Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

BIODIVERSITÀ ED EVOLUZIONE

Ciclo XXV

Settore Concorsuale di afferenza: 05/B1

Settore Scientifico disciplinare- BIO/08

Comparing genetic and linguistic diversity in African populations with a focus on the Khoisan of southern Africa

Presentata da: Chiara Barbieri

Coordinatore Dottorato

Prof. Barbara Mantovani

Relatore

Prof. Donata Luiselli

Esame finale anno 2012

University of Bologna

Doctoral Thesis

Comparing genetic and linguistic diversity in African populations with a focus on the Khoisan of southern Africa

Author:
Chiara Barbieri

Supervisor:
Prof. Donata Luiselli

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

PhD program in Biodiversity and Evolution
Department of Anthropology

March 2013

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Preface: A day in Dobe

"Chiara, today we are going to visit Dobe! Arent you excited?"

Brigitte Pakendorf, my supervisor at Max Planck Institute, conveys her expectations to me on the incoming day of fieldwork. I know Dobe had been the center of intense ethnographic study in the 70s, on which some books were published: this is how the picture of the pristine society system of the indigenous foragers San was made popular to a broad audience. We will meet the Ju|'hoan community there: will it be the same of the 70s? I dont think it will. But perhaps we will find people who remember the time spent with Richard Lee and how research was carried on for many years. On the other hand, we may find a touristic attraction spot.

Brigitte and I have a kind of a different approach on fieldwork. She had read the ethnographies, she can remember the names of the people and of the villages, and she is trying to cross-check all her notes taken from our colleagues from the KBA (Kalahari Basin Area) project, some of the best experts of Khoisan in the academic world. In a word, she is organized. I would describe my attitude as oriented towards a naïve immersion in the people and in their life, learning day by day on the field (the perfect way to become a source of desperation for your supervisor, after the fifth time that she has to explain you that the !Xuun and the !Xoon are different people). The truth is that, unfortunately, I did not have time to prepare myself too much. I have confused memories of the talks in Berlin with the KBA colleagues, I keep mixing the names and I do not remember that the same population can be called in different terms, I barely recall where the people are supposed to live, and the only thing that I really put an effort into was learning how to pronounce the five clicks, thanks to the patience of

Hirosi Nakagawa and Alan Barnard who taught me the "clicks for dummies". This was how I felt in my very early stages of the PhD: confused, excited, more disorganized than usual, trying to project my life into new responsibilities and panicking about the setting up of weeks of Spartan fieldwork in Botswana. The tent, the Swiss knife, the vaccinations, the equipment, and that young girl face which has to conduct a trip all over the country, persuade village chiefs to join the project, give orders to the driver, Justine, and the translator, Bless, and keep the mood up until the last day of work.

With a shiny sun and a rather hot stream of dry air hitting our faces, this is going to be a typical fieldwork day, but on a special location. As usual, we ask directions about where to go to from the people we meet on the street, but this time our destination appears particularly inaccessible. Suddenly, we find ourselves on a white street that goes straight as a ruler, geometrically drawn until it meets the horizon. That must be the border between Botswana and Namibia. An infinite straight line. We turn around and explore the surrounding a bit better, and finally we park the jeep near a baobab, suspecting that that compound must be Dobe. The huts, a circular wall made with adobe with a conic roof of straws, are deserted. There seems to be nobody around. We walk a bit further and finally meet an old woman, who guides us to her hut. There are children and a young couple. The old woman is happy to see new people and chats a lot, even if Bless has some difficulties in translating straight from Ju|'hoan; she does not speak Tswana, the Bantu language spoken by the majority of people in this country. She shows us some items that surely will pique curiosity in our white faces: a big arrow, some traditional decorations, ostrich eggshell ornaments, and a turtle shell that contains some cosmetic powder used as a face sunscreen. Dobe is rather uninhabited nowadays; people are leaving for less isolated places, for finding better opportunities — the young couple is also very nice and they sit with us chatting in Tswana. At the present life in Dobe is difficult: the surroundings are very arid, as we noticed, and the government aid, usually in the form of a couple of cows, do not reach as far as here. In such conditions, people have to rely on hunting; but hunting is prohibited by some recent laws, and the officers sometimes patrol the area to keep an eye on the Khoisan communities.

Occasionally, when it is necessary to do so, the men go hunting, albeit very cautiously. When they come back to the compound with some prey, usually an antelope, the children are sent inside the huts and hushed to sleep. Meanwhile, the women start to slaughter the animal and prepare the meat together with the men. A fire is set. When the food is ready, some water is sprinkled over the hut — "children, it's raining" — and then the meal is served to everybody. Sometimes the officers get close to the villages and find some children playing around: they target them and start asking questions to check how frequently the adults go out hunting. "Children, when was the last time you ate meat" and they would answer: "the last time we had meat... it was raining!" No, it does not rain that much often in the Kalahari.

Life in the Kalahari can be harsh: there is paucity of water, resources and infrastructure. Life as a Khoisan, or better a Mosarwa (the term that the Khoisan in Botswana commonly use to identify themselves) is definitely tough: in spite of some recent effort from the government to recognize their identity and their critical status, the discrimination against them is tangible, and their conditions are highly unstable. Entire villages have been relocated; their social structure has been rapidly dismantled and rebuilt into something that would better fit the post-colonialist world. Most Khoisan have been somehow integrated as herdsman or laborers for members of other ethnic groups, but the majority of the extant Khoisan communities are affected by social ills such as economic dependency, alcoholism, malnutrition, and societal breakdown, often associated with their abandoning the compounds to join the little urban centers.

My fieldwork experience in Botswana represented an occasion to reflect on the need of anthropology today — or at least to what it means to me, to a certain extent. Ethnography in the third millennia has to face an exceptional reassessment of perspective, especially when the first material of study (pristine societies with a remarkable distinct cultural background) is disappearing. The encounter with distant culture excites the same enthusiasm it exerted for the Victorian explorers. Now however, a component of mutual exchange is intrinsic. In the era of globalization, we are facing people who share the same needs all over the planet, people who want to reach a better standard of living, a better prospective, a better political and economical system. As we converse, I have to realize

that we are in the same conditions: citizens of the world with the respective traditions and cultures, which will be possibly passed to the next generations as a fading hologram. In that moment, I am the vulnerable observer described by the anthropologist Ruth Behar in her essays. Ethnography smoothly shifts towards political and social sciences.

Nevertheless, my conclusions are clearer than ever: the need to study each other and to understand our past and our roots is strong and encompasses the fundamental need of an identity: so many disciplines are involved in tracing this picture, in an intriguing dialogue. The study of the language, of the kin system, of the genetic makeup of a population, the reconstruction of their prehistory, is a fantastic storytelling; every culture is a great story to tell and not to be forgotten.

Acknowledgements

My first sincere gratitude goes to my supervisor at Max Planck Institute for Evolutionary Anthropology, Brigitte Pakendorf, for her dedication, her strength, and her endless care for me over all the big and small research steps. I am thankful for being trusted and guided, and for having had the opportunity to participate in truly exciting projects.

I also want to acknowledge the support I received from my colleagues of the Comparative Population Linguistic group and of the Population History group, and from Mark Stoneking, who also followed all the research projects I participated to, and so many times just had the right suggestion for me when I was stuck in some analysis.

To my supervisors in Bologna University, Donata Luiselli and Davide Pettener, and to the colleagues of the Anthropology department: I want to thank you for the support I felt and for always keeping an eye on me. For being models of open-minded, hardworking scientists who did not lose heart and passion.

My gratitude also goes to Tom Güldemann and to the colleagues of the Kalahari Basin Area project, who introduced me to the Khoisan universe, and with whom I had so many exciting and fruitful discussions and multidisciplinary confrontations.

A special mention to the invaluable help received by the numerous good friends and colleagues who gratuitously revised and edited pieces of this dissertation (especially the sometimes funny English that my Italian-shaped brain still produces): Mimi Arandjelovic, Ana Duggan, Linda Gerlach, Gillian Knight, Marcello Mannino, Eugenie Stapert and Mark Whitten.

Within my friends and colleagues at MPI, I have dedicate a particular thank to Cesare de Filippo, who first took care of me when arrived in the Genetic department, introduced me to the lab protocols and to the world of R. For his stubbornness and his big heart.

