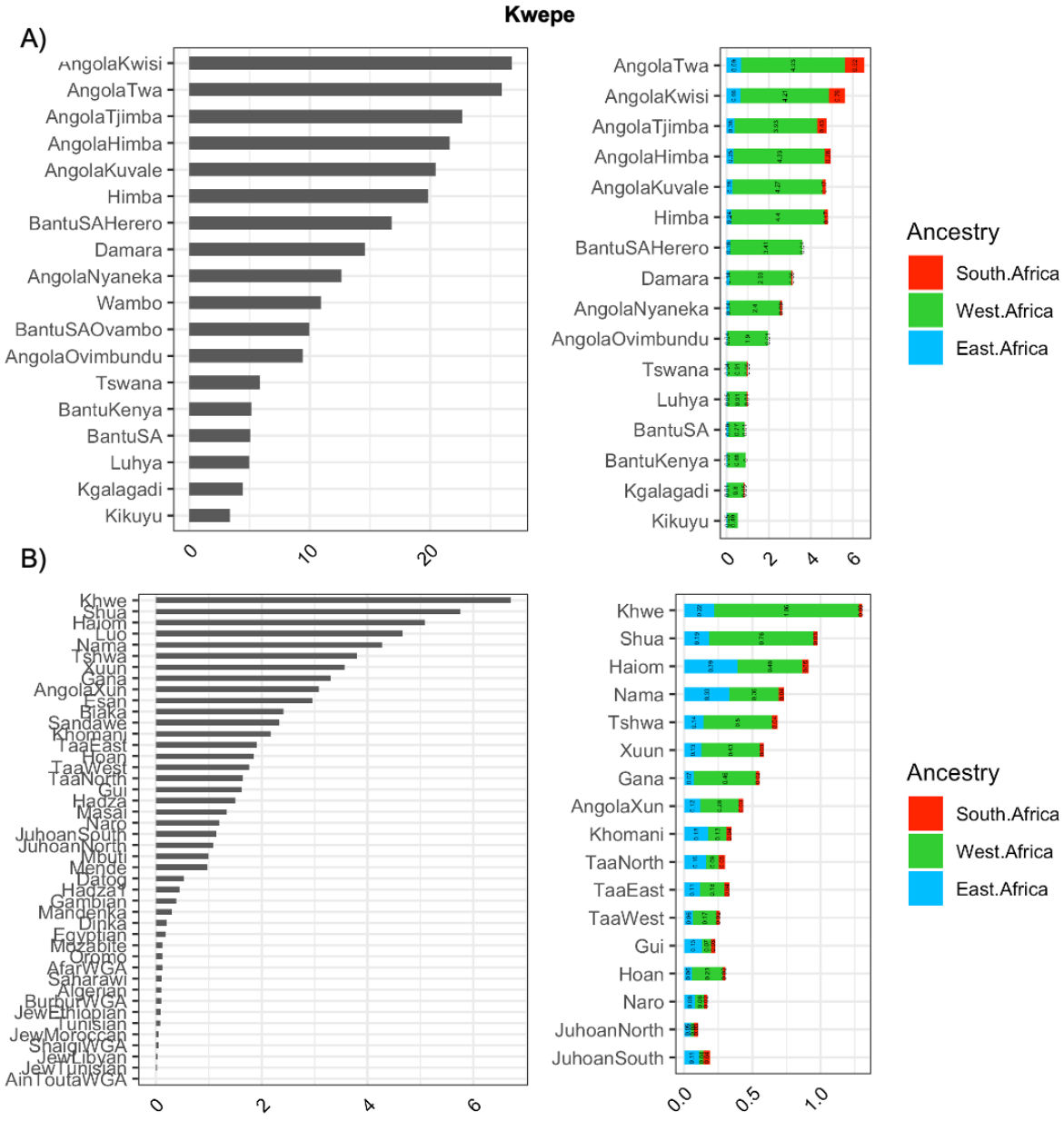

Figure 29.3 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Kwepe and Bantu (A) and Non-Bantu (B) populations.

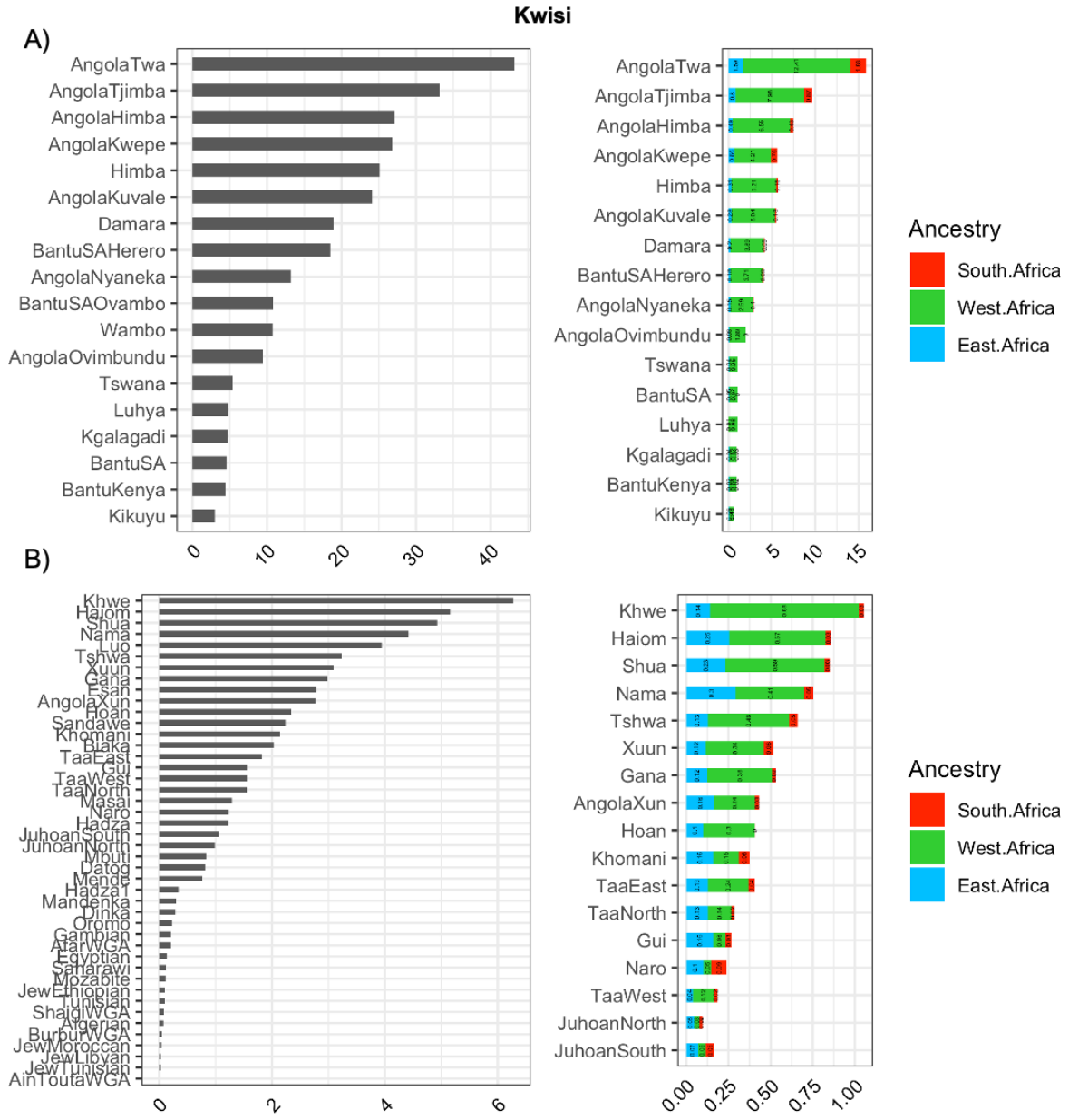

Figure 29.4 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Kwisi and Bantu (A) and Non-Bantu (B) populations.

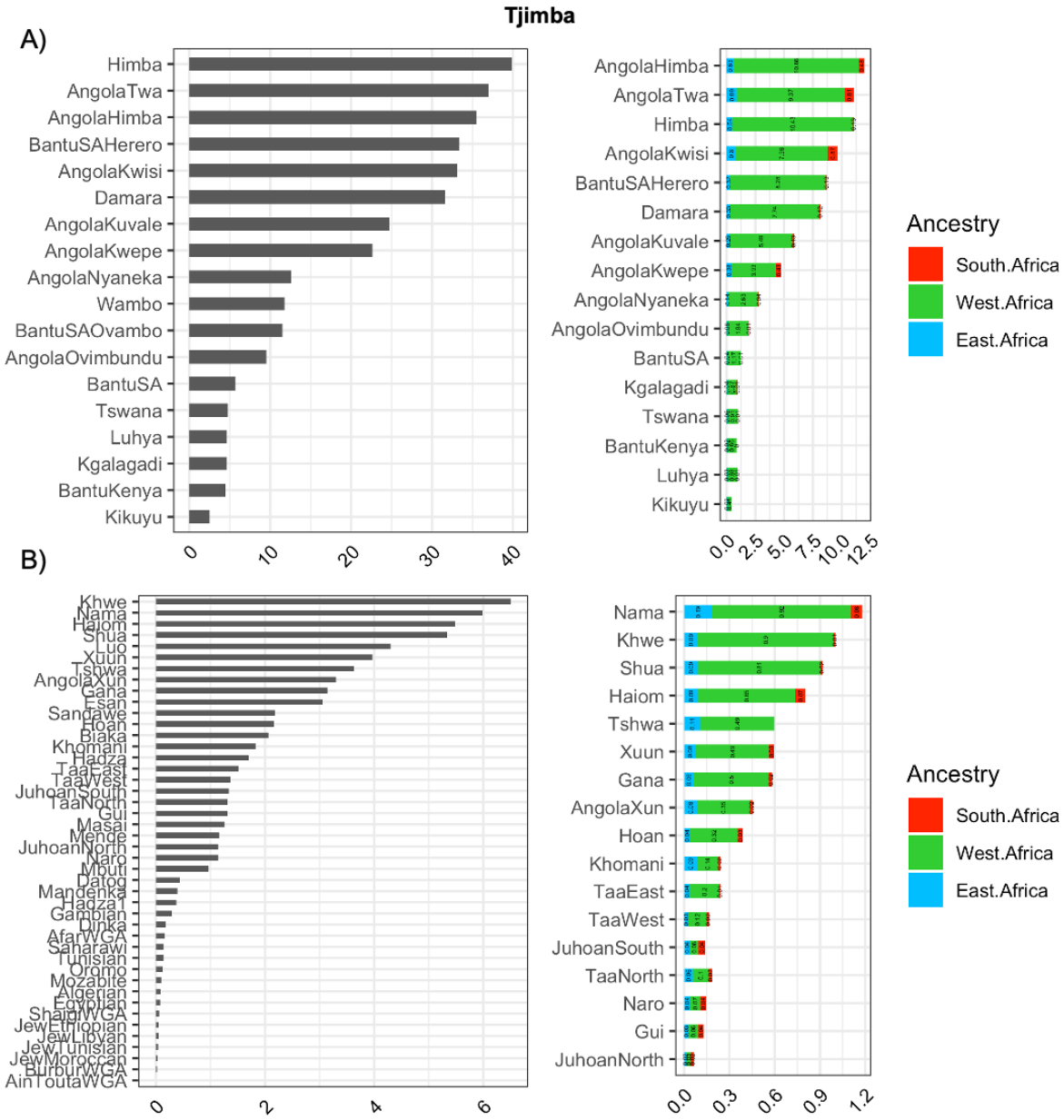

Figure 29.5 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Tjimba and Bantu (A) and Non-Bantu (B) populations.

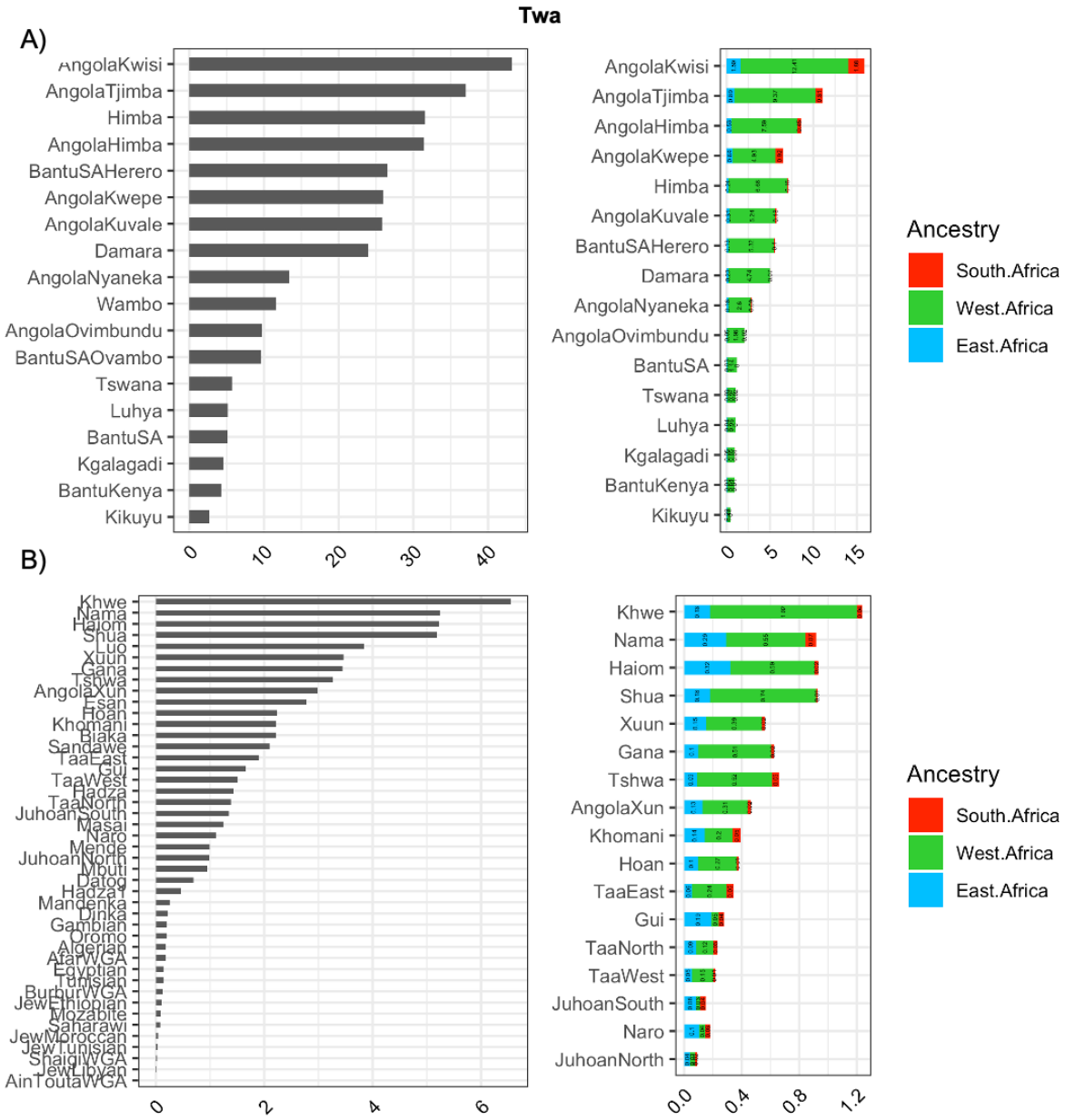

Figure 29.6 - Number of Shared IBDs (1 to 5 cM) and their respective ancestry assignment between the Twa and Bantu (A) and Non-Bantu (B) populations

[Page intentionally left blank]

Discussion

In this work, we have considered multidisciplinary data to assess the origin and historic migration of Khoe-Kwadi speakers from southern Africa. A review of existing data from archaeology, anthropology, linguistics and human genetics provides a strong case for an association between the ancestors of the Khoe-Kwadi and the introduction of Late Stone Age pastoralism from eastern Africa to the Kalahari Basin Area (see, *e.g.* Güldemann and Stoneking (2008) (§2). To test this hypothesis, we created comprehensive datasets from published and newly collected results and performed quantitative analyses of linguistic (§3.1) and genome-wide autosomal data (§3.2) in order to see whether the origin, migratory routes and branching patterns of the modern Khoe-Kwadi fit the proposed scenario.

Our results suggest that the Khoe-Kwadi are both linguistically and genetically diverse. They do not share a common genetic profile but reflect all ancestries from the Kalahari Basin area, *i.e.* autochthonous forager (Non-Khoe Khoisan) and west African food producer (Bantu). While a small genetic contribution linked to pastoral populations from eastern African could be detected in all Khoe-Kwadi speakers, its frequency varies and only reaches significant numbers in the Khoekhoe-speaking Nama. The Nama are the only Khoe-Kwadi speaking population historically associated with Late Stone Age herding and also display the highest genome-wide contribution from eastern Africa, together with elevated amounts of LP (§2.5.3), high frequencies of the light skin colour variant from SLC24A5 (§2.5.4) and of the eastern African NRY haplogroup E1b1b (§2.5.2). Therefore, it can be suggested that the genetic and cultural profile of the incoming herders was best preserved in the Nama, despite their considerable genetic contribution from Tuu-speaking foragers.

Using linguistic data as a proxy for the branching pattern and subsequent migrations of the Khoe-Kwadi within southern Africa, we performed a phylogeographic analysis which yielded a tree-like structure with three main branches previously identified with classic linguistic methods (Güldemann, 2014; Vossen, 1997): Kwadi in south-western Angola, Khoekhoe in the south-west, and Kalahari Khoe in the central Kalahari and its fringes (**Figure 17**). Contradicting previous hypotheses which located the proto-Khoe-Kwadi population in south-eastern Africa (Ehret, 2008; Elphick, 1985), our results suggest a split-off point in the west, along the border between Botswana and Namibia (Heine & Konig, 2008). A western origin of the proto-Khoe-Kwadi in southern Africa matches the earliest archaeological attestations of herding in the region, which are concentrated in northern Botswana and along the western coast of modern Namibia and South Africa (Lander & Russell, 2018; Pleurdeau *et al.*, 2012); the spread of early herder sites also seems to concur with a southward migration of the herders associated with the language's Khoekhoe branch (**Figure 17**). Such an

association is further supported by ancient DNA from a South African shepherd from the site of Kasteelberg (~1300BP), who displayed a genetic profile very similar to that of modern Nama (Skoglund *et al.*, 2017).

Focusing on newly collected data from the Namibe and Kunene provinces of south-western Angola, we found evidence for a northward migration of Khoe-Kwadi speaking pastoralists as proposed by the location of the Kwadi branch in the Khoe-Kwadi language tree. The formerly Kwadi-speaking Kwepe, their peripatetic neighbours Kwisi, Twa, and Tjimba, as well as the Khoekhoe-speaking Damara, all display predominantly Bantu genetic profiles but can be shown to have small contributions of the same eastern African component also found in Khoe-Kwadi speakers from Namibia, Botswana and South Africa. This is in line with eastern African LP and NRY previously detected in the area (Breton *et al.*, 2014; Macholdt *et al.*, 2014; Oliveira *et al.*, 2019; Pinto *et al.*, 2016). It is nevertheless difficult to determine whether genetic material from the east entered the Namibe populations due to contact with an undiluted herder population, or through interaction with a population already displaying contact influence from Bantu and/or Khoisan groups. While a Khoisan component was detected in the relevant populations, it could not be matched with either two of the main profiles found among southern African Kx'a and Tuu speakers. It therefore seems possible that the Namibe populations preserve not only an eastern African heritage but also a previously undescribed forager component related to but not identical with the southern African Khoisan.

Considering the high genetic diversity of modern Khoe-Kwadi speakers, along with the absence of obvious linguistic relatives in eastern Africa or *en route* to southern Africa, it becomes clear that their migration did not involve the linear expansion of language, genes and culture, as described for the Bantu migrations (Diamond & Bellwood, 2003; Grollemund *et al.*, 2015; Semo *et al.*, 2020). We therefore need to consider their current diversity and distribution in the light of different models which have been proposed to account for mismatches between genes, languages and culture as the result of human migration:

(1) High levels of genetic admixture between Khoe-Kwadi and their neighbouring Non-Khoe Khoisan populations indicate that intermarriage between the hunter-gatherers and the incoming herders lead to a dilution of the herders' original gene pool which became almost untraceable in later generations (Diamond & Bellwood, 2003).

(2) The fact that the majority of Khoe-Kwadi speakers are foragers who display a strong cultural link to the hunter-gatherer lifestyle may imply that they gave up their pastoral lifestyle shortly after having arrived in southern Africa. Such a reversion may have been triggered by an expansion into areas like the Okavango River Basin (*e.g.* Khwe) or the central Kalahari Desert (*e.g.* Glui-Gǁana) which are unsuitable for herding domestic animals (Diamond & Bellwood, 2003).

(3) The scenario was probably further complicated by autochthonous populations shifting to Khoe-Kwadi in a contact situation involving dominance by the herders. Interestingly, the Khoe-Kwadi may not only have triggered language shift in formerly Kx'a and Tuu-speaking populations but also in some of their Bantu neighbours, as exemplified by the Damara and Kwadi. As these shift situations probably involved a clear imbalance of power, the incoming herders would not have shared genes with their contact populations (Diamond & Bellwood, 2003), leading to Khoe-Kwadi-speaking groups with a genetic heritage which is almost exclusively Khoisan (*e.g.* Naro, Glui) or Bantu (Damara, Kwepe).

(4) Finally, it seems likely that the ancestors of the Khoe-Kwadi inhabited areas they are no longer found in. These may be located in modern-day eastern Africa, but also in areas like Malawi, Zambia and Zimbabwe which connect eastern to southern Africa and are nowadays almost exclusively populated by Bantu speakers. If the Khoe-Kwadi had settled or passed there, it seems likely that their languages and genes were subsequently replaced by the Bantu migrations, making them inaccessible to modern analysts. Furthermore, no related language can be located in modern-day eastern Africa. While Sandawe shares similarities with Khoe-Kwadi, these may be due to shared contact with an unknown third party or typological resemblance within a former contact setting. The linguistic and genetic heritage of the Khoe-Kwadi is therefore restricted to southern Africa, rendering it difficult to trace their full migratory movement as can be done for the Bantu or Indo-European expansions.

Some authors (Bousman, 1998; Diamond & Bellwood, 2003; Sadr, 2013) have proposed a further scenario in which herding was diffused to local hunter-gatherers with no or only little genetic contribution from the original pastoralists. Following this proposal, the Khoekhoe would be former !Ui-speakers who received both language and livestock from a source no longer physically present in the area. However, it has been pointed out (Smith, 2017) that foragers are reluctant to adopt food production, especially herding. Furthermore, the presence of eastern African genetic markers, including an autosomal component, can be seen as proof for a demic scenario which included the migration of the actual herding population. Hence, the Khoekhoe are best seen as a pastoral population from eastern Africa with local southern African contributions, rather than southern African foragers who adopted herding.

Taken together, our results suggest that the origin of the Khoe-Kwadi lies indeed in eastern Africa: the typological and even lexical affinity of the Khoe-Kwadi phylum to Non-Bantu languages presently spoken in eastern Africa has a counterpart in genetic markers displaying a link to pastoralists from the region (Pickrell *et al.*, 2014). Considering the archaeological evidence, these pastoralists entered southern Africa not long before the arrival of the first Bantu speakers around 2000BP with fat-tailed sheep and possibly some few heads of cattle. After having reached the western Kalahari Basin fringe in the border area between modern

Namibia and Botswana, the proto-population diverged into three main branches which continued to interact with Kx'a and Tuu-speaking groups, as well as with incoming Bantu farmers from the eastern and western branches of the Bantu migrations. While the full package (pastoralism, eastern African ancestry and Khoe-Kwadi linguistic patrimony) was largely retained in the Khoekhoe, interaction in a contact zone led to novel combinations of culture, genes and languages: Khoe-Kwadi speakers may be foragers with a genetic patrimony indistinguishable from their Kx'a and Tuu neighbours (Naro, Glui), peripatetics sharing a genetic ancestor with south-western Bantu speakers (Damara, Kwepe), or highly admixed populations carrying the cultural and genetic patrimony of all population strata found in modern southern Africa (Khwe, Shua, Tshwa). They remain united by their linguistic heritage and remind us of the complex contact scenarios which are often at odds with the widespread idea of pure, undisturbed forager populations without history.

Material and Methods

Sample collection

This study makes use of previously unpublished data from human evolutionary genetics and linguistics collected in the field. Saliva samples from nine diverse populations from south-western Angola speaking Bantu, Khoe-Kwadi and Kx'a languages were collected during three ethnographically informed field trips from 2013-2014 (Oliveira, 2019). During the same trips, linguistic data from two rememberers of Kwadi was collected (Fehn, 2019b; Oliveira, 2019; Oliveira *et al.*, 2018; Oliveira *et al.*, 2019).

Additional linguistic data from selected Khoe-Kwadi varieties used in the phylogenetic study were collected among the Ts'ixa and Shua communities (Botswana), the Khwe community (Namibia & Botswana), and the Tjwao community (Zimbabwe).

Genotyping

DNA was extracted from 208 saliva samples from individuals belonging to the nine populations from south-western Angola. These individuals were genotyped on the Affymetrix Human Origins Array. SNPs with a missing rate above 10%, and with deviations from Hardy-Weinberg equilibrium and non-autosomal markers were removed. Furthermore, 37 individuals were excluded from the dataset due to cryptic relatedness. More details about DNA extraction and genotyping can be found in Oliveira (2019).

Merging with other datasets

For comparative purposes, the genetic data was supplemented with 550 individuals from 56 additional populations genotyped on the same array published in Lazaridis *et al.* (2014), Pickrell *et al.* (2012) and Patterson *et al.* (2012). The same quality control filters were applied as described above, and 25 individuals were excluded, due to cryptid relatedness. After applying the quality filters, the final merged dataset yielded a total of 607,761 SNPs genotyped for 378 individuals (Oliveira, 2019). Details about these populations are provided in **Table S 4**.

Genetic analyses

Towards exploring population structure and differentiation, we performed a PCA with the Affymetrix Human Origins Array SNP data, using the smartpca software of the EIGENSOFT 6.0.1 package (Patterson *et al.*, 2006) (**Figure 20**). This analysis allows us to see the genetic relationships between our populations, and, therefore, see which populations are genetically closer to each other.

To explore interactions within and between 63 populations typed on the Affymetrix Human Origins Array, we evaluated Identity by Descent (IBD) fragment sharing between them. IBD sharing is defined as two or more individuals sharing identical nucleotide sequences (identical by state - IBS) that they have inherited from a common ancestor without recombination (Stoneking, 2016). Before doing the IBD analysis, the data was phased by running the program BEAGLE (Browning & Browning, 2007).

We used refinedIBD v.17Jan20.102 (Browning & Browning, 2013) to identify shared IBD and homozygous-by-descent (HBD) blocks within each individual. We used the program merge-ibd-segments v.17Jan20.102 from BEAGLE (Browning & Browning, 2013) to merge IBD and HBD blocks within a 0.6 centiMorgan (cM) gap, allowing only one inconsistent genotype between the gap and block regions. The reference genome used was the Genome Reference Consortium Human Build 37 (GRCh37) in PLINK format (Church *et al.*, 2011; Purcell *et al.*, 2007). As described in Liu *et al.* (2019), we created three datasets of the merged IBD segments according to the length of IBD blocks: 1 to 5 cM, 5 to 10 cM and over 10 cM. The population means/medians were calculated in order to explore IBD sharing between populations across generations. The different lengths of IBD blocks provide a time frame in which the smallest IBD blocks (1 to 5cM) represent the oldest contact between two populations, and the longest segments (over 10cM) indicate the most recent contact. More specifically, IBD blocks between 1 to 5cM can be attributed to 90 generations ago (~2700 years), segments 5 to 10cM long are from 23 generations ago (~675 years) and segments over 10cM are only 7.5 generations dated (~225 years) (Al-Asadi *et al.*, 2019). These results were then visualized as histograms using the R package *ggplot2* (Wickham, 2016).

For the purpose of this study, the ancestry proportions of 37 populations genotyped on the Affymetrix Human Origins Array were estimated with RFMix version 2 (Maples *et al.*, 2013). RFMix2 assigns each SNP to an ancestry, based on reference populations which we considered as possible sources. We chose the Yoruba, Ju|'hoan North and Somali as the sources of the following ancestries: west Africa ("Bantu"), south Africa ("Khoisan") and east Africa, respectively. We selected 13 individuals per source population to perform the ancestry assignment and then calculated the proportion of SNPs assigned to each ancestry per individual and population, with a correction for the number of SNPs per chromosome. The resulting ancestry proportions can be viewed in **Table S 5**.

In order to assess which proportion of each IBD shared segment between two populations belonged to which ancestry, we ran RFMix2 with 25 number of generations and the option –reanalyze-reference with 3 iterations, to consider sets of haplotypes that are admixed and not of "pure" ancestry. Our criteria for assigning an ancestry portion to an IBD segment depended on the mean of each ancestry for both haplotypes: *e.g.* if 51% of the length of the IBD segment was of "south African" ancestry in the first haplotype and 53% of the length of the IBD segment

was of “south African” in the second haplotype, then we considered that particular IBD segment to have 52% of “south African” ancestry. Some IBD segments were excluded if the ancestry calls of the pairs of haplotypes were inconsistent.

In order to see whether the Khoisan component in our populations was more closely related to either Kx’a or Tuu populations, we used the ancestry-assigned dataset from the first run of RFmix2 to perform a “masking” analysis, where we converted the alleles that belonged to Non-Khoisan ancestries (west Africa and east Africa) to missing data. We did this for all populations except for the Ju|’hoan North and Taa west, which were used to calculate the PCs and represent the Kx’a and Tuu genetic components. We then projected the masked genomes of individual Khoe-Kwadi and Namibe populations onto eigenvectors (**Figure 23**, **Figure 28**) computed using the LSQ option of smartpca from the EIGENSOFT 6.0.1 package (Patterson *et al.*, 2006), which indicates which populations to be projected.

We further tested 1-source, 2-source or 3-source models as well as admixture proportions for the Affymetrix Human Origins Array populations, using the qpAdm algorithm in the ADMIXTOOLS package (Patterson *et al.*, 2012) and following the same analysis done in Skoglund *et al.* (2017). Firstly, 19 populations (Mbuti, Dinka, Mende, South_Africa_2000BP, Tanzania_Luxmanda_3100BP, Ethiopia_4500BP, Levant_Neolithic (PPNB), Anatolia_Neolithic, Iran_Neolithic, Denisova, Loschbour, Ust_Ishim, Georgian, Iranian, Greek, Punjabi, Orcadian, Ami, and Mixe) were used to maximize the power to infer the admixture proportions for the ancient African populations. This set is composed of previously published complete genomes (Fu *et al.*, 2014; Lazaridis *et al.*, 2014; Mallick *et al.*, 2016; Meyer *et al.*, 2012) as well as enriched ancient DNA data (Lazaridis *et al.*, 2016; Mathieson *et al.*, 2015), and was chosen to capture the main threads of ancestry in the diverse populations found in sub-Saharan Africa (Skoglund *et al.*, 2017). Next, we selected Mende, South_Africa_2000BP and Tanzania_Luxmanda_3100BP from the outgroup set to test as sources in admixture models, noting that the last two belong to ancient DNA. These selected populations are not necessarily the true source populations, but they are more closely related to the true source than the remaining populations from the outgroup (Skoglund *et al.*, 2017). For each target population, we tested 3 one-source ancestry models, 3 two-source models and 1 three-source model, thus obtaining mixture models for the populations genotyped on the Affymetrix Human Origins Array (**Table S 6**).

In order to observe the genetic variation in the diverse populations of southern Africa, we collected the frequency distribution of mtDNA (Barbieri, Güldemann*, et al.*, 2014; Barbieri, Vicente*, et al.*, 2014; Barbieri *et al.*, 2013; Castrì *et al.*, 2009; de Filippo *et al.*, 2010; Marks *et al.*, 2015; Oliveira *et al.*, 2018; Schlebusch, 2010; Tishkoff *et al.*, 2007) and Y-chromosome (Bajić *et al.*, 2018; Barbieri *et al.*, 2013; Marks *et al.*, 2015; Mineiro, 2016; Oliveira *et al.*, 2019;

Schlebusch, 2010) haplogroups as well as lactase persistence alleles (in particular, variant -14010C) (Breton *et al.*, 2014; Macholdt *et al.*, 2014; Pinto *et al.*, 2016; Ranciaro *et al.*, 2014; Schlebusch *et al.*, 2012) and skin colour allele (SLC24A5) (Lin *et al.*, 2018) from published data. With this data, we created frequency plots for each population using ggplot2 (Wickham, 2016) in R Studio v1.2.5033 (RStudio Team, 2019) (**Figure 8, Figure 9B, Figure 10B, Figure 12, Figure 13B**).

Linguistic analyses

All maps presented in this thesis were created with QGIS v3.10 (QGIS.org, 2020). The base map was built using the publicly available Map of National Boundaries for Africa (https://hub.arcgis.com/datasets/geoduck::africa-boundaries) as well as detailed vectors of African Rivers (http://landscapeportal.org/layers/geonode:africa_rivers_1#more), Waterbodies (http://geoportal.rcmrd.org/layers/servir%3Aafrica_water_bodies#license-more-above) and Worldwide Protected Areas (https://hub.arcgis.com/datasets/geoduck::africa-boundaries).

For our phylogenetic analysis, we created a linguistic dataset comprising 35 Khoe-Kwadi languages, based on lexical data matching two well-known wordlists: (1) Leipzig-Jakarta (LJ), a 100-word list that was systematically created to compile meanings displaying little borrowing across languages (Tadmor *et al.*, 2010) (2) Swadesh 100 (SW), created with the same intention, but based on intuition (Tadmor *et al.*, 2010). Finally, we created a dataset consisting of the merged meanings from both wordlists. In order to gain a maximally wide coverage of the area in which Khoe-Kwadi was spoken in historical times, this dataset includes historical data from Khoekhoe varieties spoken in the western and eastern Cape (both Nienaber (1963)). The sources for all individual varieties are displayed in **Table S 1**.

In order to analyse the linguistic data of the Khoe-Kwadi languages, we coded languages for presence (1) or absence (0) of a lexical root or cognate set. Missing data were coded as <?>, following Fehn (2019b). In linguistics, cognates are understood to be words that have a common etymological origin, and that could be either inherited from a common ancestor language or borrowed from another language. If a lexical root expressing the same meaning is shared by two languages, they are coded as 1, and if it is absent, as 0. An example of coding is seen in the following **Table 6**.

The dataset comparing the linguistic varieties of Khoe-Kwadi was first converted into the NEXUS format using a custom-built python script. Before the analysis, we observed the state counts for each language in order to check the proportion of ‘0’s, ‘1’s and missing data to make sure languages with low data would not disrupt the result. We also looked at the distribution of cognate sizes, whereas it is considered ideal to have a big number of cognates shared by one or two sister languages, and fewer cognates shared by many languages (**Table**

7). These preparatory analyses were all made using custom scripts in R Studio v1.2.5033 (RStudio Team, 2019).

Table 6 - Examples of cognate coding.

Meaning	Root	Kwadi	Damara	Tshwao
Black	*nǂuu	1	1	1
House	*kx'óm̀	0	1	0

Table 7 - Characteristics of the linguistic dataset of the 35 Khoe-Kwadi varieties.