And of course, a special mention to MPI people who have been close to me over these years, often helping me in my work, or motivating each other during the hard times: like Natalia Aralova, Falko Berthold, Patricia Heyn, Michael Dannemann, Fabrizio Mafessoni, Maura Pellegrini, Roland Schröder, Mário Vicente, Erin Wessling, those already listed as dissertation reviewers, and so many others.

To all the wonderful people who supported and trusted me for the past four years, from Leipzig and from many kilometers away, to the people who thought that I could eventually make it: the respect and gratefulness I have towards you just cannot be reduced in a few words here. You are great.

This dissertation is dedicated to Angela and Tonino, my parents.

Contents

List of Figures

List of Tables

Abbreviations

aDNA	**a**ncient **DNA**
AMH	**A**natomically **M**odern **H**umans
AMOVA	**A**nalysis of **MO**lecular **VA**riance
bp	**b**ase **p**airs
BSP	**B**ayesian **S**kyline **P**lot
CKGR	**C**entral **K**alahari **G**ame **R**eserve
DNA	**D**eoxyribo**n**ucleic **A**cid
HLA	**H**istocompatibility **L**eucocyte **A**ntigene
HVS1, HVS2	**H**yper **V**ariable **S**egment **1**, Hyper Variable Segment 2
IPA	**I**nternational **P**honetic **A**lphabet
kya	**k**ilo **y**ears **a**go
LD	**L**inkage **D**isequilibrium
LGM	**L**ast **G**lacial **M**aximum
LSA	**L**ate **S**tone **A**ge
MDS	**M**ulti **D**imensional **S**caling
mtDNA	**m**itochondrial **DNA**
mya	**m**illion **y**ears **a**go
$\mathbf{N}_e$	effective population size
NRY	**N**on-recombining **R**egion of the **Y** chromosome
PCR	**P**olymerase **C**hain **R**eaction
POA	**P**lace **o**f **A**rticulation
SAK	**S**outh **A**frican **K**hoisan
SNP	**S**ingle **N**ucleotide **P**olymorphism
STR	**S**hort **T**andem **R**epeat
TMRCA	**T**ime of the **M**ost **R**ecent **C**ommon **A**ncestor

Chapter 1

INTRODUCTION

In the field of anthropology, the commonly evoked multidisciplinary approach is often successfully applied to resolve a disparate list of research questions (Renfrew 2010). In fact, the study of humankind and of its biological and cultural features represents an ideal case where many disciplines converge on a common ground; i.e. the investigation of our origin and history. This dissertation embodies the interaction between disciplines in the study of human population history, being founded in particular on the dialogue between linguistics and genetics.

The genetic variability found in extant populations goes far beyond the simple description of demographic patterns, revealing much more of the processes that affected the evolution of the current genetic make-up. With the help of uniparental genetic markers, the Y chromosome and mitochondrial DNA, which are non-recombinant and characterize a genealogy traceable in time and space, we are able to describe cases of contact between populations, sex-biased gene flow, population expansions and migrations, the amount of genetic structure and also a wide range of social contexts and demographic patterns (Underhill and Kivisild 2007).

In this dissertation, linguistic data is often incorporated into the study and compared with the genetic data. Furthermore, linguistics plays a role in designing the study, contextualizing the research questions, providing evidence of population interaction and ultimately in representing the neutral criteria chosen to

define human populations: the latter represents a delicate point for anthropological research, given that human variability cannot be simply summarized with univocal labeling. For this reason in molecular anthropological studies the unit of investigation is often identified by the research question itself (see Pakendorf et al. 2011).

The association between the study of genes and languages, already suggested by Charles Darwin himself in the Origin of Species, has been proved to represent a particularly fruitful intuition: the comparison of genetic and linguistic data can elucidate and complement both human population prehistory and the dynamics underlying language evolution (Cavalli-Sforza 2001). Chapter 2 illustrates the parallel between language evolution and population history, giving a broad overview of how the study of languages allows us to reconstruct phylogenies, calculate linguistic distances and evaluate the intensity of contact between groups. Examples of the linguistic data employed and the kind of analyses developed are also provided.

In Chapter 3, the physical context of the current investigation is presented: the African continent. Following the linguistic approach chosen for this anthropological investigation, the chapter illustrates the diversity within the African populations as inferred from the diversity of the languages spoken. The four major African linguistic phyla are briefly introduced and collocated in time and space. Special attention is given to the Bantu family (and, consequently, to the Bantu-speaking people) of the Niger-Congo phylum: its wide diffusion across much of the continent is associated with a major human migration and the spread of agriculture (Bostoen 2007, Diamond and Bellwood 2003). The Khoisan languages are also described in detail, with a focus on their linguistic peculiarity: the so-called "click" sounds.

Chapter 4 provides a deeper characterization of the Khoisan populations of southern Africa that occupy areas of Botswana and Namibia nowadays; currently extinct groups of South Africa and Angola are also mentioned. We focus on the variability within the Khoisan, in terms of the languages they speak, their modes of subsistence (foragers or pastoralists), and their different physical appearances. The typical foragers of the Kalahari Basin are distinguished by the

Khoe speakers, who were probably bringing a pastoralist economy (Güldemann 2008a). Different origins and demographic dynamics are proposed for the various groups.

Finally, Chapter 5 explores the core of this dissertation, the genetic variability of African populations. This chapter reviews the extant knowledge of genetic variation and human prehistory in the continent, starting from the importance given to Africa as the place of origin of modern humans, and following with the description of major findings coming from autosomal and uniparental markers. A high level of population structure is inferred for prehistoric times and described for modern times: the current structure is the result of population migrations and contact, and reflects the linguistic affiliations of the populations (Tishkoff et al. 2009). Lastly, the genetic data available for Khoisan populations is presented and contextualized.

In this dissertation, I present four papers that explore population prehistories on the background of sub-Saharan Africa at different geographic resolutions and time scales. All four papers share the same research approach: incorporating a linguistic background and linguistic data (when available) and testing hypotheses on population prehistory built on linguistic, archeological and cultural evidence, with the help of uniparental genetic makers.

In Paper I (6, de Filippo et al. 2011), genetic analyses are performed at a continental scale: for this perspective, the four linguistic phyla are considered in grouping the populations. In this paper, we analyze the paternal component alone: Y chromosomal haplogroup affiliation and STR profiles are typed in 1,195 individuals from areas that were previously poorly characterized. The new genetic data is included in a wider dataset retrieved from available Y chromosome studies, providing a good coverage of the continent. Linguistic affiliation is shown to be a good proxy for the genetic structure uncovered in this paper. A major finding is the identification of a haplogroup marker that is present at high frequencies in Bantu-speaking populations but absent in other Niger-Congo speakers, providing further evidence of the rapid demographic spread of Bantu-speaking people through the continent.

In Paper II (7, Barbieri et al. 2012a), we analyze the contact between Bantu-speaking immigrants and Khoisan residents from both the maternal and the paternal genetic components. The geographic setting is southwest Zambia, where only Bantu languages are currently spoken; some of these languages have click sounds, which are characteristic of Khoisan languages. With full mtDNA genomes and fine-scaled Y chromosome typing, we show how the Bantu groups with clicks likely incorporated the clicks after contact with Khoisan populations, which resulted in a sex-biased gene flow. The Khoisan-specific maternal lineages retrieved in Bantu-speakers are very divergent from those commonly found in extant Khoisan: this suggests that at the time of contact, Bantu speakers met Khoisan populations that were genetically different from the Khoisan who survived until the present time.

In Paper III (8, Barbieri et al. 2013) we analyze 500 mtDNA genomes of the deepest rooting clades of the phylogenies, commonly found in Khoisan populations. This dataset expands the previous knowledge about this ancestral clade by more than 10-fold. We are therefore able to redefine the phylogeny, pushing back in time some nodes, and discovering new deeply divergent branches present exclusively in Bantu-speaking populations. We test the probability of retaining these divergent lineages for different population sizes and conclude that deep ancestral substructure must have been reduced and diluted in modern populations.