Cognates (Characters)	Meanings	Roots	Doculects
527	136	480	35

The linguistic datasets were subsequently used in SplitsTree v4.16.1 (Huson & Bryant, 2006). A Neighbor-Joining (NJ) tree was computed, as well as the delta (Holland *et al.*, 2002) and Q-residual scores to quantify the tree-likeness of the dataset (Gray *et al.*, 2010).

The delta score method scores individual taxa from 0 to 1, according to how much each taxon is involved in conflicting signal, which appears as a network-type structure, instead of a tree-like structure (Gray *et al.*, 2010; Holland *et al.*, 2002). If the delta score equals zero, the distances between the taxa exactly fit a tree model. Otherwise, the score ranges from 0 to 1 (Gray *et al.*, 2010). These scores were computed for each doculect (**Table S 2**), as well as for the Khoe-Kwadi family and the Khoekhoe and Kalahari Khoe subgroups (**Table 4**). Additionally, we computed distance-based trees like UPGMA and Neighbor-Joining, which both use clustering algorithms (Penny & Hendy, 2004), as well as parsimony-based Maximum Parsimony, which can distinguish ancestral and derived shared traits (Page & Holmes, 2009; Penny & Hendy, 2004). These trees were built in R Studio v1.2.5033 (RStudio Team, 2019) using both ape v5.3 (E. Paradis, 2018) and paragorn v2.5.5 (Schliep *et al.*, 2016; Schliep, 2011) packages. For each tree, a likelihood score test was performed using the function 'pml' in combination with a corresponding statistical analysis; subsequently, the likelihood scores were used to optimize each tree (**Figure S 3**, **Figure S 4**, **Figure S 5**) (Schliep *et al.*, 2016; Schliep, 2011). Finally, a bootstrap analysis was performed with the function 'bootstrap.pml', using 1000 bootstrap replicates (Schliep *et al.*, 2016; Schliep, 2011).

To supplement the distance-based methods, a Bayesian phylogenetic approach was implemented, using BEAST v2.6.3 (Bouckaert *et al.*, 2019) with the Markov chain Monte Carlo (MCMC) algorithm (Drummond *et al.*, 2002). This approach is character-based, and its aim is

modelling the best-fitting evolutionary scenario considering loss and gain of cognates (Dunn, 2015).

In order to find the best-suited model for our linguistic dataset, we used the Babel package ("Babel: BEAST analysis backing effective linguistics,") to test three models of cognate evolution: (1) the Continuous Time Markov Chain (CTMC) model providing the most simple scenario in which cognate sets may be gained and lost at the same time (Bouckaert, 2015; Bouckaert *et al.*, 2012; Gray *et al.*, 2009); (2) the Covarion model, which allows the cognates to be in “slow” and “fast” states of change, being better suited for scenarios of language diversification as it involves changes in population size and contact (Atkinson, 2011; Bouckaert, 2015; Penny *et al.*, 2001); (3) the stochastic Dollo model, which assumes that each cognate can rise only once but be lost multiple times (Bouckaert, 2015; Bouckaert *et al.*, 2012; Nicholls & Gray, 2006). We also considered two forms of rate variation of branches in the tree and across cognate sets: (1) the strict model, that assumes no rate variation and uses only the clock rate as a scale factor for all branches in a tree (Bouckaert, 2015; Bouckaert *et al.*, 2012); (2) the uncorrelated relaxed clock model, which allows variation of clock rates across branches (Bouckaert, 2015; Bouckaert *et al.*, 2012). This resulted in a total of six files.

In the prior section, we considered Yule (pure birth) as the tree prior, commonly used with model species diversification, which assumes that all lineages have been sampled at the same time (Bouckaert, 2015; Bouckaert *et al.*, 2012). We also adopted a constant birth rate through time, meaning a uniform (0,1) prior to birth and death events (Bouckaert, 2015; Bouckaert *et al.*, 2012). We also added a TMRCA prior for a second analysis in order to put Kwadi as an outgroup, thereby grouping the Kalahari-Khoe and the Khoekhoe as a monophyletic clade.

We ran each model assuming a chain length of 10000000 generations, sampling every 1000 generations. Afterwards, each .log file was assessed in Tracer v1.7.1 (Rambaut *et al.*, 2018) to visualize the resulting summary statistics as well as to evaluate ESS values and traces.

Subsequently, we evaluated the performance of each model in two different ways: (1) path sampling (PS), one of the most recent approaches to model selection in phylogenetics which outperforms previous methods like the harmonic mean estimator (HME) and improves the accuracy of model selection (Baele, Li*, et al.*, 2012; Leaché *et al.*, 2014); and (2) Akaike Information Criteria through Markov Chain Monte Carlo (AICM), based on posterior simulation, where lower AICM values indicate a better model fit (Baele, Lemey*, et al.*, 2012); AICM performs better than HME but is considered unreliable.

The analyses were computed with the MODEL_Selection v1.0.2 (Bouckaert, 2014) package from the Beauti v2.6.3 (Bouckaert *et al.*, 2019) package library using the path sampler app. The .xml file for each corresponding model was analysed using an alpha of 0.3, with 20 steps to provide good ESS values, and burn-in as 10% of the total chain length (1000000).

The corresponding marginal likelihood (logML) estimate was retrieved from each run and used to compare between the best (with the lowest logML) and worst models (with the highest logML) by using the Bayes Factor (BF) estimator calculated as logBF=logML1-logML2 (Bouckaert, 2014; Kolipakam *et al.*, 2018; Leaché *et al.*, 2014). The logBF compares two individual models and indicates the better one, depending on its signal and value: if logBF is positive, then logML1 is favoured; if it is negative, logML2 will be the better model (Kass & Raftery, 1995). The support for this statistic varies according to the value of the BF (which we standardized as 2*|logBF| (Kolipakam *et al.*, 2018)), as described in Kass and Raftery (1995). The AICM scores were computed using the app Analyser Akaike Information Criterion by MCMC, where we ran the .log file from each run specifying burn-in as 10% and bootstraps as 10000. AICM and logML values for each model are shown in **Table S 3**. The corresponding statistical analysis can be found in **Table S 7**.

For each model, each tree posterior was visualized in DensiTree v2.2.7 (Bouckaert & Heled, 2014) (**Figure S 7**). A maximum clade credibility tree was computed with TreeAnnotator v2.2.7 (Drummond & Rambaut, 2007) by keeping burn-in as 10% as assessed using Tracer v1.7.1 (Rambaut *et al.*, 2018) and, finally, visualized using FigTree v1.4.4 (Rambaut, 2016-2018) (**Figure S 6**, **Figure 16**).

To infer if a correlation is mainly driven by the geographic distance between the three main branches of Khoe-Kwadi (Kwadi, Kalahari-Khoe and Khoekhoe), a Mantel test was performed with the R package *vegan* (Jari Oksanen, 2019), considering both Pearson (which evaluates the linear relationship between two continuous variables, *i.e.* whether a linear equation describes the relationship between language and geographic distance, and if that relationship is positively or negatively correlated) and Spearman correlations (which evaluates the monotonic relationship between two variables, *i.e.* whether language distance increases or decreases when geographic distance increases), with 1e+06 permutations (Haynie, 2014; Legendre & Legendre, 2012) (**Table 5**). In theory, both coefficients take values between -1 and 1, where r=1 indicates perfect positive correlation (Haynie, 2014; Legendre & Legendre, 2012).

In the final step, we performed phylogeographic hypothesis testing. We experimented with different language evolution and spatial diffusion models (Lemey *et al.*, 2009) using BEAST 2.6.3 (Bouckaert *et al.*, 2019) in combination with the GEO_SPHERE package (Bouckaert, 2015, 2016). This package allows the addition of geographic information to our cognate data to produce a posterior distribution of Khoe-Kwadi language trees with location estimates at the root and internal nodes sampled in proportion to their posterior probability (Bouckaert, 2016). This analysis allows us to locate the internal nodes of the phylogenetic tree in

geographic space. The inferred posterior probability location represents the possible range across which the ancestral language of the Khoe-Kwadi was spoken (Bouckaert, 2016).

In the course of the spatial analysis, we opted to experiment with two models of cognate evolution: (1) the Continuous Time Markov Chain (CTMC) model (Bouckaert, 2015; Bouckaert *et al.*, 2012; Gray *et al.*, 2009) and (2) the Covarion model (Atkinson, 2011; Bouckaert, 2015; Penny *et al.*, 2001), accounting for both (1) strict and (2) relaxed log normal clock rates as rate variation (Bouckaert, 2015; Bouckaert *et al.*, 2012). This resulted in four files in total. We adopted a constant birth rate through time (0,1) and considered Yule (pure birth) as the tree prior (Bouckaert, 2015; Bouckaert *et al.*, 2012). Each model was run assuming a chain length of 10000000 generations, sampling every 1000 generations. AICM and logML values for each model are shown in **Table S 3**.

To evaluate the support for each model, a Nested Sampling (NS) analysis was computed using the NS v1.1.0 package (Bouckaert, 2019b). Nested sampling is faster than path sampling and provides an estimate of the marginal likelihood along with its variance, which allows us to see the confidence in the corresponding BFs (Baele, Li*, et al.*, 2012; Kass & Raftery, 1995; Leaché *et al.*, 2014). To perform the analysis, on the app MCMC to XML, we chose Particle Count, Sub Chain Length and Epsilon as followed: 10; 10000; 1.0E-6 (Bouckaert, 2019a, 2019b). However, in order to provide an accurate estimation of these values, we observed the Standard Deviation (SD) of the marginal likelihoods and made sure that SD was below or equal to 2. In consequence, Particle Count was increased to 20 in all files in order to decrease SD and provide a fair analysis with accompanying good ESS values (Bouckaert, 2019a, 2019b). The resulting XML file was run on BEAST v2.6.3 (Bouckaert *et al.*, 2019) and SD and ESS values were inspected. The resulting .log files were input in NSLogAnalyser with N=20 and the marginal L estimate was retrieved for each model. Afterwards, a Bayes Factor analysis was conducted, as it was previously described (Bouckaert, 2014; Kolipakam *et al.*, 2018; Leaché *et al.*, 2014). The corresponding statistical analysis can be found in **Table S 7**.

We then visualized the tree posterior in DensiTree v2.2.7 (Bouckaert & Heled, 2014) to observe overall tree topology, and then proceeded to create a maximum clade credibility tree (MCCT) using TreeAnotator v2.2.7 (Drummond & Rambaut, 2007). For both tree burn-in and posterior probability limit, we kept 0 as a value in order to annotate all trees and chose mean heights for the node heights. The resulting MCCT was used to produce a continuous tree in SPREAD v1.0.7 (Bielejec *et al.*, 2011), where a KML file was generated and then visualized in a map built on QGIS v3.10 (QGIS.org, 2020) (**Figure 17**).

References

Abu-Amero, K. K., Larruga, J. M., Cabrera, V. M., & González, A. M. (2008). Mitochondrial DNA structure in the Arabian Peninsula. *BMC evolutionary biology, 8*(1), 1-15.

Al-Asadi, H., Petkova, D., Stephens, M., & Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS genetics, 15*(1), e1007908.

Almeida, A. d. (nd). *Kwadi fieldnotes and recordings*. Unpublished fieldnotes.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science, 332*(6027), 346-349.

Babel: BEAST analysis backing effective linguistics. Retrieved from https://github.com/rbouckaert/Babel

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution, 29*(9), 2157-2167.

Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2012). Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular biology and evolution, 30*(2), 239-243.

Bajić, V., Barbieri, C., Hübner, A., Güldemann, T., Naumann, C., Gerlach, L., . . . Roewer, L. (2018). Genetic structure and sex-biased gene flow in the history of southern African populations. *American journal of physical anthropology, 167*(3), 656-671.

Barbieri, C., Güldemann, T., Naumann, C., Gerlach, L., Berthold, F., Nakagawa, H., . . . Pakendorf, B. (2014). Unraveling the complex maternal history of Southern African Khoisan populations. *American journal of physical anthropology, 153*(3), 435-448.

Barbieri, C., Vicente, M., Oliveira, S., Bostoen, K., Rocha, J., Stoneking, M., & Pakendorf, B. (2014). Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. *PLoS One, 9*(6), e99117.

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

Barnard, A. (1992). *Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples*: Cambridge University Press.

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., . . . Comas, D. (2008). The dawn of human matrilineal diversity. *The American Journal of Human Genetics, 82*(5), 1130-1140.

Bielejec, F., Rambaut, A., Suchard, M. A., & Lemey, P. (2011). SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics, 27*(20), 2910-2912.

Bleek, W. H. I. (1851). *De nominum generibus linguarum Africae australis*. *Copticae, Semiticarum aliarumque sexualium. Bonn: Adolphum Marcum*.

Blench, R. (1993). Ethnographic and linguistic evidence for the prehistory of African ruminant livestock, horses and ponies. *The archaeology of Africa. Food, metals and towns*, 71-103.

Blench, R. (2007). Was there and interchange between Cushitic pastoralists and Khoisan speakers in the prehistory of Southern Africa and how can this be detected. *Sprache und Geschichte in Afrika*.

Bollig, M. (2004). Hunters, foragers, and singing Smiths: The metamorphoses of peripatetic peoples in Africa. *Customary Strangers. New Perspectives on Peripatetic Peoples in the Middle East, Africa, and Asia*, 195-231.

Boonzaier, E. (1997). *The Cape herders: a history of the Khoikhoi of southern Africa*: New Africa Books.

Bostoen, K. (2018). The bantu expansion. In *Oxford research encyclopedia of African history*: Oxford University Press.

Bouckaert, R. (2014). PATH SAMPLING WITH A GUI. Retrieved from https://beast2.blogs.auckland.ac.nz/tag/model-selection-v-1-0-2/

Bouckaert, R. (2015). Spherical Phylogeography with BEAST2.2. *Continuous phylogeography on a sphere.* Retrieved from https://www.beast2.org/tutorials/

Bouckaert, R. (2016). Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ, 4*, e2406.

Bouckaert, R. (2019a). Model selection with nested sampling. Retrieved from https://taming-the-beast.org/tutorials/NS-tutorial/

Bouckaert, R. (2019b). Nested sampling packagin for BEAST. Retrieved from https://github.com/BEAST2-Dev/nested-sampling

Bouckaert, R., & Heled, J. (2014). DensiTree 2: Seeing Trees Through the Forest. bioRxiv, 12401. In.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., . . . Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science, 337*(6097), 957-960.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., . . . De Maio, N. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology, 15*(4), e1006650.

Bouckaert, R. R., Bowern, C., & Atkinson, Q. D. (2018). The origin and expansion of Pama–Nyungan languages across Australia. *Nature ecology & evolution, 2*(4), 741-749.

Bousman, C. B. (1998). The chronological evidence for the introduction of domestic stock into southern Africa. *African Archaeological Review, 15*(2), 133-150.

Breton, G., Schlebusch, C. M., Lombard, M., Sjödin, P., Soodyall, H., & Jakobsson, M. (2014). Lactase persistence alleles reveal partial East African ancestry of southern African Khoe pastoralists. *Current Biology, 24*(8), 852-858.

Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics, 194*(2), 459-471.

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics, 81*(5), 1084-1097.

Bryant, D., & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution, 21*(2), 255-265.

Burrett, R. (2007). Beyond the pots and bones: subtle changes in the archaeological record of Bambata cave, Matobo Hills. *Zimbabwean Prehistory*(27), 31-39.

Campbell, M. C., & Tishkoff, S. A. (2010). The evolution of human genetic and phenotypic variation in Africa. *Current Biology, 20*(4), R166-R173.

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature, 325*(6099), 31-36.

Cashdan, E. (1986). Competition between foragers and food-producers on the Botletli River, Botswana. *Africa, 56*(3), 299-318.

Castrì, L., Tofanelli, S., Garagnani, P., Bini, C., Fosella, X., Pelotti, S., . . . Luiselli, D. (2009). mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: Implications for peopling and migration patterns in sub-Saharan Africa. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 140*(2), 302-311.

Chebanne, A., & Mathes, T. K. (2013). *Tsua lexicon (Unpublished fieldnotes)*. University of Botswana & New York University.

Chiaroni, J., King, R. J., Myres, N. M., Henn, B. M., Ducourneau, A., Mitchell, M. J., . . . Nik-Ahd, M. (2010). The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. *European Journal of Human Genetics, 18*(3), 348.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., . . . Ritchie, G. R. (2011). Modernizing reference genome assemblies. *PLoS Biol, 9*(7), e1001091.

Coelho, M., Sequeira, F., Luiselli, D., Beleza, S., & Rocha, J. (2009). On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC evolutionary biology, 9*(1), 80.

Collins, C. C., Andy. (nd.). *Kua wordlist (Unpublished fieldnotes)*. University of Botswana & New York University.

Cooke, C. (1965). Evidence of human migrations from the rock art of southern Rhodesia. *Africa*, 263-285.

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

De Almeida, A. (1965). *Bushmen and other non-Bantu peoples of Angola: three lectures*: Witwatersrand University Press for the Institute for the Study of Man in Africa.

de Filippo, C., Heyn, P., Barham, L., Stoneking, M., & Pakendorf, B. (2010). Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 141*(3), 382-394.

Diamond, J., & Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science, 300*(5619), 597-603.

Dornan, S. S. (1917). The Tati Bushmen (Masarwas) and Their Language. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 47*, 37-112.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics, 161*(3), 1307-1320.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology, 7*(1), 214.

Dunn, M. (2015). Language phylogenies. *The Routledge handbook of historical linguistics*, 190-211.

E. Paradis, K. Schliep. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*, 526-528.

Ehret, C. (1967). Cattle-keeping and milking in eastern and southern African history: the linguistic evidence. *Journal of African History*, 1-17.

Ehret, C. (1982). The first spread of food production to southern Africa.

Ehret, C. (2008). The early livestock-raisers of southern Africa. *Southern African Humanities, 20*(1), 7-35.

Elderkin, E. D. (2014). Clicks, prosodies and Khoisan. In *Beyond'Khoisan'* (pp. 103-124): John Benjamins.