In Paper IV (9, Barbieri et al. in preparation), our dataset of southern African Khoisan populations is fully analyzed for the purposes of reconstructing their prehistory at a fine scale. With mtDNA sequences for 700 individuals of 26 Khoisan- and Bantu-speaking populations, we are able to describe a different maternal profile for prehistoric foragers, pastoralist pre-Bantu immigrants, and Bantu-speaking agriculturalists: a complex pattern of woven histories is emerging for previously understudied indigenous populations. Our major findings include 1) A revision of the split between northwest and southeast Kalahari foragers, proposed by autosomal data in a parallel study and tested here for mtDNA with simulations: the effect of recent contact through the Kalahari Basin blurred this initial divergence, and is confirmed in the sharing of linguistic features between different families; 2) The high structure within Khoisan populations, which is explained by socio-cultural factors like small, semi-nomadic bands and uxorilocal

post-marital residence patterns; 3) The discovery of a signal of expansion in a specific lineage, which could be linked to the Khoe-speaking pastoralist migration from East Africa, and to a strong gene flow in the ancestors of Bantu-speaking populations of Namibia.

Chapter 2

LINGUISTICS

2.1 Evolution of languages and population history

The evolution of human languages can be paralleled to the evolution of human populations as both are subjected to similar patterns of transmission of traits, in cultural and biological terms. Languages evolve along with the human populations and their development is influenced by the same demographic changes. The first intuition about this parallel can be found in the Origin of Species (1859: 422-423), when Darwin suggests that biological and linguistic data could describe similar genealogies.

However, the parallel between cultural and biological evolution can only provide a fruitful area of investigation if we also understand the main points of incongruence between the two processes, the most obvious one being the modality of transmission: while biological evolution proceeds mainly via vertical transmission of the genetic material, cultural evolution occurs also through horizontal diffusion of information or behavior. Languages tend to diffuse to their immediate environment, which may lead to the spread of certain linguistic features across that region, generating a so-called areal effect. Holman et al. (2007) evaluate the dissimilarities between related and unrelated languages controlling for geographical distance, and they found that there is a positive correlation between dissimilarity and distance, that nevertheless does not affect the major signal coming from language relatedness.

The interest in recognizing the causes of language change and the course of language history overcomes the mere reconstruction of phylogenies for related languages, and opens up for application to broader anthropological research questions. Since the past decade, there has been an increasing interest in the investigation of 1) how languages are related, and 2) how, where and when they originated; for these studies, scholars started examining quantitative data with the support of appropriate statistical methods and validating them with the help of computational simulation models, thus defining a new field that might be designated as "language dynamics" (Wichmann 2008). The methods of this qualitative analysis rely on stable language features that can be compared between languages to measure the level of proximity, an approach that is more consistent when the comparison is performed within the same language family. Languages can be classified in families according to the presence of genetic relationships, i.e. features that are retained as the result of genealogical inheritance, applying the comparative method. Thomason and Kaufman (1988 :9-12) provide a list of the theoretical assumptions behind the concept of genetic relationship, which must be taken into consideration when paralleling language change and population dynamics. First, it must be accepted that languages can evolve: they can change through time as a result of result of drift (due to structural imbalance), dialect interference and foreign (external) interference. A language can become fragmented through a geographic area, which may lead to increased internal diversity, and eventually to the development of different languages (whereby it needs to be kept in mind that such 'language splits' is far from categorical). Second, language change can occur at any level of the linguistic system, e.g. in the phonological, morphological or syntactic domain. Third, vertical transmission of language in one-generation step is accompanied by relatively small changes in normal social context; drastic changes happening in particular contexts are typically associated with language shift. Fourth, genetic relationships lose strength when transmission is imperfect, for example after a language shift where language structures from the substrate are interfering, or in cases of bilingualism. From this assumption comes the fifth and last one: a language cannot have multiple ancestors in the course of normal transmission.

2.2 Which linguistic data?

After defining the rules of language relatedness, attention should be paid to the raw data that the linguist choses to analyze. Stable features within languages that could prove vertical relatedness or horizontal contact, can be brought back to two widely used categories: lexical features and structural features. Morris Swadesh developed the concept of lexicostatistics from the idea that language relatedness can be measured quantitatively on a lexical base, employing a list of basic concepts that are present in every language (called basic vocabulary) and are less subject to change (Swadesh 1950): this list of stable words includes among others pronouns, numerals, body parts, geographical features, and was coded in the so called "Swadesh list", which most common version includes a hundred items (Swadesh 1971). This vocabulary is compared across languages in search for cognate words, i.e. words that have the same etymological origin. According to Swadesh, the proportion of meanings that are cognate is proportional to the degree of relatedness, or to the linguistic distance. However, recent borrowings can mask the real degree of relatedness between two languages, as well as chance similarities: an example of the latter is the Modern Greek and Maori word for "eye", mata and mati, which are apparently similar but surely do not reflect any historical connection (Gray 2005).

Another application of the Swadesh list is glottochronology, which measures the time depth of linguistic divergences based on the assumption of a constant rate of vocabulary loss for all languages. These methods have been widely used for analysis of various linguistic families, and a representative example is the work of Isidore Dyen, who created ad hoc word lists to classify Austronesian languages (Dyen 1965) and subsequently Indo-European languages (Dyen, Kruskal and Black 1992). A major limitation of this method concerns the time depth that we can reach, since most linguists agree that language families cannot be traced back after an estimated age of 10 kya: beyond that time limit, there will be no detectable similarities between pairs of languages (Gray 2005). In a more simple perspective, languages evolve faster than genes (Cavalli-Sforza et al. 1994), and language phylogenies coalesce to the common root (proto-language) within a narrow prehistorical depth.

In contrast to lexical features, structural features of languages are claimed to provide connections between languages on a deeper time depth. Linguistic typology aims at finding those deep connections analyzing language universals, i.e features that should be present in all human languages, and it is defined as "the classification of languages or components of languages based on shared formal characteristics" (Whaley 1997: 7). A more practical definition of the field comes from Johanna Nichols, for whom typology engages in "developing framework-neutral grammatical theory and applying that theory to crosslinguistic distribution and their implications" (Nichols 2007: 235). Language comparisons based on typology has certainly succeeded in addressing questions about broad language distribution and ancient linguistic history (cf. Dunn et al. 2005), because of their codable definitions that are suitable for computational analysis and because they employ stable features that are resistant to change and borrowing (Nichols 2007).

Other research in this field focuses on the stability of language features. The aim is to identify features that change at a slower rate and to use them to establish deep linguistic connections; for example, a highly stable feature is the subject-verb-object word order (Nichols 2008). Another practical example connects population structure to the rate of language evolution: Nettle (1999) shows with computer simulations that languages change faster in smaller populations. A limitation of this approach resides in the possibility of confusing genealogical features with the effect of horizontal diffusion, especially for distantly related languages. Finally, Wichmann and Saunders (2007) insist on a combination of both typological and lexical features for extending the time depth at which we can reliably investigate phylogenies; Wichmann and Holman (2009) provide further evidence of this possible parallel, calculating a similar retention rate for stable typological traits and for the core vocabulary of the Swadesh list — therefore validating Swadesh glottochronology itself.

2.3 The effect of population contact in linguistic and genetic data

Mismatches between information coming from the biological evolution of populations and the linguistic data should not be ignored: they can be informative of historical processes that brought a change of the cultural profile. A typical example is language shift, which can be detected when two populations speak the same language but have different genetic profiles, or conversely when two populations speak different languages but have a similar genetic profile; language shift in Khoisan populations is discussed in Paper IV (9). An example of the first scenario is the case of Pygmies of Central Africa, who adopted the language of the Bantu-speaking colonizers but present very distinct genetic features. In fact, Pygmies harbor exclusive lineages deeply rooted in the human genealogy, which is in line with an early divergence of these populations (Patin et al., Schlebusch et al. 2012, see also Paper I (6) for the Y chromosome paternal relationships between Pygmies and the rest of Africa). There is no trace left of the original languages spoken by Pygmy populations before the shift to Bantu: it remains mysterious. The second case, which involves the permanence of a language (and possibly of a cultural identity) in a community which is genetically homogeneous with their neighbors who speak different languages, is best exemplified by the Hungarians: their language belongs to the Finno-Ugric family, a pre-Indoeuropean family, but they are genetically similar to other European populations (Tömöry et al. 2007).

More detail about the degree of contact and exchange between populations can be obtained from comparison of languages and genetic data on a finer scale: direct exchange can be quantified by the degree of exchange of linguistic features (phonological, morpho-syntactic or lexical features) and of identical or similar genetic motifs (gene-flow). In Paper II (7), the degree of contact between Khoisan and Bantu-speakers is measured by the amount Khoisan specific haplogroups detected in Bantu groups, for the maternal or paternal line. Linguistically, most of these Bantu-speaking populations use click sounds, which are characteristic of Khoisan languages. The match with Khoisan gene flow and presence of clicks confirms the presence of ancestral contact between the two groups.

In conclusion, the field of linguistics provides methodological approaches that are suitable to investigate population dynamics and can be easily paralleled to genetic analyses. Population prehistory can be inferred through the reconstruction of language phylogenies, the estimation of linguistic distances, and the evaluation of cases of contact and exchange. The integration of information from archeology, linguistics and genetics together with cultural anthropology in a broad sense is the ultimate goal of the multidisciplinary approach, which is employed to achieve a better understanding of our diversity and prehistory (Renfrew 2010, Blench and Spriggs 1999).

Chapter 3

AFRICAN LANGUAGES

The African continent harbors a great linguistic diversity, in line with the overall high cultural variability characteristic of the territory occupied by modern humans for the longest time on earth. More than 2000 languages are spoken, roughly one third of all the languages of the world (Heine and Nurse 2000), although this cannot be considered a definitive number, as new languages are being discovered, others are on the verge of extinction, and there is always debate in terms of what defines a language versus a dialect. The 'autochthonous' African languages, i.e. excluding those introduced in historical time (such as English, Portuguese, French, Afrikaans, Arabic, etc.) are grouped into four main clusters, phylogenetically separated and named as 'phyla': these are Afroasiatic, Nilo-Saharan, Niger-Congo, and Khoisan (see Greenberg 1964). Their distribution is shown in Figure 3.1.

An important caveat when studying the variation and the diffusion of the different African languages is that in Africa, multilingualism is almost a norm (Childs 2003) which makes it difficult to link the ethnolinguistic affiliation of the various populations. Paper I (6) investigates the genetic variation within a subset of sub-Saharan African populations, focusing on the linguistic diversity between and within phyla: samples of populations speaking languages from all the four African phyla are compared, with a specific interest on the Niger-Congo phylum.

FIGURE 3.1: Map showing the distribution of the four African language phyla and major languages (modified from Heine and Nurse 2000)

3.1 Afroasiatic

Afroasiatic is the least controversial of the four phyla: the common origin of this language group was determined even before Greenberg's classification, and collocates over 10 kya according to Hayward (in Heine and Nurse 2000). Six major branches compose this phylum, namely Chadic, Berber, Egyptian, Semitic, Cushitic and Omotic. Regarding their geographic distribution, these languages are mainly present in the north of the continent and in the Horn of Africa (Figure 3.2). The most ancient families (Omotic and Cushitic) are located in the Horn of Africa, in agreement with an eastern African origin of the phylum, probably in pre-Neolithic times (Ehret 1979). A second hypothesis links the origin and dispersal of Afroasiatics with the spread of agriculture and cattle rising from the Levant, after the Neolithic revolution (Diamond and Bellwood 2003).

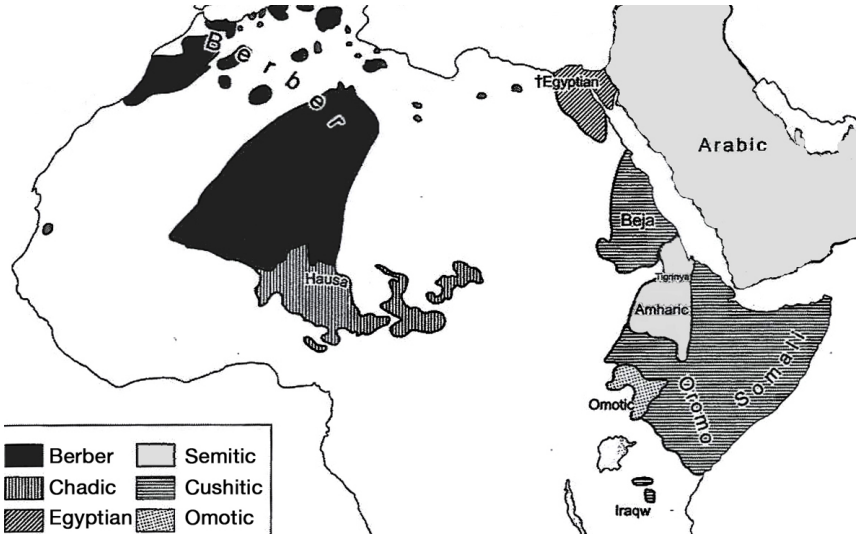


FIGURE 3.2: Distribution of Afroasiatic linguistic families (modified from Heine and Nurse 2000)

3.2 Nilo-Saharan

Nilo-Saharan represents the most controversial phylum: its unity is in fact very debated, with many scholars considering it mostly a collection of unrelated groups (Bender in Heine and Nurse 2000). Some scholars suggested that Nilo-Saharan should be merged with the Niger-Congo phylum, with the former being the earlier branch splitting from the phylogenetic classification (Gregersen 1972), based on morphological (non-lexical) evidence (Blench 1995). The difficulties in classifying this phylum have been evident since Greenberg's time and are highlighted in Ruhlen (1987): for him these difficulties consists in a lack of descriptive literature coming from a late contact with Europeans, in the presence of relatively fewer languages than Afroasiatic or Niger-Congo, and in the high internal heterogeneity and presence of small isolated groups. Bender (1996) describes it as the less known of the four African phyla. In the classification, the most studied and well documented group is the Nilotic family, which includes languages spoken in Sudan, Kenya and Tanzania by predominantly pastoralist populations, such as the Masaai, Turkana, Dinka, Luo and Datooga. The geographic distribution of the whole phylum is scattered over at least 15 countries of central/north Africa and is often discontinuous (Figure 3.3). The greatest diversity is found in Chad, Sudan and Ethiopia.

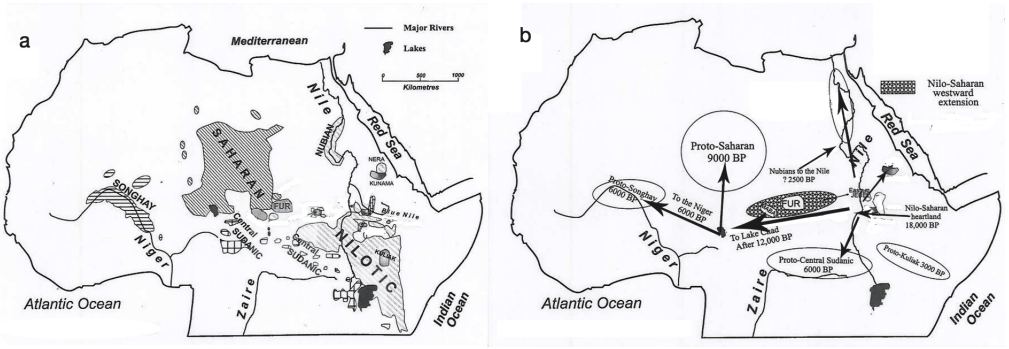


FIGURE 3.3: Distribution of Nilo-Saharan languages (a) and routes of dispersal (b) (Modified from Blench 2006)

3.3 Niger-Congo

Niger-Congo is the largest phylum in the world in terms of the number of languages it contains, and occupies most of sub-Saharan Africa (Williamson and Blench in Heine and Nurse 2000; see Figure 3.4). In the literature it is also referred to as Niger-Kordofanian, from Greenberg's classification (1963), but the term is now considered obsolete (Williamson 1989). A predominant linguistic feature of the phylum is its complex noun classification system. Within the phylum, there is a general agreement on an early split of three major branches. Kordofanian is the first split: a small group of poorly documented languages on the verge of extinction, and whose connection with the rest of the phylum is the weakest. The second split is Mande, a larger group of languages of West Africa whose affiliation to the phylum is also debated by some authors (Mukarovsky 1966, Dimmendaal 2008). This split is followed by the rest of the families: the Atlantic, the small families Dogon and Ijoid, and the Volta-Congo branch, which comprise most of the languages of the phylum and for which there is the strongest unity in terms of shared linguistic features (Williamson and Blench in Heine and Nurse 2000). The most widespread of the Niger-Congo linguistic families is Bantu, a family of the Benue-Congo branch.