Elphick, R. (1985). *Khoikhoi and the founding of white South Africa*: Ravan Press.

Engelbrecht, J. A. (1928). Studies oor Korannataal. *Annale van die universiteit van stellenbosch, 6* (Kaapstad: Kapstadt: Nasionale Pers), 45pp.

Epstein, H. (1971). *The origin of the domestic animals of Africa*: Africana publishing corporation.

Estermann, C. (1962). Les twa du sud-ouest de l'Angola. *Anthropos*(H. 3./6), 465-474.

Estermann, C. (1976). *The Ethnography of Southwestern Angola: The Nyaneka-Nkumbi Ethnic Group* (Vol. 2): Africana Pub.

Estermann, C. (1981). The Ethnography of Southwestern Angola: The Herero People (translated and edited by Gibson GD). *Africana Pub. Co*.

Fauvelle-Aymar, F.-X. (2008). Against the'Khoisan paradigm'in the interpretation of Khoekhoe origins and history: a re-evaluation of Khoekhoe pastoral traditions. *Southern African Humanities, 20*(1), 77-92.

Fehn, A.-M. (2014). *Kwadi wordlist*. Unpublished fieldnotes.

Fehn, A.-M. (2016). *A grammar of Ts'ixa (Kalahari Khoe).* (Ph.D.). University of Cologne.

Fehn, A.-M. (2019a). *A dictionary of Ts'ixa: a Khoe language of northern Botswana*. Unpublished manuscript.

Fehn, A.-M. (2019b). Kuvale: A Bantu language of southwestern Angola. *Journal of African Languages and Linguistics, 40*(2), 235-270.

Fehn, A.-M. (2019c). *Survey of Namibian Khwe dialects*. Unpublished fieldnotes.

Fehn, A.-M., & Phiri, A. (2017). Nominal marking in Northern Tshwa (Kalahari Khoe). *Stellenbosch papers in linguistics, 48*, 105-122.

Fehn, A.-M. M., William B.; Blesswell, Kure. (2013). *Survey of Northern Kalahari Khoe*. Unpublished fieldnotes.

Fernandes, V., Triska, P., Pereira, J. B., Alshamali, F., Rito, T., Machado, A., . . . Soares, P. (2015). Genetic stratigraphy of key demographic events in Arabia. *PLoS One, 10*(3), e0118625.

Frobenius, L. (1933). Kulturgeschichte Afrikas Phaidon. In: Zürich.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., . . . de Filippo, C. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature, 514*(7523), 445-449.

Gray, R. D., Bryant, D., & Greenhill, S. J. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1559), 3923-3933.

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science, 323*(5913), 479-483.

Greenberg, J. H. (1963). Universals of language.

Greenberg, J. H. (1972). Linguistic evidence regarding Bantu origins. *The Journal of African History, 13*(2), 189-216.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., & Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences, 112*(43), 13296-13301.

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.

Güldemann, T. (2008). A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *Southern African Humanities, 20*(1), 93-132.

Güldemann, T. (2014). 'Khoisan' linguistic classification today*. In T. Güldemann & A.-M. Fehn (Eds.), *Beyond'khoisan': Historical relations in the kalahari basin* (Vol. 330): John Benjamins Publishing Company.

Güldemann, T. (2020). Changing Profile when encroaching on forager territory. In T. Güldemann, P. McConvell, & R. A. Rhodes (Eds.), *The language of hunter-gatherers* (pp. 114-146): Cambridge University Press.

Güldemann, T., & Elderkin, E. D. (2010). On external genealogical relationships of the Khoe family. In M. K. Brenzinger, Christa (Ed.), *Proceedings of the 1st International Symposium, January 4-8, 2003, Riezlern/Kleinwalsertal*. Cologne: Rüdiger Köppe.

Güldemann, T., & Fehn, A.-M. (2017). The Kalahari Basin area as a "Sprachbund" before the Bantu expansion - an update. In R. Hickey (Ed.), *The Cambridge handbook of areal linguistics* (pp. 500-526). Cambridge: Cambridge University Press.

Güldemann, T., & Stoneking, M. (2008). A historical appraisal of clicks: a linguistic and genetic population perspective. *Annual Review of Anthropology, 37*, 93-109.

Haacke, W. H. (2002). *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*: Gamsberg Macmillan.

Haacke, W. H. (2018). Khoekhoegowab (Nama/Damara). In *The Social and Political History of Southern Africa's Languages* (pp. 133-158): Springer.

Haynie, H. J. (2014). Geography and spatial analysis in historical linguistics. *Language and Linguistics Compass, 8*(8), 344-357.

Heine, B. (1976). *A typology of African languages: Based on the order of meaningful elements* (Vol. 4): D. Reimer.

Heine, B., & Honken, H. (2010). The Kx'a family: a new Khoisan genealogy. *Journal of Asian and African Studies, 79*, 5-36.

Heine, B., & Konig, C. (2008). What can linguistics tell us about early Khoekhoe history? *Southern African Humanities, 20*(1), 235-248.

Henn, B. M., Gignoux, C., Lin, A. A., Oefner, P. J., Shen, P., Scozzari, R., . . . Underhill, P. A. (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proceedings of the National Academy of Sciences, 105*(31), 10693-10698.

Holland, B. R., Huber, K. T., Dress, A., & Moulton, V. (2002). δ plots: a tool for analyzing phylogenetic distance data. *Molecular biology and evolution, 19*(12), 2051-2059.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution, 23*(2), 254-267.

Ingman, M., Kaessmann, H., Pääbo, S., & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature, 408*(6813), 708-713.

Jari Oksanen, F. G. B., Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, and Eduard Szoecs and Helene Wagner. (2019). vegan: Community Ecology Package (Version 2.5-6) [R package]. Retrieved from https://CRAN.R-project.org/package=vegan

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90*(430), 773-795.

Kilian-Hatz, C. (2003). Khwe dictionary (Namibian African Studies 7). *Cologne: Rüdiger Köppe Verlag*.

Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., . . . Villems, R. (2004). Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *The American Journal of Human Genetics, 75*(5), 752-770.

Köhler, O. (1973/74). Neuere Ergebnisse und Hypothesen der Sprachforschung in ihrer Bedeutung für die Geschichte Afrikas. *Paideuma, 19 & 20*, 162-199.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., & Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society open science, 5*(3), 171504.

Lander, F., & Russell, T. (2018). The archaeological evidence for the appearance of pastoralism and farming in southern Africa. *PLoS One, 13*(6), e0198941.

Lander, F., & Russell, T. (2020). A southern African archaeological database of organic containers and materials, 800 cal BC to cal AD 1500: Possible implications for the transition from foraging to livestock-keeping. *PLoS One, 15*(7), e0235226.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., . . . Sirak, K. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature, 536*(7617), 419-424.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., . . . Lipson, M. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature, 513*(7518), 409-413.

Le Quellec, J.-L. (2011). Provoking lactation by the insufflation technique as documented by the rock images of the Sahara. *Anthropozoologica, 46*(1), 65-125.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic biology, 63*(4), 534-542.

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology*: Elsevier.

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput Biol, 5*(9), e1000520.

Lin, M., Siford, R. L., Martin, A. R., Nakagome, S., Möller, M., Hoal, E. G., . . . Henn, B. M. (2018). Rapid evolution of a skin-lightening allele in southern African KhoeSan. *Proceedings of the National Academy of Sciences, 115*(52), 13324-13329.

Liu, D., Duong, N. T., Ton, N. D., Van Phong, N., Pakendorf, B., Van Hai, N., & Stoneking, M. (2019). Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity. *bioRxiv*, 857367.

MacCalman, H. R., & Grobbelaar, B. (1965). *Preliminary Report of Two Stone-Working Ovatjimba Groups in the Northern Koakoveld [sic] of South West Africa*.

Macholdt, E., Lede, V., Barbieri, C., Mpoloka, S. W., Chen, H., Slatkin, M., . . . Stoneking, M. (2014). Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Current Biology, 24*(8), 875-879.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., . . . Tandon, A. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature, 538*(7624), 201-206.

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics, 93*(2), 278-288.

Marks, S. J., Montinaro, F., Levy, H., Brisighelli, F., Ferri, G., Bertoncini, S., . . . Mitchell, P. (2015). Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. *Molecular biology and evolution, 32*(1), 29-43.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., . . . Novak, M. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature, 528*(7583), 499-503.

MAUENSTEIN, A. (1980). Rites et coutumes liés à l'élevage du bétail dans le sud de l'Angola. *Collectanea Instituti Anthropos St Augustin*(17), 9-222.

McGregor, W. B. F., Anne-Maria; Christfried Naumann. (nd). *Shua data collected with speaker Blesswell Kure from Nata*. Unpublished fieldnotes.

Meinhof, C. (1930). Das Verhältnis der Buschmannsprachen zum Hottentottischen. *Wiener Zeitschrift für die Kunde des Morgenlandes, 37*, 219-229.

Mellars, P. (2006). Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proceedings of the National Academy of Sciences, 103*(25), 9381-9386.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., . . . De Filippo, C. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science, 338*(6104), 222-226.

Mineiro, N. (2016). *The link between Y chromosome variation and the surnames and clans of the patrilineal hunter-gatherer Tshwao population in Zimbabwe.* Faculdade de Ciências, Universidade do Porto.

Mitchell, P. (1997). Late Pleistocene and Holocene Hunter-Gatherers of the Matopos: An Archaeological Study of Change and Continuity in Zimbabwe. In: JSTOR.

Mitchell, P. (2010). Genetics and southern African prehistory: an archaeological view. *J Anthropol Sci, 88*, 73-92.

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

Nakagawa, H. (2011). ǂ*Haba fieldnotes*. Unpublished manuscript.

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

Nicholls, G. K., & Gray, R. D. (2006). Quantifying uncertainty in a stochastic model of vocabulary evolution. *Phylogenetic methods and the prehistory of languages*, 161-171.

Nienaber, G. S. (1963). *Hottentots*. Pretoria: Van Schaik.

Oliveira, S. (2019). *Inferring the demographic history of southern Angola: a key region for understanding human settlement in Southern Africa.* (phD). Faculdade de Ciências, Universidade do Porto.

Oliveira, S., Fehn, A.-M., Aço, T., Lages, F., Gayá-Vidal, M., Pakendorf, B., . . . Rocha, J. (2017). The maternal genetic history of the Angolan Namib Desert: a key region for understanding the peopling of southern Africa. *bioRxiv*, 162230.

Oliveira, S., Fehn, A. M., Aço, T., Lages, F., Gayà-Vidal, M., Pakendorf, B., . . . Rocha, J. (2018). Matriclans shape populations: Insights from the Angolan Namib Desert into the maternal genetic history of southern Africa. *American journal of physical anthropology, 165*(3), 518-535.

Oliveira, S., Hübner, A., Fehn, A.-M., Aço, T., Lages, F., Pakendorf, B., . . . Rocha, J. (2019). The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola. *European Journal of Human Genetics, 27*(3), 475.

Orton, J., Mitchell, P., Klein, R., Steele, T., & Horsburgh, K. A. (2013). An early date for cattle from Namaqualand, South Africa: implications for the origins of herding in southern Africa. *Antiquity, 87*(335), 108-120.

Page, R. D., & Holmes, E. C. (2009). *Molecular evolution: a phylogenetic approach*: John Wiley & Sons.

Pakendorf, B., de Filippo, C., & Bostoen, K. (2011). Molecular perspectives on the Bantu expansion: a synthesis. *Language Dynamics and Change, 1*(1), 50-88.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., . . . Froment, A. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science, 356*(6337), 543-546.

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

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet, 2*(12), e190.

Penny, D., & Hendy, M. (2004). Phylogenetics: parsimony and distance methods. *Handbook of statistical genetics*.

Penny, D., McComish, B. J., Charleston, M. A., & Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution, 53*(6), 711-723.

Phiri, A. (2015-2019). *Tjwao Dictionary*. Unpublished fieldnotes.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., . . . Naumann, C. (2012). The genetic prehistory of southern Africa. *Nature communications, 3*, 1143.

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

Pinto, J. C., Oliveira, S., Teixeira, S., Martins, D., Fehn, A. M., Aço, T., . . . Rocha, J. (2016). Food and pathogen adaptations in the Angolan Namib desert: Tracing the spread of lactase persistence and human African trypanosomiasis resistance into southwestern Africa. *American journal of physical anthropology, 161*(3), 436-447.

Pleurdeau, D., Imalwa, E., Détroit, F., Lesur, J., Veldman, A., Bahain, J.-J., & Marais, E. (2012). “Of sheep and men”: earliest direct evidence of caprine domestication in southern Africa at Leopard Cave (Erongo, Namibia). *PLoS One, 7*(7), e40340.

Pratchett, L. J. (2020). Language contact and change in eastern Botswana: New insights from the pronominal system of an undocumented Kalahari Khoe language. *Language in Africa, 1*(1), 34-64.

Prussin, L. (1995). African nomadic architecture: Space, place, and gender.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

QGIS.org. (2020). QGIS Geographic Information System. Retrieved from http://qgis.org/

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., & Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature genetics, 23*(4), 437-441.

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

Rambaut, A. (2016-2018). FigTree: Tree Figure Drawing Tool. (Version 1.4.4). Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology, 67*(5), 901.

Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W.-Y., Froment, A., Anagnostou, P., . . . Omar, S. A. (2014). Genetic origins of lactase persistence and the spread of pastoralism in Africa. *The American Journal of Human Genetics, 94*(4), 496-510.

Raven-Hart, R. (1967). *Before Van Riebeeck: Callers at South Africa from 1488 to 1652*: C. Struik.

Redinha, J. (1975). *Etnias e culturas de Angola*: Instituto de investigação científica de Angola com a colaboração do Banco de ….

Robbins, L., Campbell, A., Murphy, M., Brook, G., Srivastava, P., & Badenhorst, S. (2005). The advent of herding in southern Africa: early AMS dates on domestic livestock from the Kalahari Desert. *Current Anthropology, 46*(4), 671-677.

Robbins, L. H., Campbell, A. C., Murphy, M. L., Brook, G. A., Liang, F., Skaggs, S. A., . . . Badenhorst, S. (2008). Recent archaeological research at Toteng, Botswana: early domesticated livestock in the Kalahari. *Journal of African Archaeology, 6*(1), 131-149.

Rocha, J. (2012). The evolution of lactase persistence. *Antropologia Portuguesa*(29), 121-137.

Rocha, J., & Fehn, A. M. (2016). Genetics and Demographic History of the Bantu. *eLS*, 1-9.

RStudio Team. (2019). RStudio: Integrated Development for R. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Sadr, K. (2013). The archaeology of herding in southernmost Africa. *Oxford Handbook of African Archaeology*, 645-655.

Sadr, K. (2015). Livestock first reached southern Africa in two separate events. *PLoS One, 10*(8), e0134215.

Sadr, K., & Sampson, C. G. (2006). Through thick and thin: early pottery in southern Africa. *Journal of African Archaeology, 4*(2), 235-252.

Salas, A., Richards, M., De la Fe, T., Lareu, M.-V., Sobrino, B., Sánchez-Diz, P., . . . Carracedo, Á. (2002). The making of the African mtDNA landscape. *The American Journal of Human Genetics, 71*(5), 1082-1111.

Schlebusch, C. M. (2010). *Genetic variation in Khoisan-speaking populations from southern Africa.* (phD). University of the Witwatersrand Johannesburg, South Africa.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., . . . Soodyall, H. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science, 358*(6363), 652-655.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., . . . Blum, M. G. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science, 338*(6105), 374-379.

Schliep, K., Potts, A. A., Morrison, D. A., & Grimm, G. W. (2016). *Intertwining phylogenetic trees and networks* (2167-9843). Retrieved from

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics, 27*(4), 592-593.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., . . . Oefner, P. J. (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *The American Journal of Human Genetics, 74*(5), 1023-1034.

Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., . . . Schlebusch, C. (2019). Mozambican genetic variation provides new insights into the Bantu expansion. *bioRxiv*, 697474.

Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., . . . Schlebusch, C. (2020). Along the Indian Ocean coast: genomic variation in Mozambique provides new insights into the Bantu expansion. *Molecular biology and evolution, 37*(2), 406-416.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., . . . Peltzer, A. (2017). Reconstructing prehistoric African population structure. *Cell, 171*(1), 59-71. e21.

Smith, A. B. (2017). Why would southern African hunters be reluctant food producers? *Hunter Gatherer Research, 2*(4), 415-435.

Smith, B., Ouzman, S., Chippindale, C., Dowson, T., Mitchell, P., Morris, D., . . . Ouzman, S. (2004). Taking stock: identifying Khoekhoen herder rock art in southern Africa. *Current Anthropology, 45*(4), 499-526.

Soares, P., Rito, T., Pereira, L., & Richards, M. B. (2016). A genetic perspective on African prehistory. In *Africa from MIS 6-2* (pp. 383-405): Springer.

Stanyon, R., Sazzini, M., & Luiselli, D. (2009). Timing the first human migration into eastern Asia. *Journal of biology, 8*(2), 18.

Stoneking, M. (2016). *An introduction to molecular anthropology*: John Wiley & Sons.

Tadmor, U., Haspelmath, M., & Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica, 27*(2), 226-246.

ten Raa, E. (2012). *A Dictionary of Sandawe: the Lexicon and Culture of a Khoesan People of Tanzania*: ed. by Christopher Ehret & Patricia Ehret. Cologne: Rüdiger Köppe.

Tishkoff, S. A., Gonder, M. K., Henn, B. M., Mortensen, H., Knight, A., Gignoux, C., . . . Ramakrishnan, U. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular biology and evolution, 24*(10), 2180-2195.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., . . . Doumbo, O. (2009). The genetic structure and history of Africans and African Americans. *Science, 324*(5930), 1035-1044.

Tofanelli, S., Ferri, G., Bulayeva, K., Caciagli, L., Onofri, V., Taglioli, L., . . . Berti, A. (2009). J1-M267 Y lineage marks climate-driven pre-historical human displacements. *European Journal of Human Genetics, 17*(11), 1520-1524.