The Niger-Congo phylum possibly originated in Central-West African some 15 kya (Ehret 2000: 293) and spread in concomitance with Holocene climate change determined from paleoclimatic data 10 kya (Dimmendaal 2008). This timing also coincides with cultural and technical innovations like the bow and arrow,

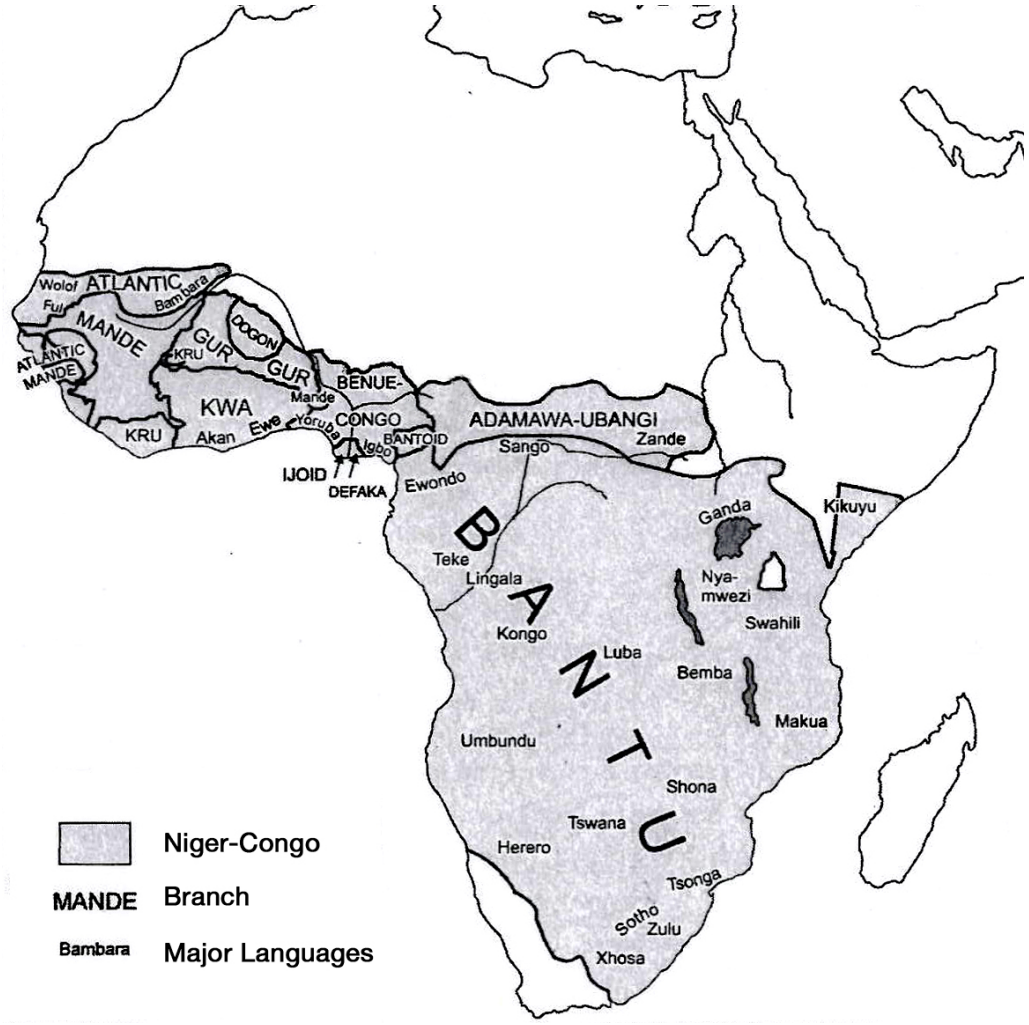


FIGURE 3.4: Distribution of Niger-Congo languages (from Heine and Nurse 2000)

poisoned arrow tips, and dog domestication (Blench 2006), all improvements that suggest the presence of a hunting-based economy.

3.3.1 Bantu Family

3.3.1.1 Classification

Bantu languages are the most diffused in sub-Saharan Africa, despite being only a small sub-group of the Bantoid branch of the Benue-Congo node, within the Niger-Congo classification (Williamson and Blench in Heine and Nurse 2000). The number of Bantu languages varies between 440 (Guthrie 1971) and 680 (Mann et al. 1987). The idea of a Bantu entity was first introduced by Bleek

(1862), while the unity of Bantu languages was unanimously established in early classifications (see Doke and Cole 1961). The boundaries between Bantu in the strictest sense and the broad Bantoid linguistic family are still a matter of debate: the two groups form a linguistic continuum (Blench 2006). Most linguists refer to the Bantu in the strict sense corresponding to the classification from Guthrie (Narrow Bantu), which includes only languages south and east of Cameroon (Maho 2003). Guthrie's classification applied a geographic subdivision over the distribution of the Bantu languages, and named the resulting zones with an alphanumerical code (Guthrie 1948; see Figure 3.5).

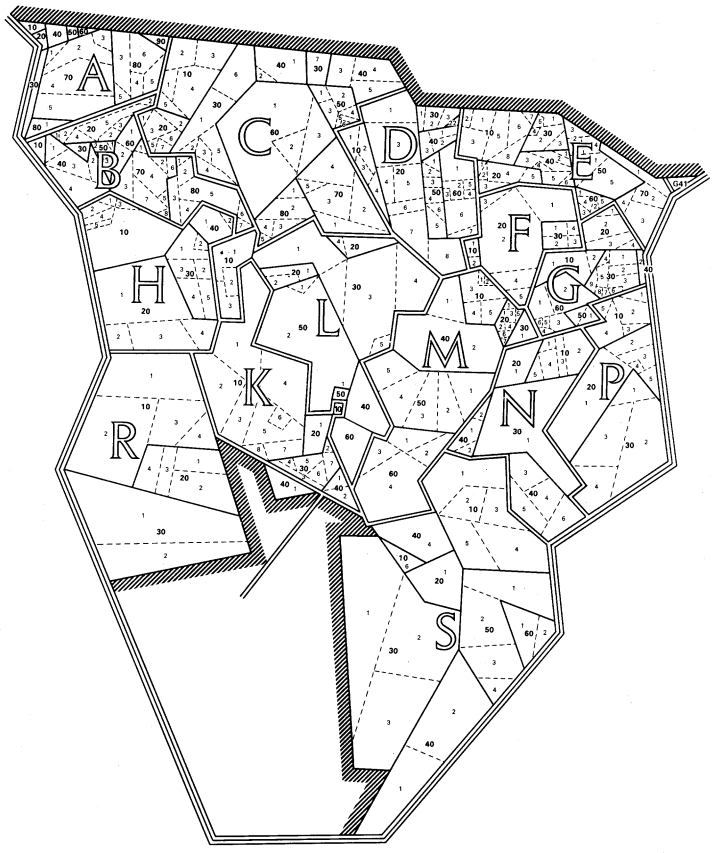


FIGURE 3.5: Map of Bantu Zones according to Guthrie's classification (Guthrie 1967)

While an internal genealogy of the languages cannot be reconstructed unanimously, two main branches have been proposed, namely West and East Bantu, but clear evidence for this split is lacking (Nurse and Philippson 2003). Nevertheless, there is more diversity in the Western branch than in the Eastern one, which forms more of a coherent unit (Nurse and Philippson 2003). The reconstruction

of how the West-East split happened currently remains unsolved, though two main hypotheses have been proposed: the first describes the evolution of Bantu with an early split of the Eastern Branch, which then separated from Cameroon and dispersed to the East following the northern border of the equatorial forest (Bastin et al. 2009, Holden 2002); the second suggests a later split of the Eastern branch, which would have then separated from the Western branch at a later stage and dispersed to the East along the south of the equatorial forest (Ehret 2001, Rexova et al. 2006; see Figure 3.6).

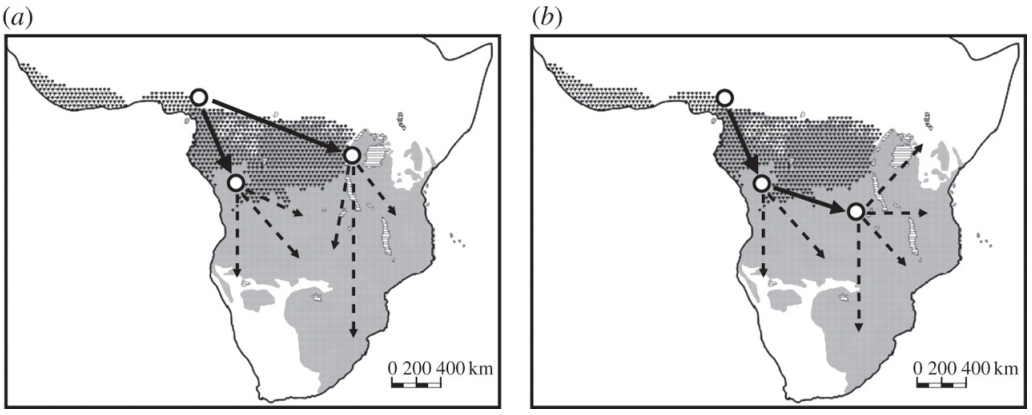


FIGURE 3.6: Routes of the Bantu diffusion, according to the (a) early split or (b) late split. Dark shading corresponds to the extent of the rainforest (from de Filippo et al. 2012)

3.3.1.2 Origin and diffusion

The Bantu language family itself is relatively young and is estimated to have originated between 4 kya (Blench 2001) and 5 kya (Vansina 1995, on glottochronology bases), in the Grassfields of Cameron, where the highest linguistic diversity is currently described. The collocation in time of the Bantu origin correlates with the spread of Neolithic technologies such as macrolithic tools, polishing, and pottery (Bostoen 2007) and with the diffusion of agriculture and a sedentary lifestyle (Diamond and Bellwood 2003). The relatively shallow time from the origin of Bantu languages and the beginning of the Neolithic cultural spread to their current widespread distribution (up to the eastern areas of Kenya and southern areas of South Africa) raises questions on how this fast spread could occur (Eggert 2005, Bostoen 2007). The most accepted hypothesis is that human migration alongside farming culture underlies the spread of Bantu languages, but

some scholars support a cultural diffusion scenario instead of an actual movement of people who replaced the previous inhabitants, suggesting that a massive language shift phenomena would have fostered this diffusion (Nichols 1997). Molecular anthropology approaches have investigated the demographic profile of the Bantu migration in several studies. The majority of studies support this massive demic movement of people, from uniparental (Salas et al. 2002, Wood et al. 2005, Gignoux et al. 2011, de Filippo et al. 2012) and autosomal (Tishkoff et al. 2009, Henn et al. 2011) markers. The West-East separation was also addressed, suggesting a complicated system of migrations and major support for the late Eastern split (de Filippo et al. 2012, Pour et al. 2012). In Paper I (6), we show how fine-resolution Y chromosome analysis reveal a substantial homogeneity for the Bantu-speaking populations, who are differentiated from other representatives of the Niger-Congo phylum, in support of a rapid demographic spread.

3.4 Khoisan Languages

Khoisan is the smallest of the African phyla, with approximately 30 languages still spoken today, and with a large number of languages currently on the verge of the extinction (Güldemann and Vossen in Heine and Nurse 2000). The first reports of Khoisan languages were performed superficially, for the difficulties in reaching a deep comprehension of these very particular languages. The most prominent feature of Khoisan languages is the extensive usage of so-called click sounds as part of the phoneme inventory. The presence of clicks, among other mainly phonological shared features, shared between the majority of Khoisan languages was the main force driving the classification of Khoisan languages into a single phylum.

3.4.1 Click sounds

Clicks are commonly found in some languages with paralinguistic functions (Gil 2002), emphasizing the communication of feelings such as disapproval, irritation, exasperation and regret. In English, for example, they are used as interjections,

POA	phonetic (IPA)
bilabial	ʘ
dental	|
lateral	‖
alveolar	!
palatal	ǂ

TABLE 3.1: Click influxes and respective symbols used in IPA

while in many Italian regions a simple dental click means "no". Apart from the nowadays extinct ritual language Damin spoken in Australia, the only languages that use clicks as phonemic speech sounds (i.e. as "normal" consonants), are the Khoisan languages alongside some southern African Bantu languages. It is likely that these Bantu languages acquired clicks through contact with the neighboring Khoisan languages (Barnard 1992). Paper III (8) analyses the degree of contact between Khoisan and Bantu-speakers from southeast Zambia which presumably led to the acquisition of clicks in some of the Bantu languages spoken in the area. The Southern African Khoisan (SAK) languages have the highest proportion of click words in the lexicon (50% and more) and the highest functional load (Güldemann 2007), in comparison to other languages with clicks.

The production of clicks involves two closures in the mouth, one in the front (called influx) and one in the back (so-called accompaniment). This creates a pocket of air between the two closures; lowering the tongue while keeping the closures creates low-pressure in the cavity. If then the front closure is released the are rushes into the mouths and creates the characteristic click sound (cf. Ladefoged and Maddieson 1996).

Concerning the orthography of clicks, the symbols that are nowadays part of the International Phonetic Alphabet (IPA) were originally developed by Bleek in 1911 and subsequently standardized by Ladefoged & Traill (1984). The front closure of clicks (the influx) can be produced at five different places in the mouth (also called place of articulation, POA). Table 3.1 shows the five click influxes and the respective symbols used in IPA.

The five clicks can be described as follows. i) ʘ: bilabial. Produced by releasing air through the lips, like in giving a kiss. It is the only click that is found only in a subset of Khoisan languages; ii) |: dental. Produced in a sucking motion

with the tip of the tongue pressing on the incisive teeth (transcribed as "c" in the Bantu system); iii) ǂ: palatal. Produced by pulling the tongue quickly away from the front of the palate; iv) !: alveolar. Produced by pulling the blade of the tongue sharply away from the alveolar ridge (transcribed as "q" in the Bantu system); v) ǁ: lateral. Produced by releasing air with the tongue adjacent to the teeth on the lateral side of the mouth (transcribed as "x" in the Bantu system) (Barnard 1992).

3.4.2 Classification

It is difficult to find evidence of shared cognate or reconstructed proto-forms for this phylum, maybe because of the extremely ancient origin, or more simply because such a genealogical relationship does not subsist (Traill 1986). In more recent times, the existence of Khoisan as a linguistic phylum has been doubted or even rejected by most linguists working on Khoisan languages (Güldemann & Vossen 2000:101): in fact, the similarities between the main branches of Khoisan may be due to areal contact and exchange.

Today, the term Khoisan subsumes three different and genealogically unrelated families: Kx'a, Khoe-Kwadi, and Tuu. Figure 3.7 reports an updated classification of the Khoisan languages (Güldemann, forthcoming). Kx'a (Heine & Honken 2010, previously labeled Ju-ǂHoan) includes what was formerly called Northern Khoisan plus the geographically isolated language ǂHoan; Tuu (Güldemann 2005) corresponds to what was formerly called Southern Khoisan; and Khoe-Kwadi (Güldemann 2004; Güldemann & Elderkin 2010) includes what was formerly called Central Khoisan, or Khoe, plus the extinct Kwadi language of Angola (Barnard 1992, Güldemann 2008a). These three families are located in southern Africa, and are referred to as Southern African Khoisan (SAK in literature), to distinguish them from the two languages spoken in eastern Africa (Tanzania): Hadza and Sandawe. These two click languages are spoken by a few communities that used to lead a foraging way of subsistence; they were originally included in Greenberg's Khoisan phylum. Hadza appears to be unrelated to all

the other Khoisan languages, (i.e. a clear case of a language isolate). Furthermore, Hadza people experienced a severe demographic bottleneck which is detectable also in their genetic makeup (Güldemann and Stoneking 2008, Lachance et al. 2012). Sandawe, on the other hand, shows some closer linguistic relationship to SAK and is most likely genealogically related to Khoe-Kwadi (Güldemann and Elderkin, 2010).