Traill, A. (1982). *Notecards from a cross-Khoisan survey*. Unpublished fieldnotes.

Traill, A., & Nakagawa, H. (2000). A historical! Xóõ-| Gui contact zone: linguistic and other relations. *The state of Khoesan languages in Botswana*, 1-17.

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., van Helden, P. D., . . . Henn, B. M. (2016). Fine-scale human population structure in southern Africa reflects ecogeographic boundaries. *Genetics, 204*(1), 303-314.

Vansina, J., & Fansina, G. (2004). *How societies are born: Governance in West Central Africa before 1600*: University of Virginia Press.

Visser, H. (2001). *Naro Dictionary: Naro-English, English-Naro*: Naro Language Project.

Vivelo, F. R. (1977). *The Herero of Western Botswana: Aspects of change in a group of Bantu-speaking cattle herders (American Ethnological Society Monographs 61)*: New York: West Publishing Co.

Vossen, R. (1997). *Die Khoe-Sprachen: Ein Beitrag zur Erforschung der Sprachgeschichte Afrikas* (Vol. 12): Rüdiger Köppe Verlag.

Westphal, E. O. (1963). The linguistic prehistory of southern Africa: Bush, Kwadi, Hottentot, and Bantu linguistic relationships. *Africa*, 237-265.

Westphal, E. O. J. (1971). The click languages of Southern and Eastern Africa. In J. G. Berry, Joseph H. (Ed.), *Linguistics in Sub-Saharan Africa* (pp. 367-420). Current Trends in Linguistics 7 (ed. by T. Sebeok). The Hague/ Paris: Mouton.

Westphal, E. O. J. (nd-a). *Kwadi fieldnotes and recordings*. University of Cape Town Archives.

Westphal, E. O. J. (nd-b). *Shua fieldnotes and recordings*. University of Cape Town Archives.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.: Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wuras, C. F. (1920). *An outline of the Bushman language*: Reimer.

Appendix

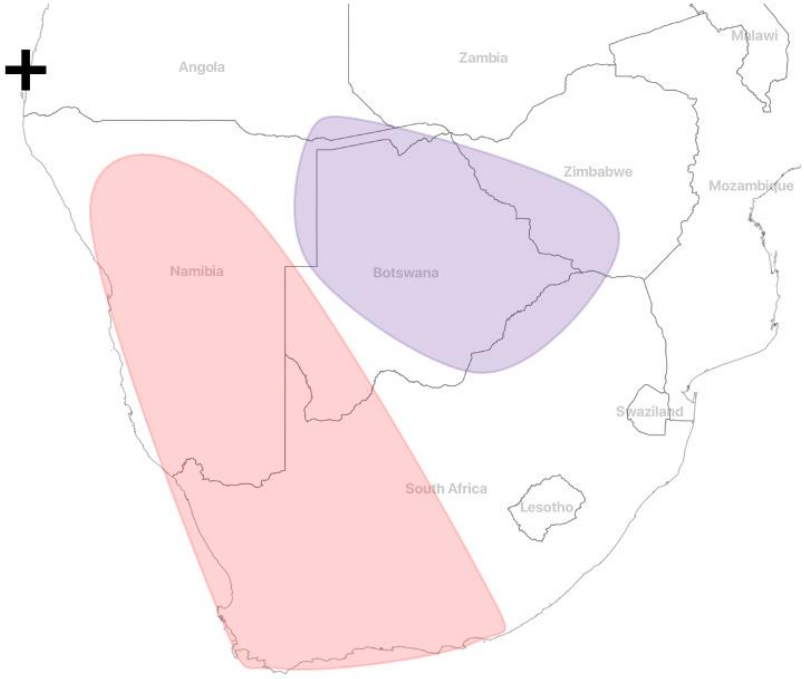

Figure S 1 - Areal coverage of the Khoe-Kwadi languages. The Khoekhoe sub-branch is in red, while Kalahari-Khoe is in purple. The black cross indicates the location of Kwadi.

Table S 1 - Source and origin of the 35 Khoe-Kwadi dialects from the linguistic analysis.

Population	Country	Source
!Ora (Engelbrecht)	South Africa	Engelbrecht (1928)
!Ora (Meinhof)	South Africa	Meinhof (1930)
!Ora (Wuras)	Namibia	Wuras (1920)
||Ani (Fehn)	Botswana	Fehn (2013)
||Ani (Traill)	Botswana	Traill (1982)
||Xoo	Namibia	Fehn (2019c)
Buga	Botswana	Fehn (2013)
Cape east	South Africa	Nienaber (1963)
Cape west	South Africa	Nienaber (1963)
Damara	Namibia	Traill (1982)
Danisi	Botswana	Fehn (2013)
Deti	Botswana	Traill (1982)
G|ui	Botswana	Nakagawa (2014)
Hai||om	Namibia	Haacke (2002)
Khute	Botswana	Traill (1982)
Khwe	Namibia	Kilian-Hatz (2003)
Kua (Collins)	Botswana	Collins (nd.)
Kua (Lephephe)	Botswana	Traill (1982)
Kua (Matsetlharobega)	Botswana	Traill (1982)
Kua (Traill)	Botswana	Traill (1982)
Kwadi	Angola	Almeida (nd); Fehn (2014); Westphal (nd-a)
Nama (Haacke)	Namibia	Haacke (2002)
Nama (Traill)	Namibia	Traill (1982)
Naro (Traill)	Botswana	Traill (1982)
Naro (Visser)	Botswana	Visser (2001)
Shua (Fehn)	Botswana	Fehn (2013)
Shua (Kure)	Botswana	McGregor (nd)
Shua (Westphal)	Botswana	Westphal (nd-b)
Tati	Botswana	Dornan (1917)
Tcire	Zimbabwe	Traill (1982)
Tjwao	Zimbabwe	Phiri (2015-2019)
Ts'ao	Botswana	Fehn (2013)
Ts'ixa	Botswana	Fehn (2019a)
Tsua	Botswana	Chebanne and Mathes (2013)
ǂHaba	Botswana	Nakagawa (2011)

Table S 2 - Delta and Q-residual scores computed for individual varieties of Khoe-Kwadi. Kwadi is highlighted in grey.

Group	Delta score	Q-residual score
!Ora (Engelbrecht)	0.194	0.0101
!Ora (Meinhof)	0.190	0.0086
!Ora (Wuras)	0.188	0.0089
||Ani (Fehn)	0.237	0.0088
||Ani (Traill)	0.252	0.0094
||Xoo	0.239	0.0086
Buga	0.260	0.0121
Cape (east)	0.254	0.0219
Cape (west)	0.222	0.0123
Damara	0.197	0.0107
Danisi	0.263	0.0120
Deti	0.330	0.0132
G|ui	0.281	0.0138
Hai||om	0.195	0.0099
Khute	0.268	0.0128
Khwe	0.256	0.0103
Kua (Collins)	0.288	0.0134
Kua (Lephephe)	0.264	0.0107
Kua (Matselharobega)	0.256	0.0096
Kua (Traill)	0.287	0.0103
Kwadi	0.225	0.0108
Nama (Haacke)	0.191	0.0089
Nama (Traill)	0.187	0.0087
Naro (Traill)	0.304	0.0159
Naro (Visser)	0.298	0.0168
Shua (Fehn)	0.279	0.0115
Shua (Kure)	0.265	0.0108
Shua (Wuras)	0.292	0.0138
Tati	0.303	0.0125
Tcire	0.303	0.0145
Tjwao	0.308	0.0138
Ts'ixa	0.268	0.0106
Tsao	0.285	0.0138
Tsua	0.264	0.0095
ǂHaba	0.286	0.0148

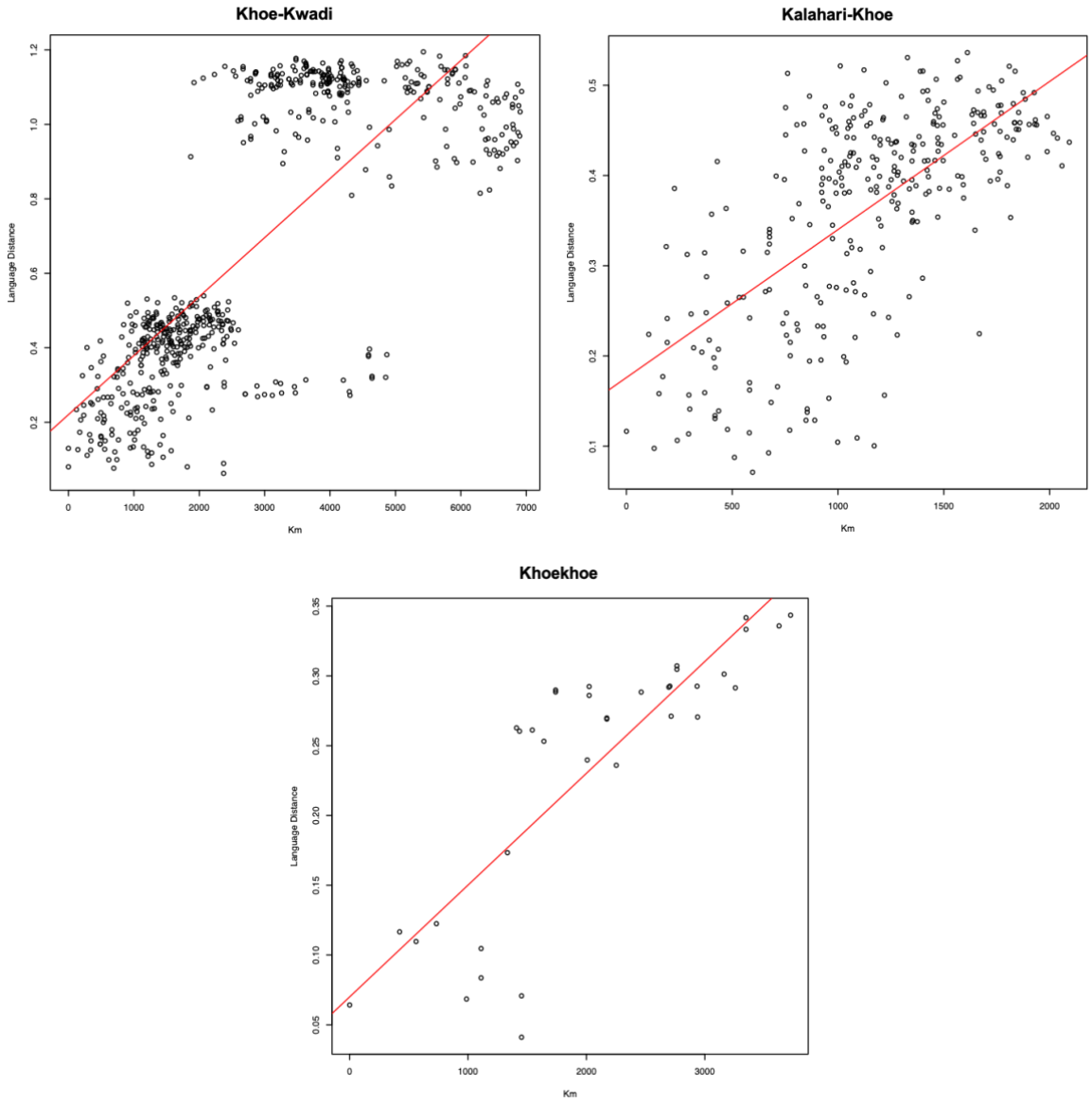

Figure S 2 - Geographic and linguistic distance correlation for the Khoe-Kwadi, Kalahari-Khoe and Khoekhoe groupings. The red line represents the linear model.

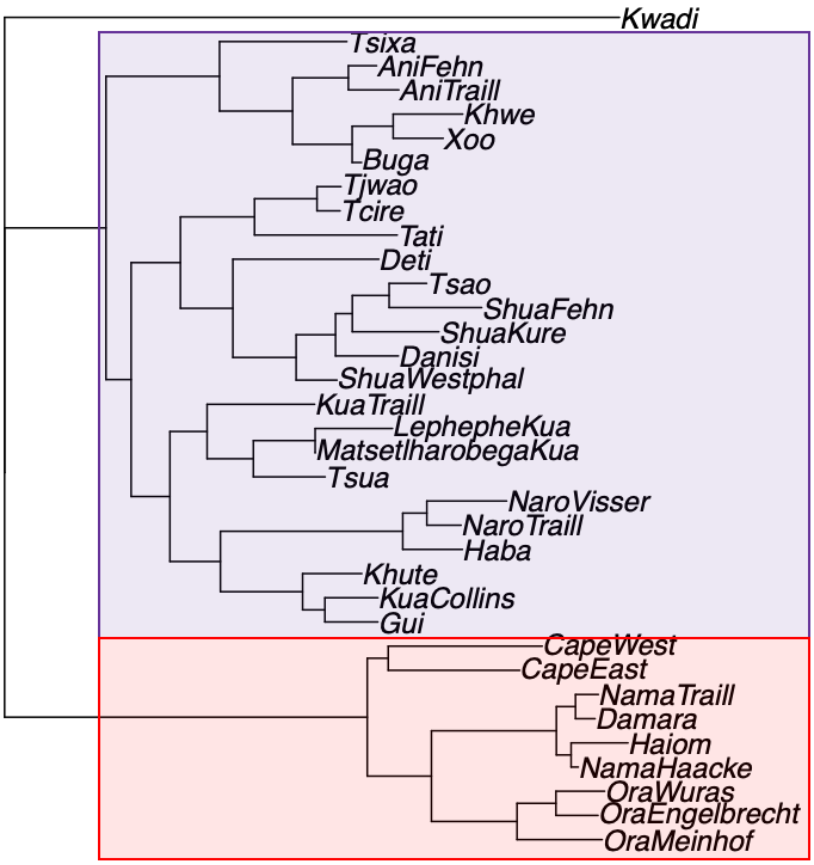

Figure S 3 - UPGMA optimized tree for the 35 Khoe-Kwadi varieties. Red and purple squares indicate, respectively, the Khoekhoe and Kalahari-Khoe sub-branches.

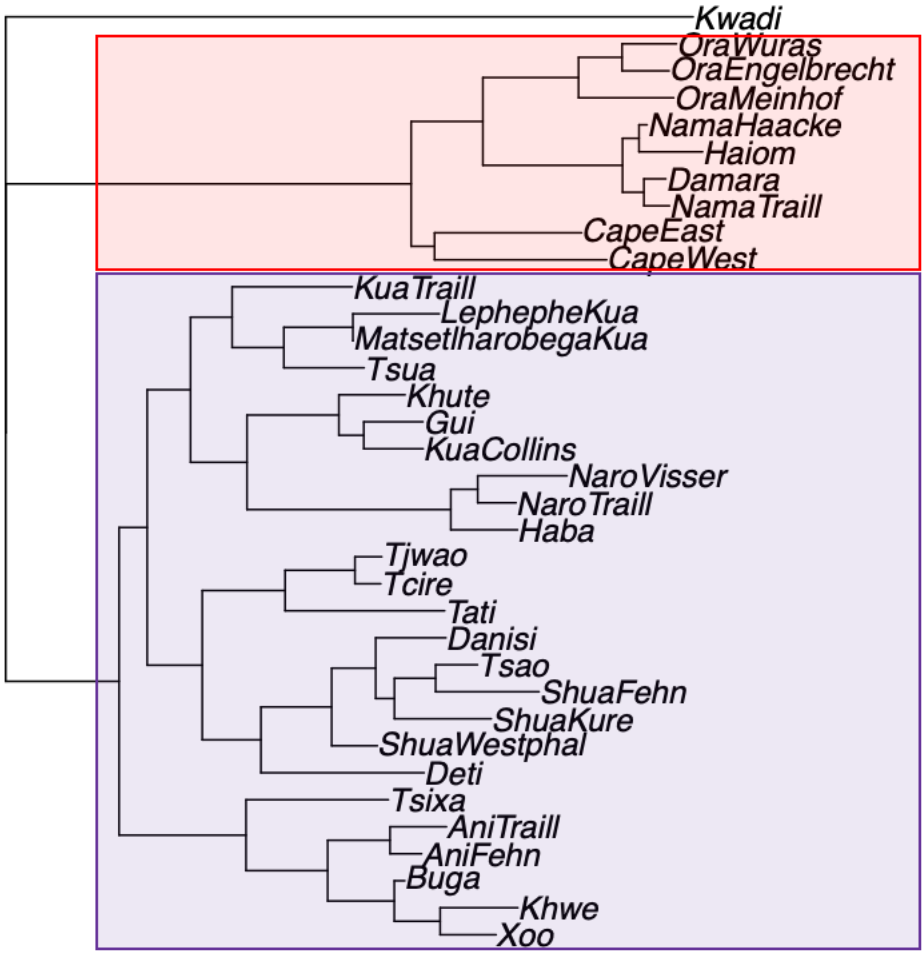

Figure S 4 - Neighbor-Joining-optimized tree for the 35 Khoe-Kwadi varieties. Red and purple squares indicate, respectively, the Khoekhoe and Kalahari-Khoe sub-branches.