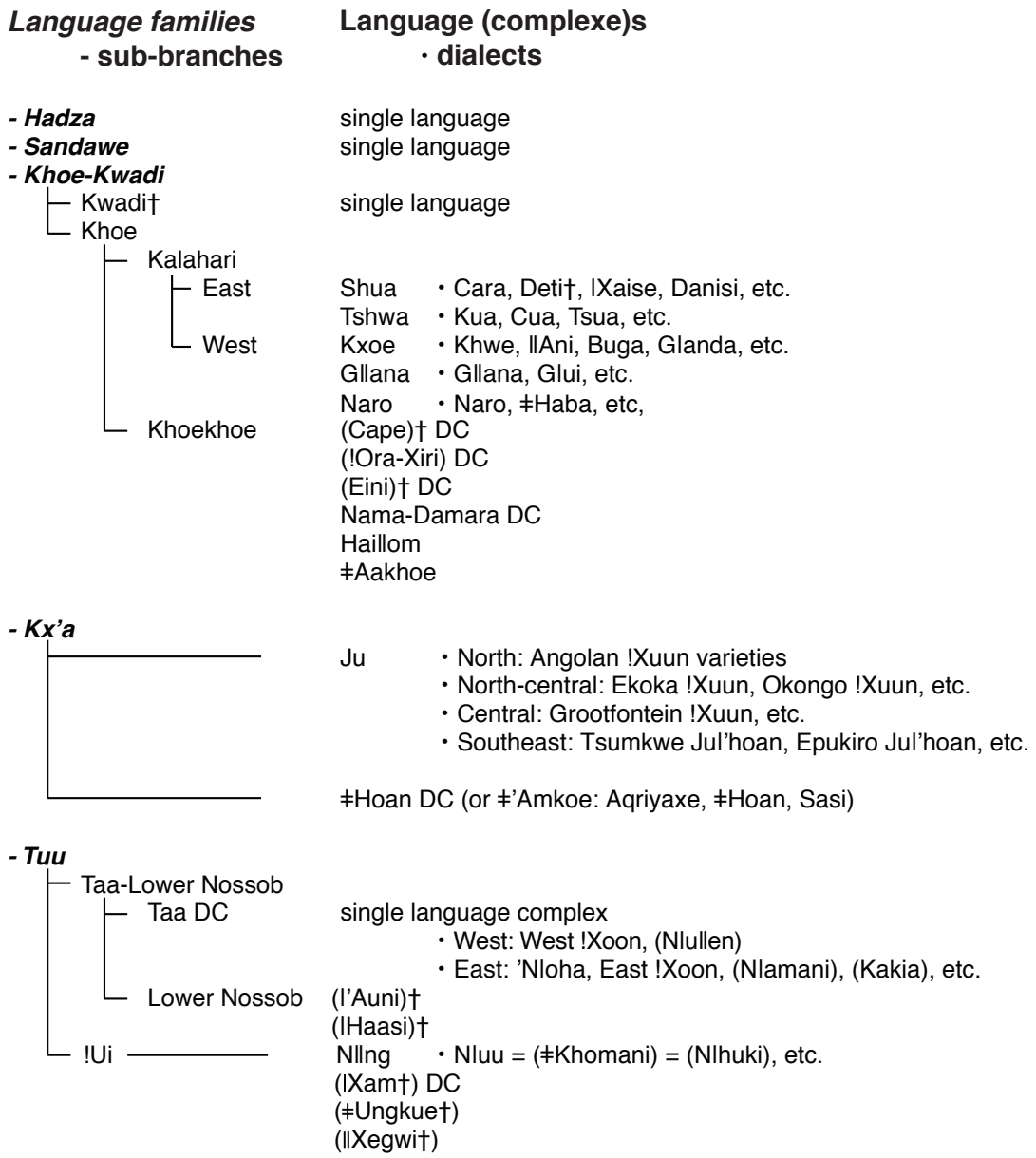


FIGURE 3.7: Lineages subsumed under Khoisan and internal composition. DC: dialect cluster, †: extinct, (Only older sources). (adapted from Güldemann, forthcoming)

3.4.3 The three SAK linguistic families: origin and distribution

Kx'a and Tuu share some linguistic features as well as lexical items. The relation between the two families is in line with a strong effect of contact. A possible historical common origin of the two language families was hypothesized, but can so far not be proved. Their distribution nowadays covers the center of the Kalahari Basin and its immediate surrounding. The Ju language family (Kx'a) is localized slightly in the North, with !Xuun speakers to the West in Namibia and up to Angola. The presence of !Xuun speakers in South Africa (Schmidtsdrift military camp) is due to the recruitment of people from Angola. The origin of the ǂHoan, which language was affiliated to the Ju family only in recent time, is still debated: possibly the high rate of contact and borrowing from neighboring Tuu and Khoe-Kwadi languages obscures a clear interpretation of its genealogical position (Gerlach, in preparation).

The Tuu linguistic family was historically distributed over a wider territory covering most of South Africa as well as parts of Botswana and Namibia, but in South Africa the territory has been drastically reduced, with cultural assimilation and language shift of many former Khoisan speakers. Nowadays, Taa is the largest Tuu language, which consists of several dialects spoken in the arid territory of southern Botswana, northern South Africa and central Namibia.

The Khoe-Kwadi languages are currently distributed over a large geographic area, including the Central Kalahari, the western territories in Namibia, the Okavango river delta, and the salt pans to the east of the Kalahari. The Kwadi language, once spoken in the coastal region of Angola, died out in the past century: the only information available is found in Westphal (1971). The Khoe-Kwadi language family is more distinct and harbors more diversity than the Kx'a and the Tuu family, with two distinct major branches: Khoekhoe and Kalahari Khoe, with the latter composed of East and West Kalahari Khoe. Cape Khoekhoe was once spoken at the southern cape and along the southern coast, which can be considered the place where the Nama language (and its close relative Hai‖om) developed before migrating North towards Namibia. East Kalahari Khoe is spoken in eastern Botswana. West Kalahari Khoe includes several sub-branches:

Kxoe, spoken around the Okavango delta, while the remaining languages such as G‖ana, G|ui and Naro are spoken in the central Kalahari.

Concerning the origin of the Khoe-Kwadi language family as a whole, linguistic evidence points towards an immigration of people from eastern Africa that were already familiar with cattle. It can be assumed that these pastoralists arrived in the region inhabited by the Khoisan hunter-gatherers before the Bantu; the linguistic link to a pastoralist migration comes for the shared presence of terms related to food production (Ehret 1982, Vossen 1996, Güldemann 2008a).

Chapter 4

KHOISAN POPULATIONS

4.1 Terminology and historical perspective

Khoisan is a term widely used to define populations who live in southern Africa and who speak non-Bantu languages characterized by heavy use of click consonants. The term was first proposed by the physical anthropologist Leonard Schultze (1928) to incorporate in a single definition the pastoralist Khoe (in the past also called "Hottentots"), and the foraging San (commonly referred to as "Bushmen"), and it was further popularized by the cultural anthropologist Isaac Schapera in the 1930s (Schapera 1930). The word "Khoi" or "Khoe" means "person" in Nama. Two surviving pastoralist groups, the Nama and Korana, use the word "Khoenkhoen", meaning "people of the people". The word "San" is the Khoe word for "foragers" or "bushmen" (Barnard 1992). Despite its early biological connotation, the term "Khoisan" became widespread by association with linguistics when Greenberg adopted it to name one of the linguistic phyla of Africa. The supposed genealogical unity of the languages (implied by Greenbergs classification) and of the people who speak these languages, for which both here we will refer to with the conventional term "Khoisan", is contradicted by linguistic, anthropological and genetic evidence (Güldemann 2008b, Güldemann and Stoneking 2008). The terms employed here to identify the single populations refer to the linguistic affiliation of the various ethnic groups, mainly following the spelling reported in Güldemann (forthcoming). It must be noted that many

similar terms are reported in literature, often to identify different groups: the definitive nomenclature system is still debated (Güldemann, p.c.).

The "Khoisan paradigm" has influenced our perception of the variability of these Khoisan populations. It represents a heritage of the first ethnographic records, when the tendency was to disregard dissimilarities between groups, their different origins and populations trajectories, while instead favoring the examination of their shared features (Fauvelle-Aymar 2008). This point of view is strongly opposed nowadays, and current focus has been on all the diversity within the Khoisan groups (summarized in Figure 4.1).