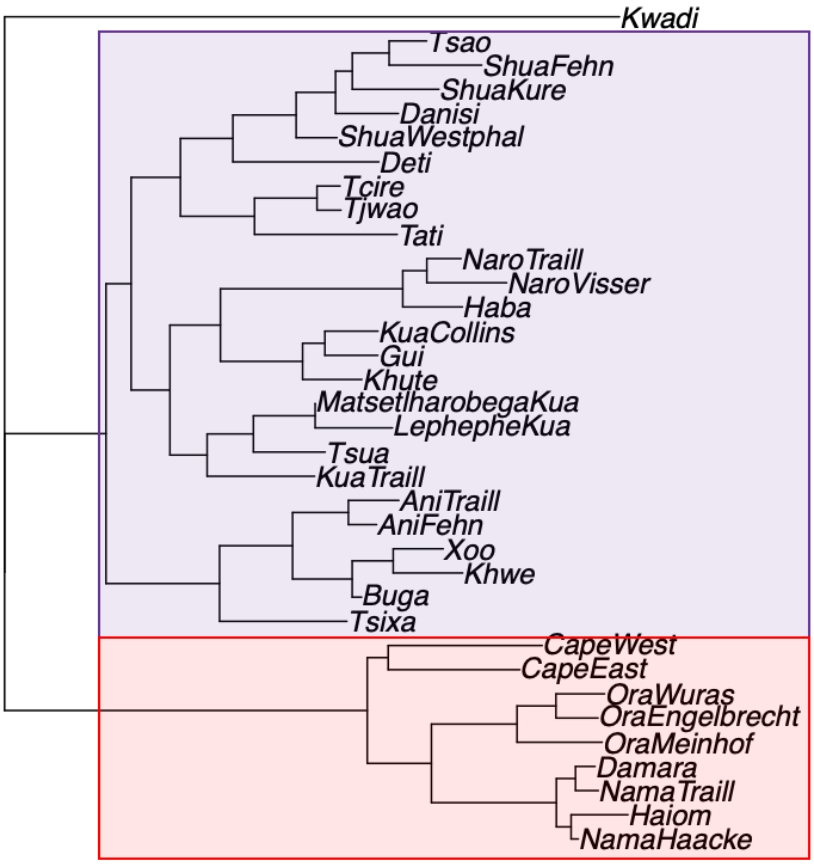

Figure S 5 - Maximum Parsimony-optimized tree for the 35 Khoe-Kwadi varieties. Red and purple squares indicate, respectively, the Khoekhoe and Kalahari-Khoe sub-branches.

Table S 3 - AICM and logML scores for each tested Bayesian model. The chosen models for each analysis are highlighted in grey. The two best-fitting models with the lowest logMLs are indicated with a "*".

Analysis	Evolutionary model	Clock rate	AICM score (with stdev)	logML
Phylogeny	CTMC	Strict	8085.7773 +/- 0.8436	-4180.4217
Phylogeny	Covarion	Strict	8025.7588 +/- 0.8015	-4171.8112
Phylogeny	Dollo	Strict	8074.2526 +/- 0.7715	-4189.1682
Phylogeny	CTMC	Relaxed	8046.3386 +/- 1.3353	-4164.5232*
Phylogeny	Covarion	Relaxed	7989.0360 +/- 1.4007	-4131.2144*
Phylogeny	Dollo	Relaxed	8043.9335 +/- 1.3291	-4171.7913
Phylogeography	CTMC	Strict	7682.9072 +/- 0.8530	-4523.1895
Phylogeography	Covarion	Strict	7623.0118 +/- 0.8651	-4500.2869*
Phylogeography	CTMC	Relaxed	7666.7359 +/- 2.2352	-4507.9685
Phylogeography	Covarion	Relaxed	7618.8052 +/- 2.0156	-4492.8205*

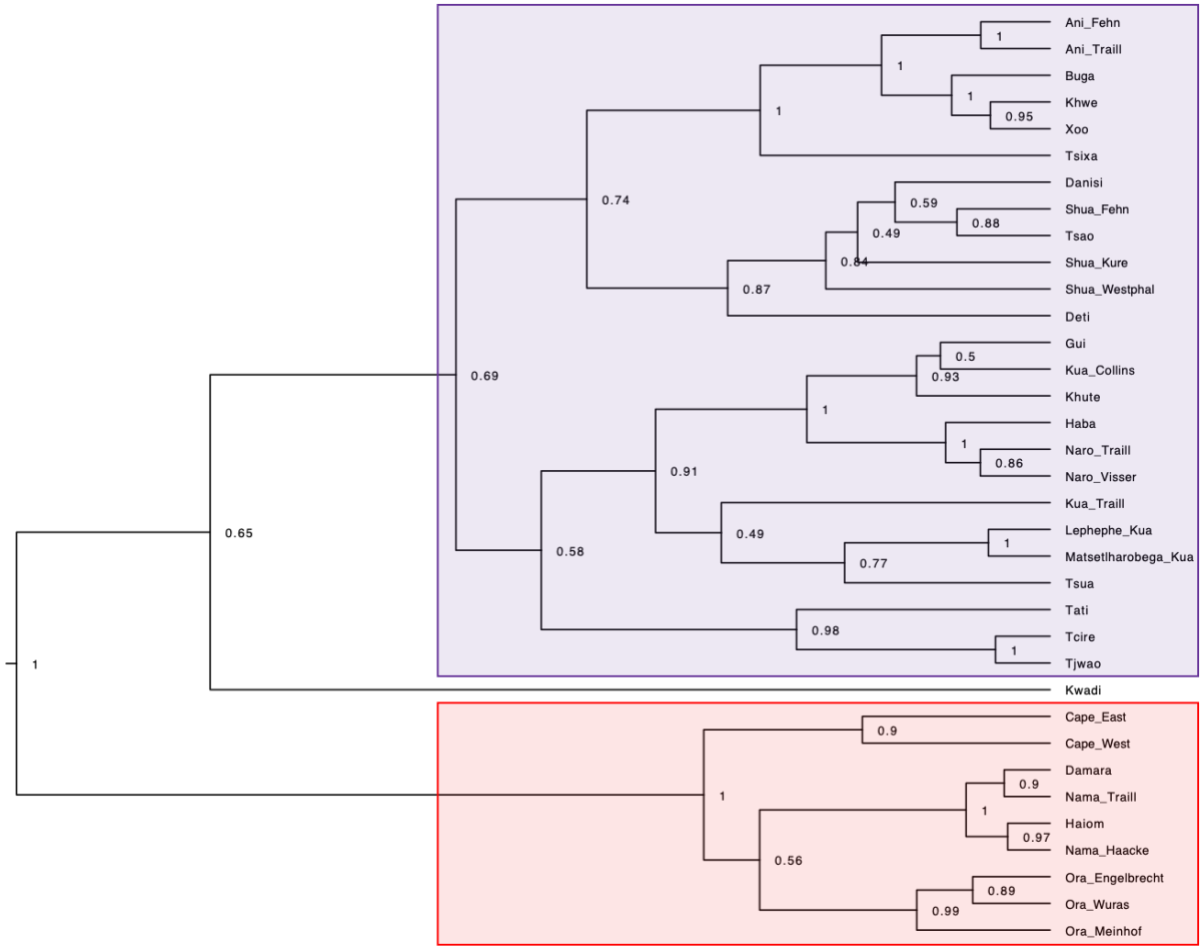

Figure S 6 - Bayesian consensus tree with posterior probabilities in each internal node. Red and purple squares indicate, respectively, the Khoekhoe and Kalahari-Khoe sub-branches.

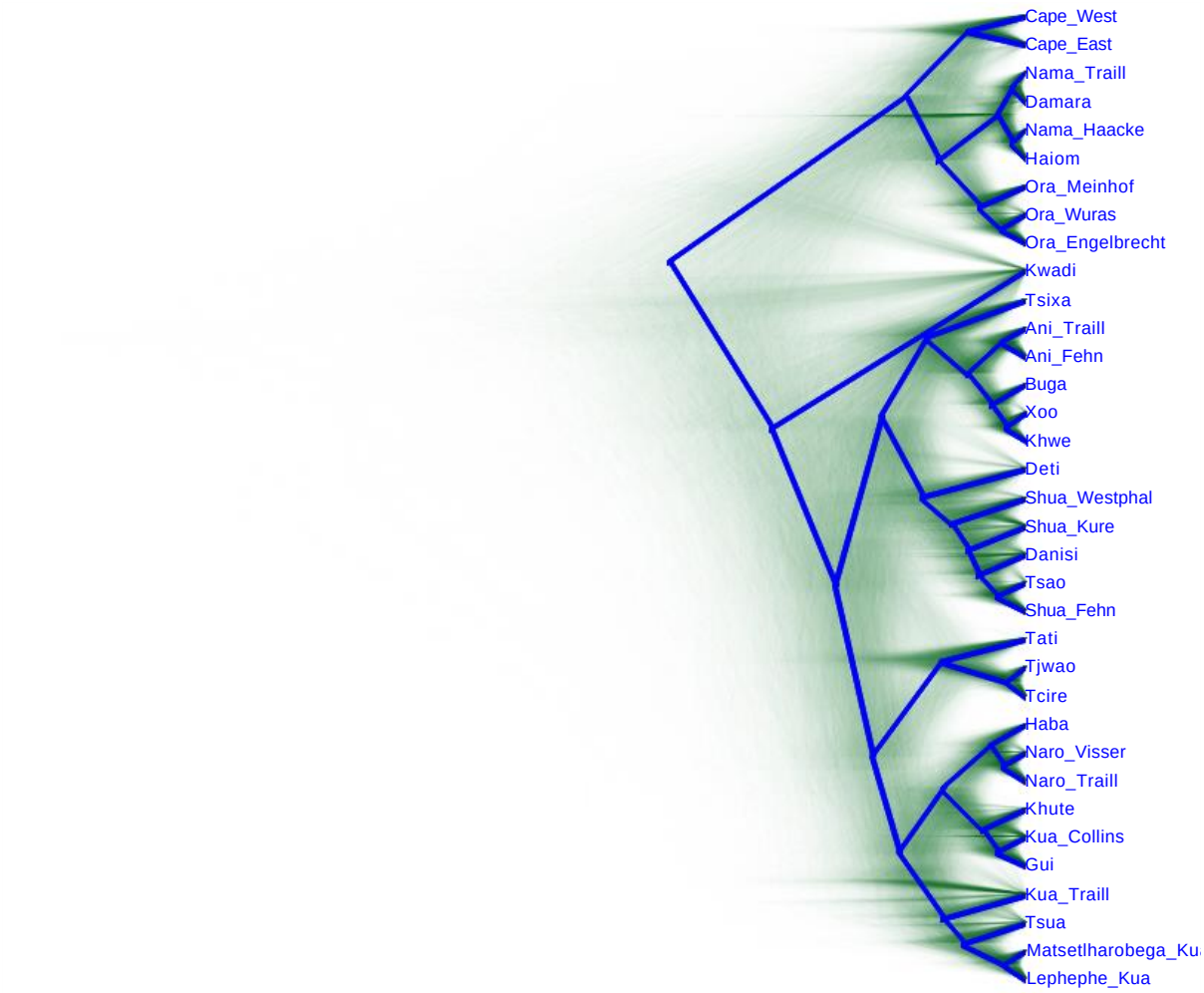

Figure S 7 - DensiTree plot of Bayesian trees displaying variation in internal nodes and alternative topologies.

Table S 4 - Populations genotyped on the Affymetrix Human Origins Array. Khoe-Kwadi populations are coloured in red and the Angolan populations genotyped by our group are coloured in green.

Population	Country	Language Family	Language Sub-Family	Language Complex	Subsistence pattern[6]
!Xun	Angola	Kx'a	Ju		Forager
!Xuun	Namibia	Kx'a	Ju		Forager
AfarWGA	Ethiopia	Afro-Asiatic	Cushitic		Pastoralist
AinToutaWGA	Algeria				
Algerian	Algeria				
BantuKenya	Kenya	Bantu	east		Agropastoralist
BantuSA	South Africa	Bantu	east		Agropastoralist
BantuSA_Herero	Namibia	Bantu	west		Agropastoralist
BantuSA_Ovambo	Namibia	Bantu	west		Agropastoralist
Biaka	Republic of the Congo				Forager
BurburWGA	Morocco				
Damara	Namibia	Khoe-Kwadi	Khoekhoe	Nama-Damara	Peripatetic
Datog	Tanzania	Nilo-Saharan	Nilotic		Pastoralist
Dinka	South Sudan	Nilo-Saharan	Nilotic		Agropastoralist
Egyptian	Egypt				
Esan	Nigeria	Niger-Congo			
G||ana	Namibia	Khoe-Kwadi	Kalahari Khoe west	G||ana	Forager
G|ui	Namibia	Khoe-Kwadi	Kalahari Khoe west	G||ana	Forager
Gambian	Chad				
Hadza	Tanzania	Hadza			Forager
Hadza1	Tanzania	Hadza			Forager
Hai||om	Namibia	Khoe-Kwadi	Khoekhoe west	Hai||om	Forager
Himba	Angola	Bantu	west		Pastoralist
Himba	Namibia	Bantu	west		Pastoralist
JewEthiopian	Ethiopia				
JewLibyan	Libya				
JewMoroccan	Morocco				
JewTunisian	Tunisia				
Ju|'hoan North	Namibia	Kx'a	Ju	south-east	Forager
Ju|'hoan South	Namibia	Kx'a	Ju	south-east	Forager
Kgalagadi	Botswana	Bantu	east		Agropastoralist
Khwe	Namibia	Khoe-Kwadi	Kalahari Khoe west	Khwe	Forager
Kikuyu	Kenya	Bantu	east		Agriculturalist
Kuvale	Angola	Bantu	west		Pastoralist
Kwepe	Angola	Bantu	west		Pastoralist
Kwisi	Angola	Bantu	west		Peripatetic
Luhya	Kenya	Bantu	east		Agriculturalist
Luo	Kenya	Nilo-Saharan	Nilotic		Agropastoralist
Mandenka	Guinea	Niger-Congo			Agropastoralist
Masai	Kenya	Nilo-Saharan	Nilotic		Agropastoralist

[6] The term peripatetic aims to classify low-status, primarily non-food-producing populations.

Table S 4 (cont.) - Populations genotyped on the Affymetrix Human Origins Array. Khoe-Kwadi populations are coloured in red and the Angolan populations genotyped by our group are coloured in green.

Population	Country	Language Family	Language Sub-Family	Language Complex	Subsistence pattern[7]
Mbuti	Democratic Republic of the Congo	Nilo-Saharan			Forager
Mende	Sierra Leone	Niger-Congo			
Mozabite	Algeria	Afro-Asiatic			Herder
Nama	Namibia	Khoe-Kwadi	Khoekhoe	Nama-Damara	Pastoralist
Naro	Botswana	Khoe-Kwadi	Kalahari Khoe west	Naro	Forager
Nyaneka	Angola	Bantu	west		Agropastoralist
Oromo	Ethiopia	Afro-Asiatic	Cushitic		Agropastoralist
Ovimbundu	Angola	Bantu	west		Agropastoralist
Saharawi	western Sahara (Morocco)				
Sandawe	Tanzania	Sandawe			Forager
ShaigiWGA	Sudan				
Shua	Botswana	Khoe-Kwadi	Kalahari Khoe east	Shua	Forager
Somali	Somalia	Afro-Asiatic	Cushitic		
Taa east	Botswana	Tuu	Taa - Lower Nossob	Taa	Forager
Taa north	Botswana	Tuu	Taa - Lower Nossob	Taa	Forager
Taa west	Botswana	Tuu	Taa - Lower Nossob	Taa	Forager
Tjimba	Angola	Bantu	west		Peripatetic
Tshwa	Botswana	Khoe-Kwadi	Kalahari Khoe east	Tshwa - north	Peripatetic
Tswana	Botswana	Bantu	east		Agropastoralist
Tunisian	Tunisia				
Twa	Angola	Bantu	west		Peripatetic
Wambo	Namibia	Bantu	west		
Yoruba	Kenya	Niger-Congo	west		Agropastoralist
ǂHoan	Botswana	Kx'a	ǂ'Amkoe°	west	Forager
ǂKhomani	South Africa	Tuu	!Ui	N|ng°	Forager/Pastoralist

[7] The term peripatetic aims to classify low-status, primarily non-food-producing populations.

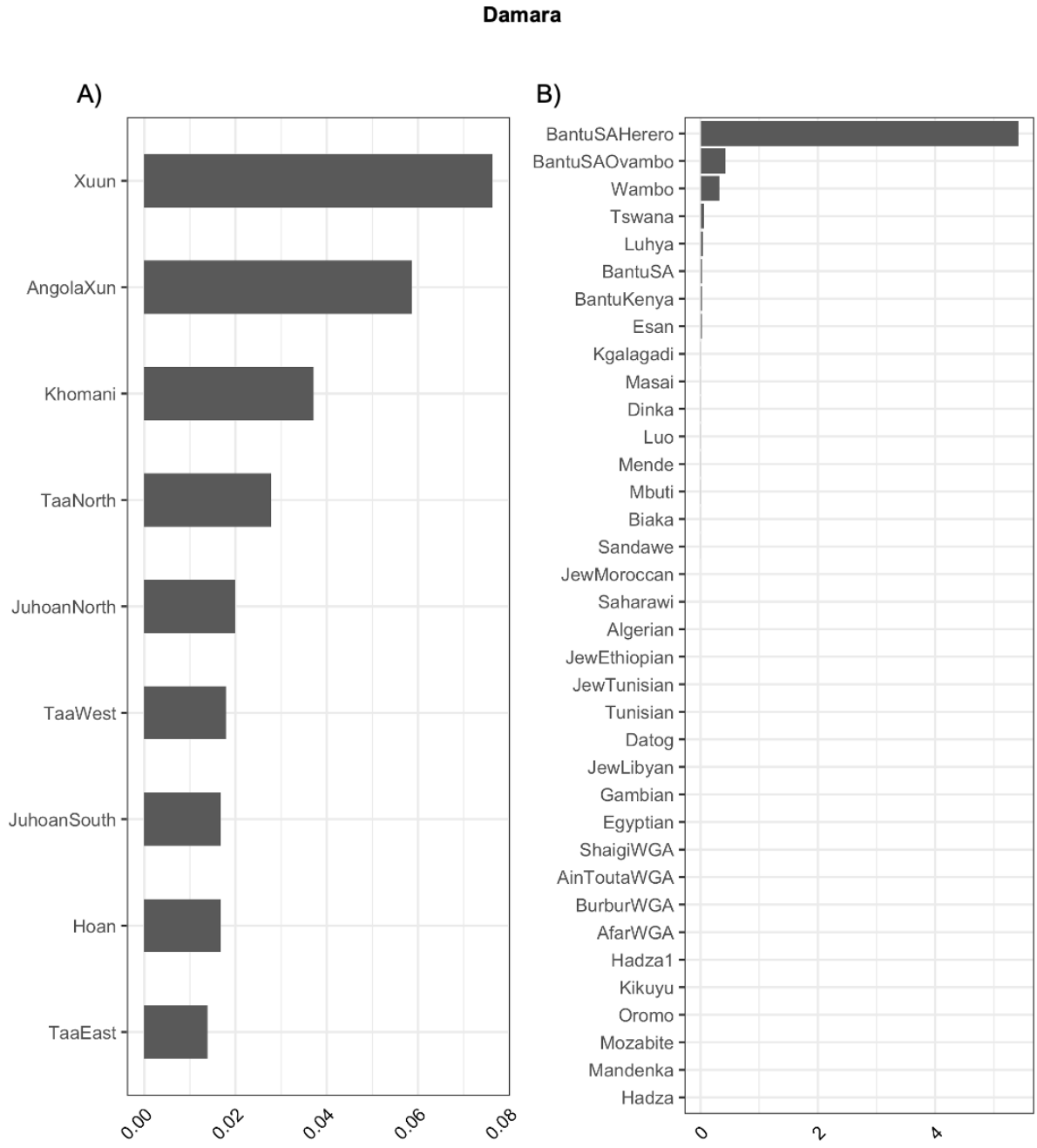

Figure S 8.1 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Damara and Khoisan (A) and Non-Khoisan (B) populations.

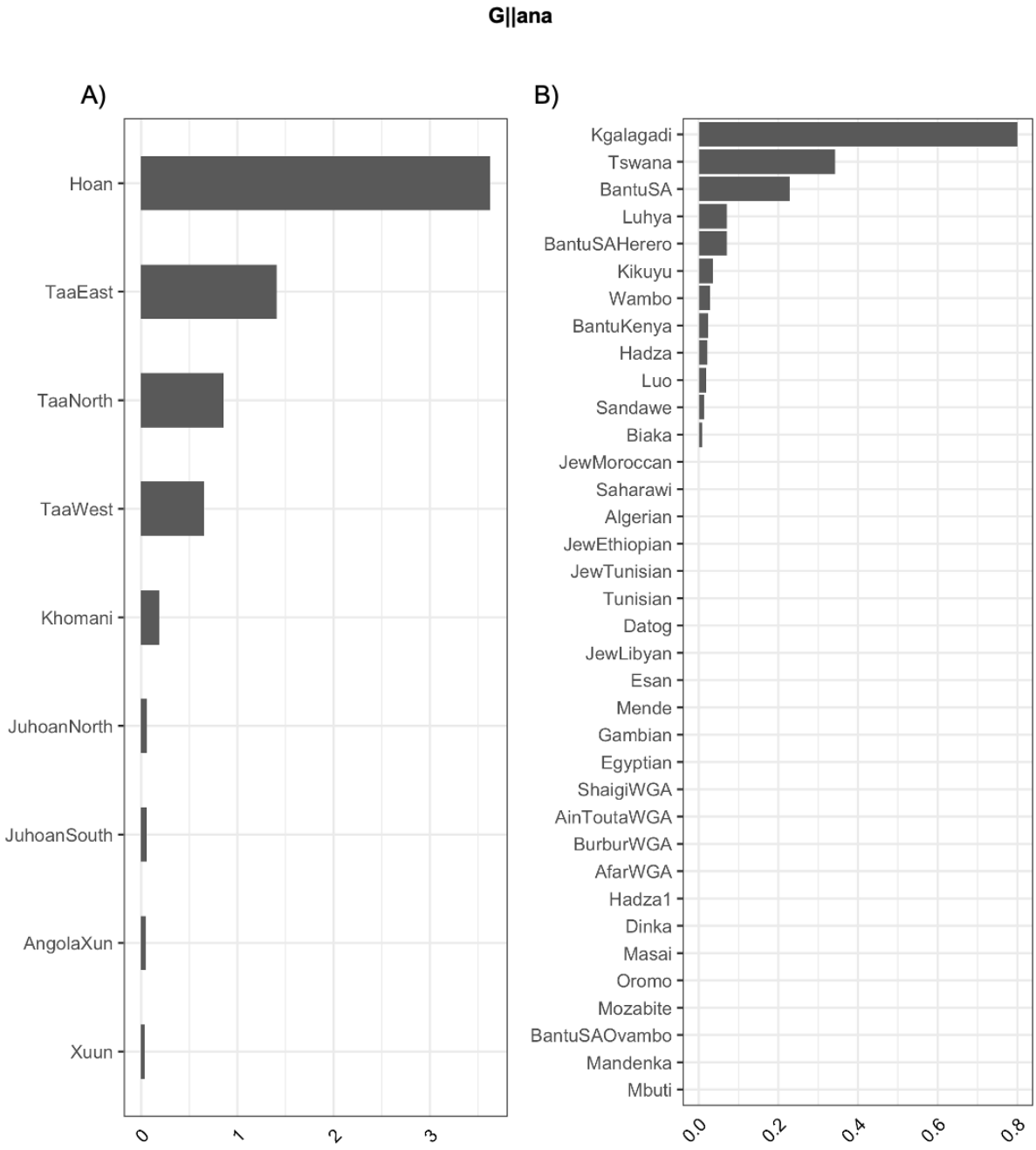

Figure S 8.2 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the G||ana and Khoisan (A) and Non-Khoisan (B) populations.

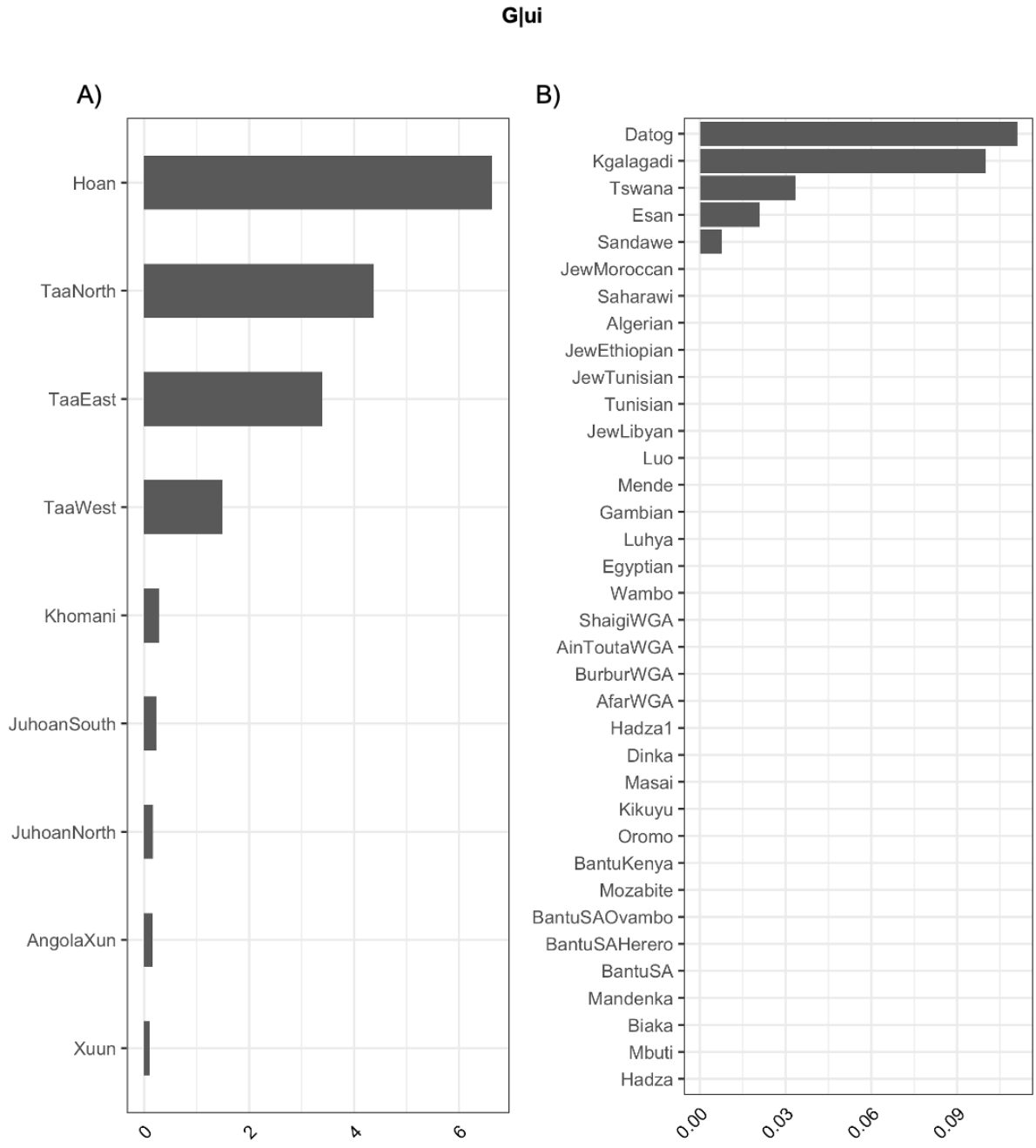

Figure S 8.3 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the G|ui and Khoisan (A) and Non-Khoisan (B) populations.

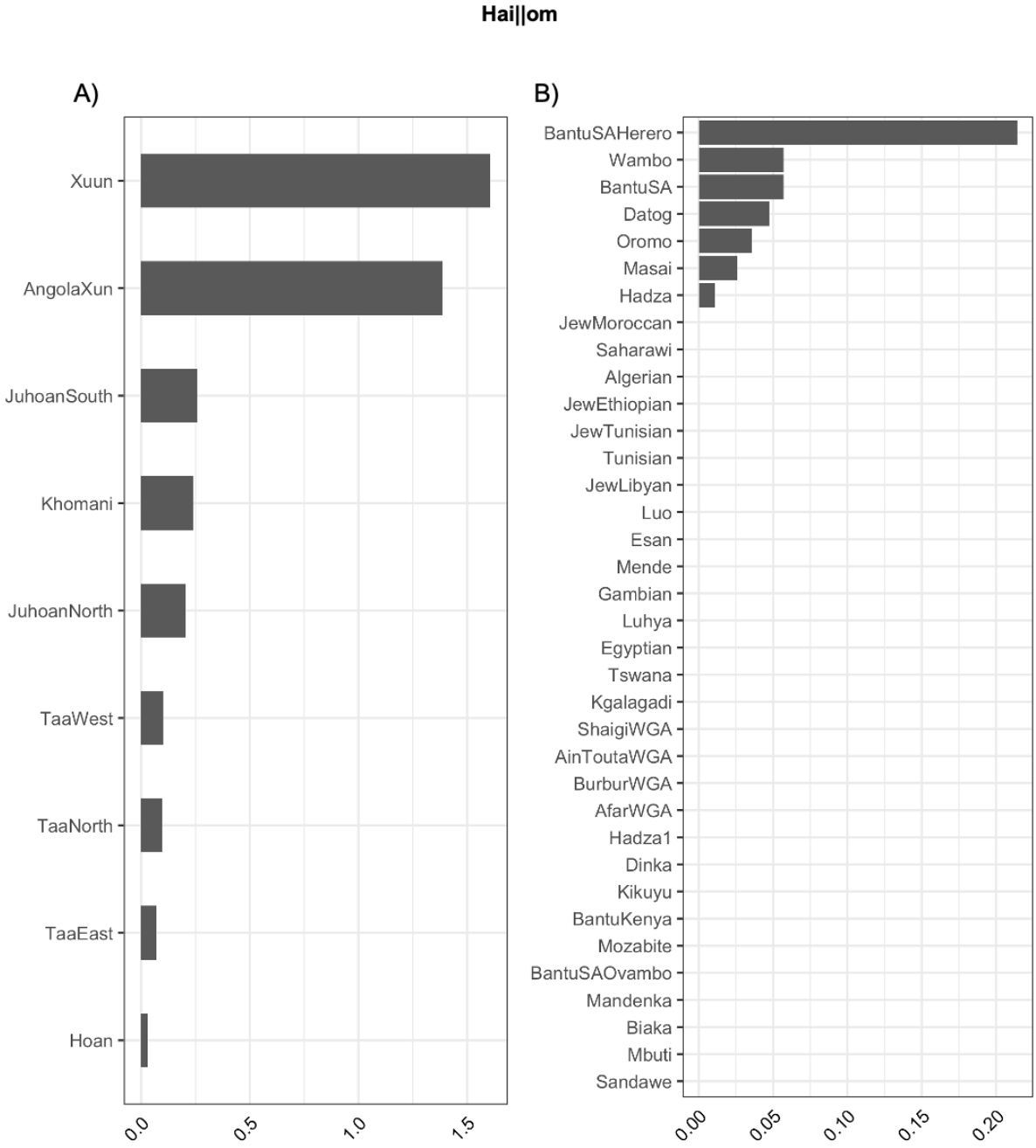

Figure S 8.4 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Hai||om and Khoisan (A) and Non-Khoisan (B) populations.

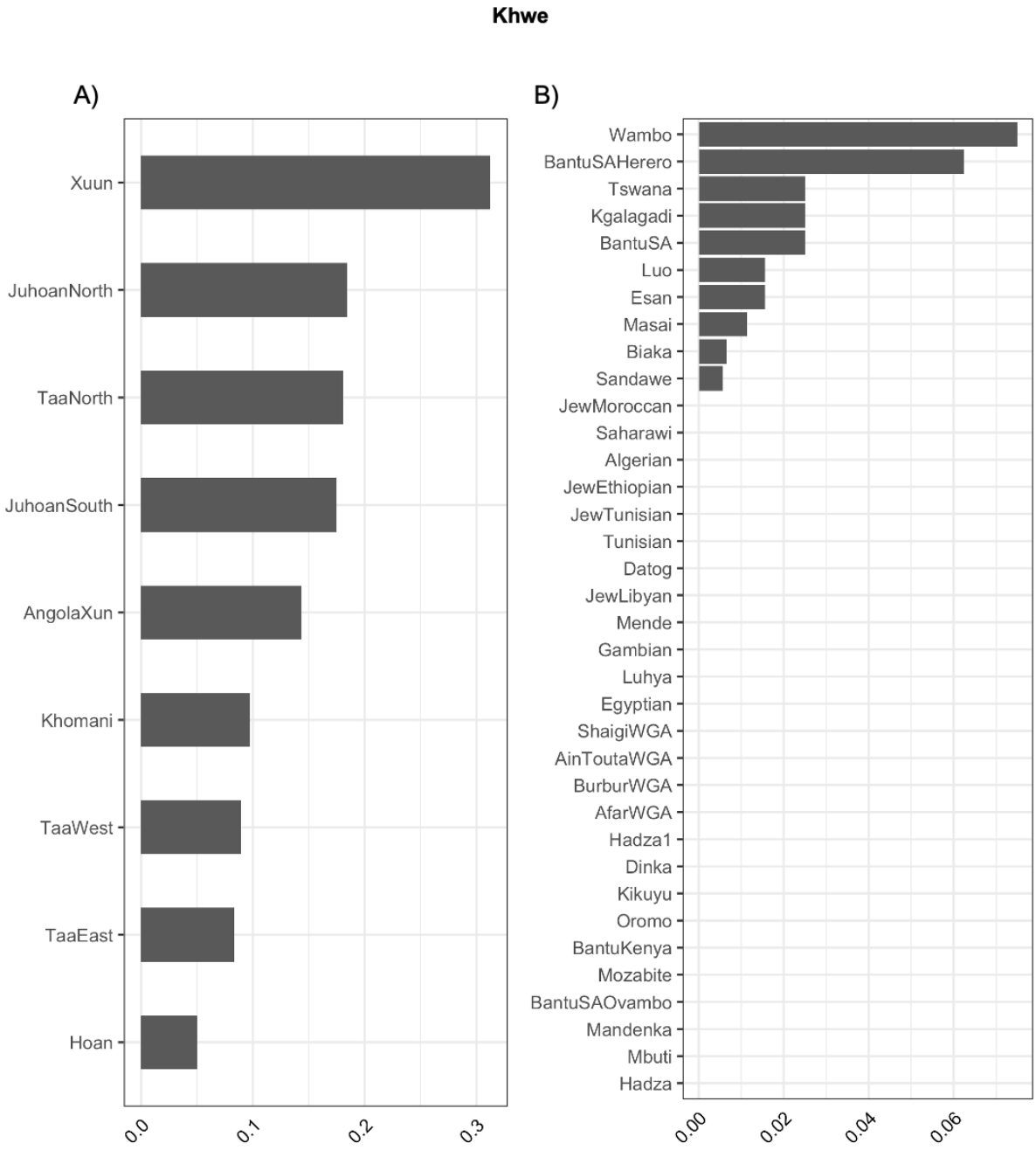

Figure S 8.5 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Khwe and Khoisan (A) and Non-Khoisan (B) populations.

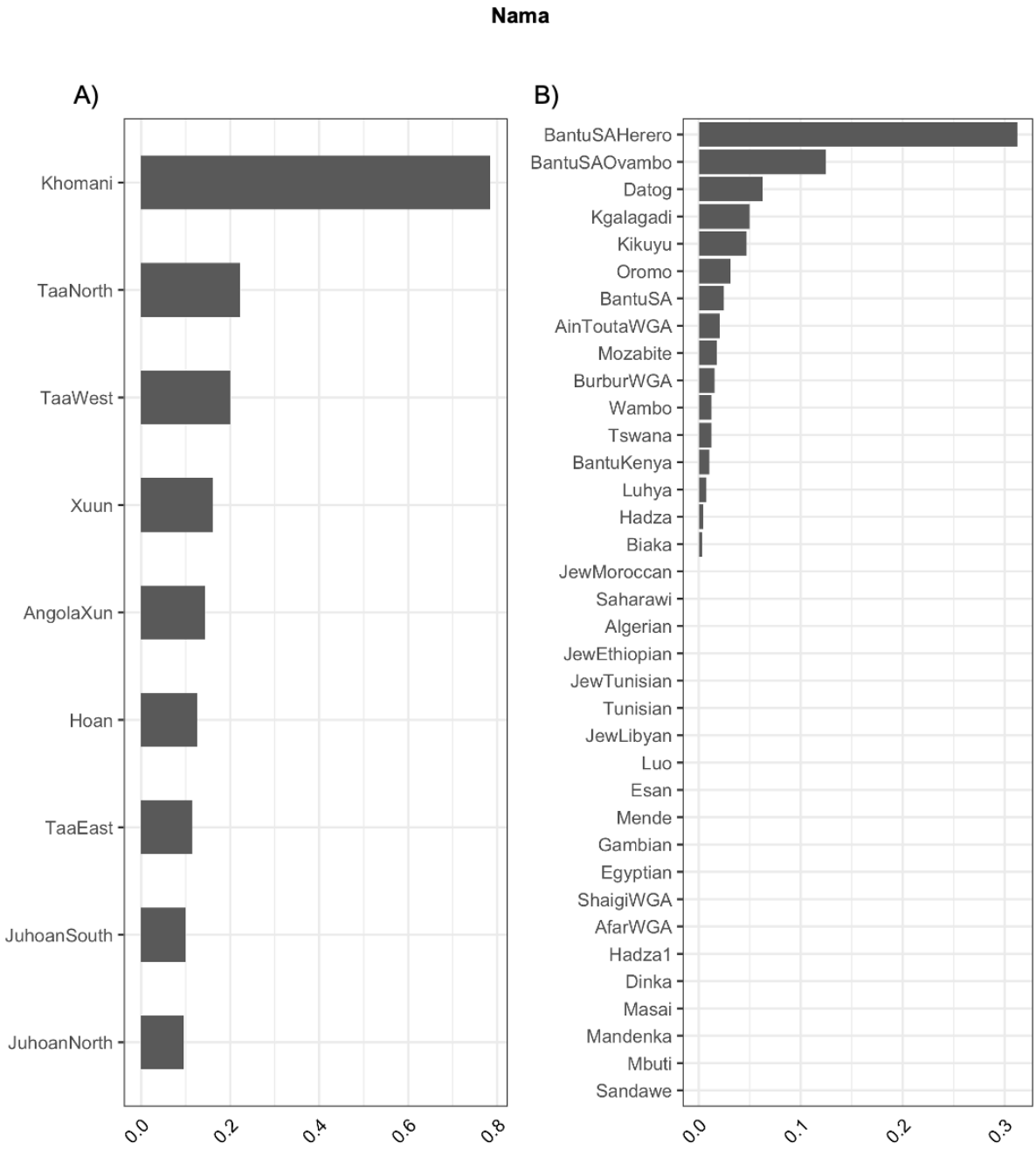

Figure S 8.6 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Nama and Khoisan (A) and Non-Khoisan (B) populations.

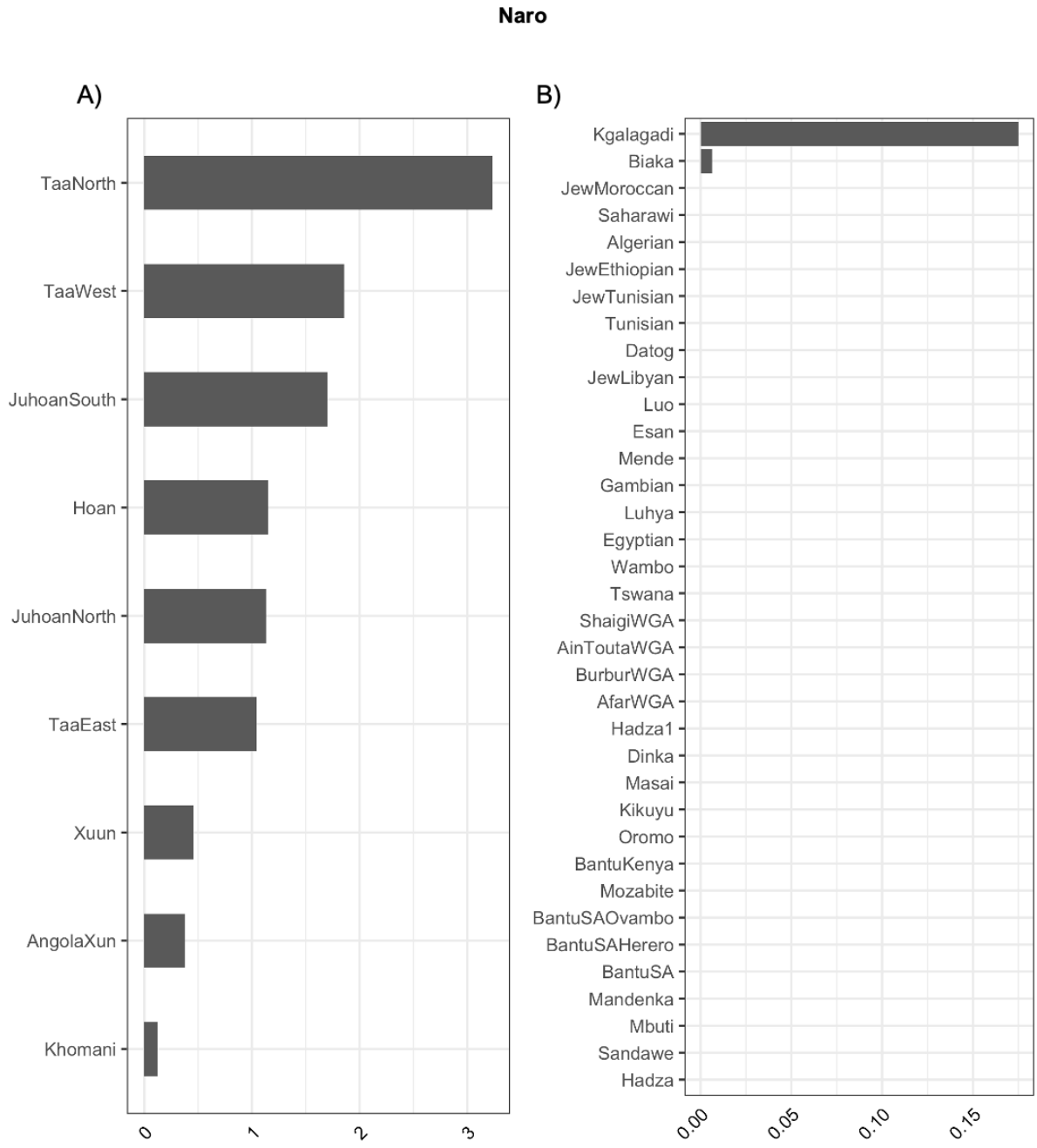

Figure S 8.7 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Naro and Khoisan (A) and Non-Khoisan (B) populations.

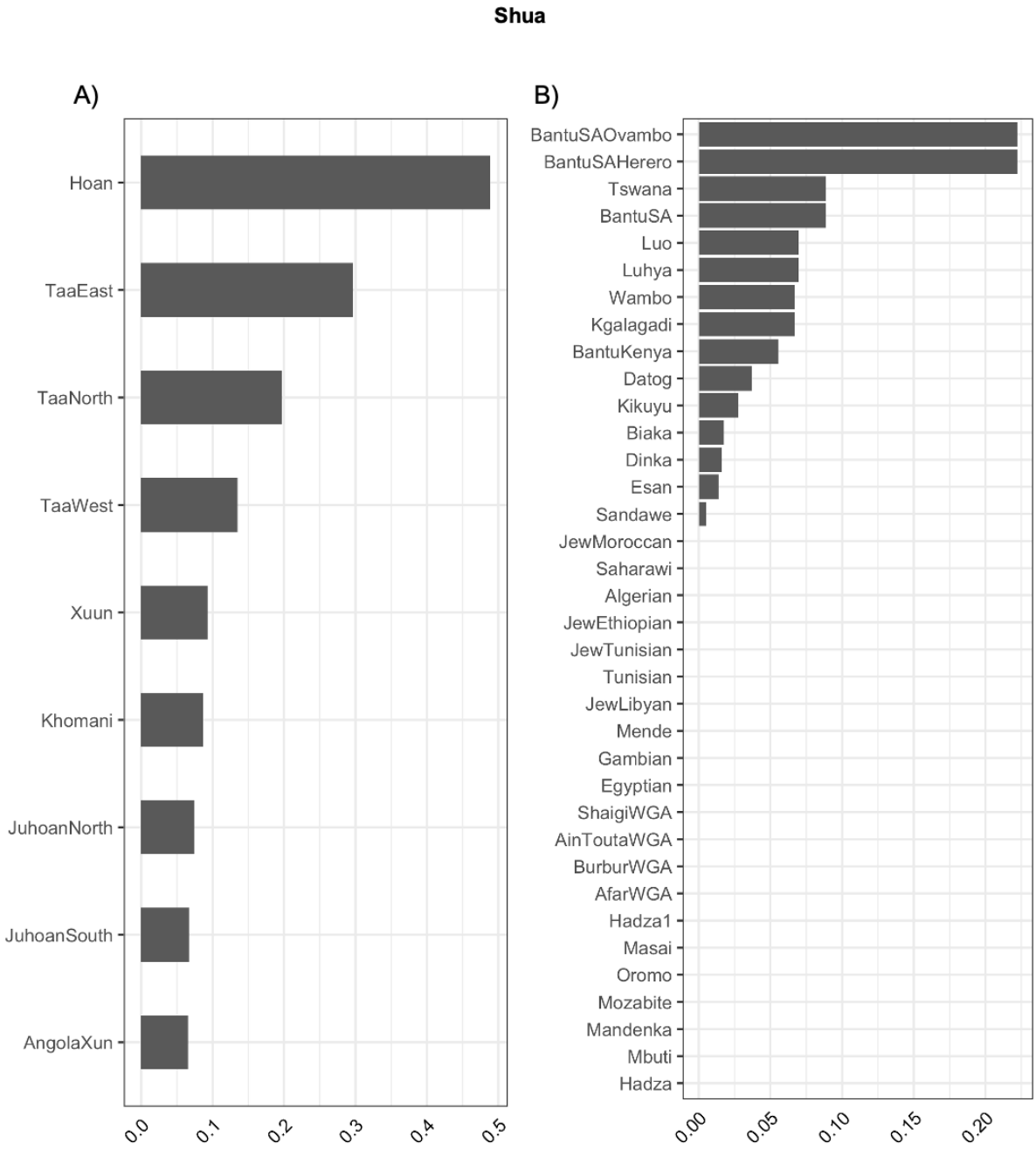

Figure S 8.8 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Shua and Khoisan (A) and Non-Khoisan (B) populations.

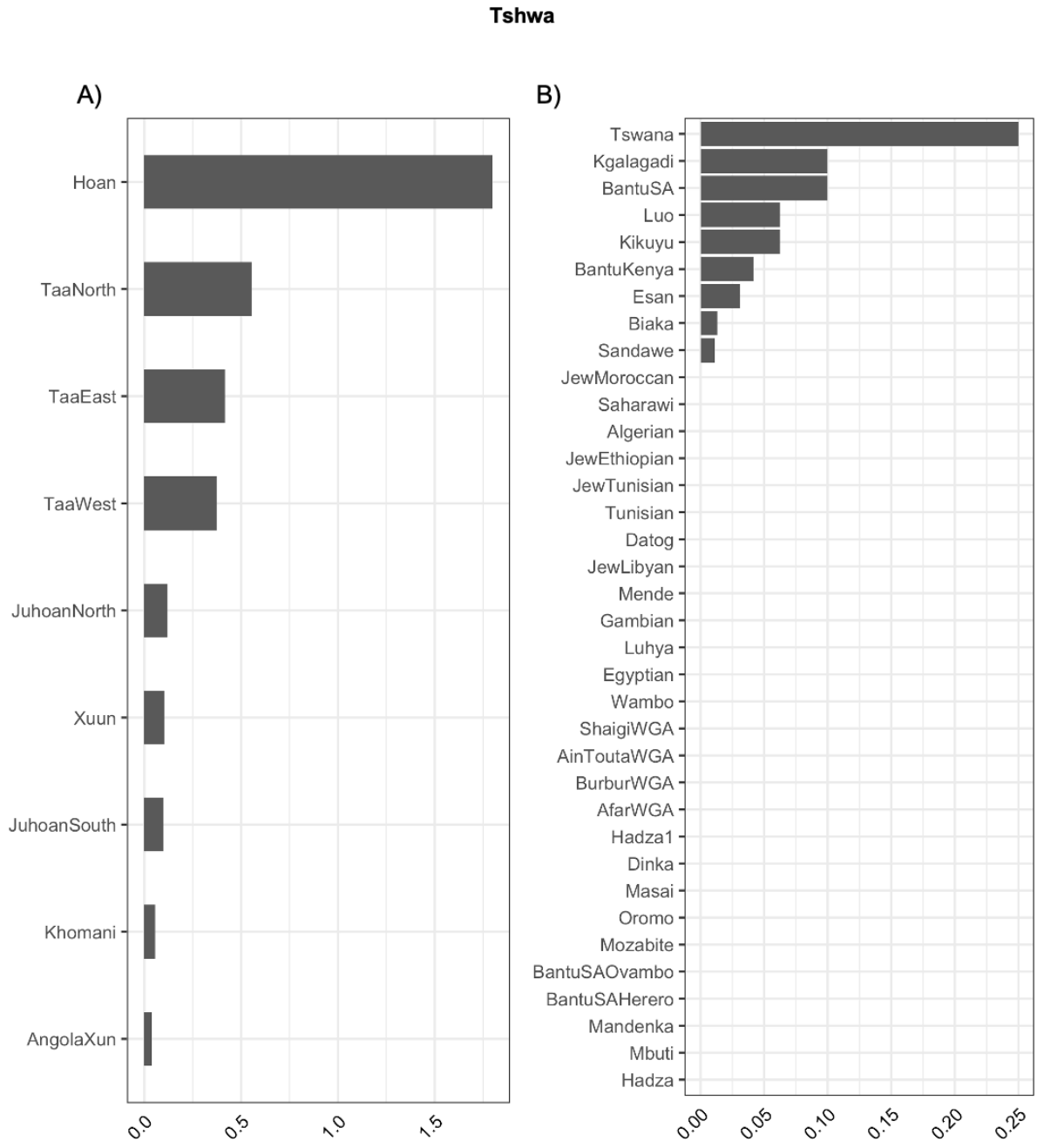

Figure S 8.9 - Number of Shared IBDs (5 to 10 cM) and their respective ancestry assignment between the Tshwa and Khoisan (A) and Non-Khoisan (B) populations.

Table S 5 - Frequencies of ancestry proportions in the populations typed on the Affymetrix Human Origins Array.

Population	west Africa	south Africa	east Africa
BantuSA	0.813	0.167	0.020
BantuSA_Herero	0.928	0.046	0.026
BantuSA_Ovambo	0.971	0.012	0.017
Biaka	0.921	0.064	0.015
Damara	0.940	0.036	0.024
Datog	0.100	0.003	0.897
G_ana	0.325	0.618	0.057
G_ui	0.052	0.883	0.065
Hadza	0.449	0.036	0.516
Hai_om	0.320	0.548	0.132
Himba_Angola	0.921	0.046	0.033
Himba_Namibia	0.949	0.029	0.022
Hoan	0.224	0.735	0.041
Ju_hoan_South	0.023	0.941	0.036
Kgalagadi	0.643	0.337	0.020
Khomani	0.119	0.711	0.170
Khwe	0.540	0.371	0.088
Kuvale	0.937	0.038	0.025
Kwepe	0.845	0.104	0.051
Kwisi	0.832	0.114	0.055
Mbuti	0.730	0.235	0.035
Nama	0.134	0.659	0.207
Naro	0.029	0.916	0.055
Nyaneka	0.967	0.016	0.017
Ovimbundu	0.988	0.002	0.010
Sandawe	0.402	0.023	0.575
Shua	0.523	0.370	0.106
Taa_east	0.144	0.819	0.036
Taa_north	0.056	0.889	0.055
Taa_west	0.119	0.848	0.033
Tjimba	0.880	0.082	0.039
Tshwa	0.338	0.595	0.067
Tswana	0.782	0.194	0.023
Twa	0.857	0.095	0.048
Wambo	0.982	0.008	0.009
Xun	0.156	0.789	0.055
Xuun	0.171	0.772	0.057

Table S 6 - Frequency of ancestry proportions using ancient DNA in the Khoe-Kwadi and Namibe populations typed on the Affymetrix Human Origins Array.

Population	Mende	SouthAfrica_2000BP	Tanzania_Luxmanda_3000BP
Damara	0.943	0.057	0.000
G_ana	0.284	0.630	0.086
G_ui	0.000	0.866	0.134
Hai_om	0.326	0.499	0.174
Himba_Angola	0.868	0.092	0.039
Khwe	0.502	0.372	0.126
Kuvale	0.917	0.083	0
Kwepe	0.778	0.166	0.056
Kwisi	0.741	0.189	0.07
Nama	0.028	0.574	0.398
Naro	0.000	0.887	0.113
Shua	0.458	0.369	0.173
Tjimba	0.793	0.162	0.044
Tshwa	0.279	0.599	0.122
Twa	0.812	0.147	0.042

Table S 7 - Bayesian model comparison results. Models coloured in red represent the chosen model. Comparisons are ordered from best to worst models.

Analysis	Comparison	logML1	logML2	logBF	BF
Phylogeny	relbcov relbct	-4131	-4165	33	67
Phylogeny	strbcov relpdc	-4172	-4172	0	0
Phylogeny	strbct strpdc	-4180	-4189	9	17
Phylogeography	strbcov relbcov	-4500	-4493	-7	15
Phylogeography	strctmc relctmc	-4523	-4508	-15	30