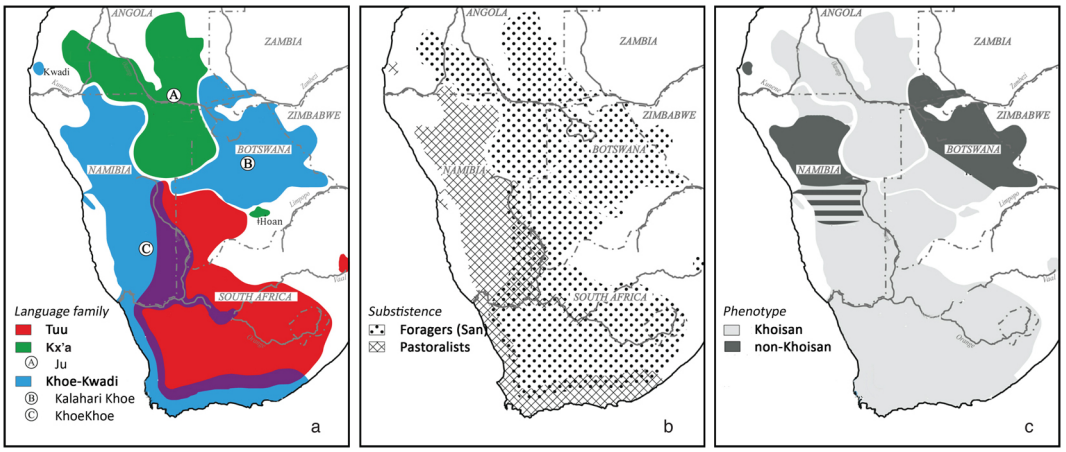


FIGURE 4.1: Historical distribution of Khoisan populations, according to a) language affiliation, b) type of subsistence, c) phenotypic characters (modified from Güldemann 2008a)

4.2 Foragers of the Kalahari Basin

The Kalahari is a semi-desert area with a long dry season and limited natural resources, where hunter-gatherers have possibly been resident since the Late Stone Age (LSA), ~30 kya (Denbow 1984, Deacon and Deacon 1999). The beginning of the LSA is marked with the transition toward a technology that included specialized tools like bows and arrows, needles, bored stones, tools for fishing, and microlithic tools. The paleoethnographic record, along with evidence of art, burials and various archeological remains suggest cultural continuity in the area, for which the extant Khoisan inhabitants represent the direct descendants of this foraging culture. All Khoisan people of the Kalahari possess a deep

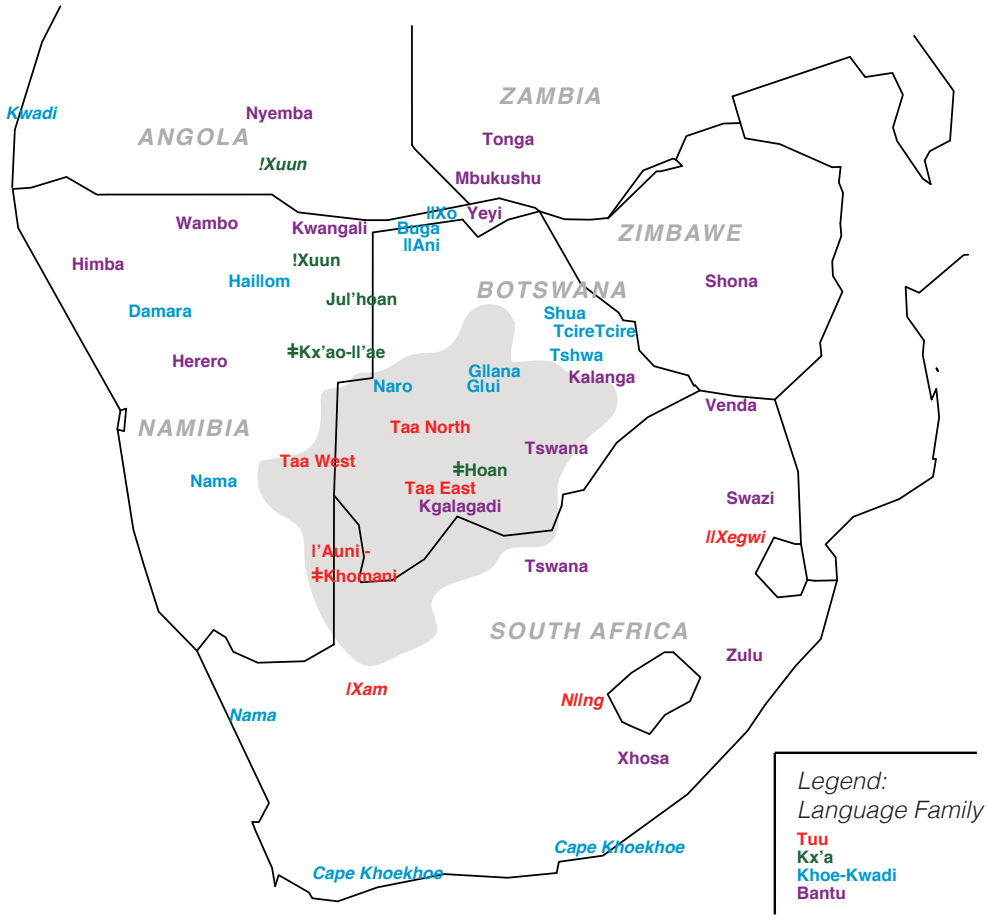


FIGURE 4.2: Approximate location of Khoisan populations and neighboring Bantu-speaking populations, colored according to linguistic affiliation. Extinct groups in italic font. The gray area indicates the core of the Kalahari semi-desert.

knowledge of the environment and of the fauna and flora, in particular of the variety, ecology and use of available plants (Barnard 1992).

These Khoisan populations share a similar "Khoisan phenotype" (Figure 4.1.c): a set of physical characteristics described in the first ethnographic records, that includes on average a short stature, a fairly light ("yellow") skin color, tightly coiled "peppercorn" hair, high cheekbones, and epicanthic eye folds (Huxley 1870), which makes them distinct from the majority of Sub-Saharan African populations, who are on average darker in skin tone. Currently, foragers of the Kalahari speak languages of the Kx'a, Tuu and Khoe families.

4.2.1 Kx'a

Within the Kx'a family, speakers of the Ju linguistic dialect cluster are distinguished in three units on the base of cultural identities and environmental zones (Figure 4.2): these are the !Xoon, the Ju|'hoan, and the ǂKx'ao-ǁ'ae (or ǂKxauǁ'en). The Ju|'hoan (called "Ju|'hoan North" in Paper III (8) and IV (9)) are the most numerous, and occupy the northern areas of Botswana, where the Kalahari experiences a transition toward a wetter environment, as well as part of Namibia. They are also the most studied of the San populations, referring in particular to the extensive ethnography collected in Nyae Nyae and Dobe (Marshall 1976, Lee and Marshall 1979, 1984), which are still important documents on the ecology, subsistence and society of these foraging communities. Ju|'hoan used to live in small bands (25 individuals) with tight boundaries and high level of endogamy, with preferential uxorilocal post-marital residence (Marshall 1959, Lee and Marshall 1984).

The !Xuun live actually outside of the Kalahari, in forested areas of Namibia and southern Angola and are the most widespread population of the Ju dialect cluster. They live in association with Namibian Bantu-speaking agriculturalists, such as the Ovambo. The ǂKx'ao-ǁae (called "Ju|'hoan South" in Paper III (8) and IV (9)) live in the south of the Ju|'hoan, in a territory largely shared with the Naro.

The ǂHoan also speak a language of the Kx'a family; their existence was discovered only in 1970 by linguist Anthony Traill (Traill 1973). The status of the language is severely endangered, being currently spoken by less than 50 people (Gerlach in preparation). The origin and history of the ǂHoan remains unclear. In recent times, with scarcity of game to hunt, they have become more dependent on the Bantu-speaking Kgalagadi people, working for them as herders and laborers (Batibo 2005).

4.2.2 Tuu (Taa)

People speaking Tuu languages live in southern areas, mainly between south Botswana and north South Africa, and neighboring areas of Namibia (Figure

4.2). It is also here that dialects of the Taa language, like !Xoon, !Ama, Tshaasi and ǂHuan are spoken. In this region the Kalahari is particularly arid and harsh, and survival depends on the ability to locate and identify the species of plants that are the only sources of water in the dry season. This territory has also been inhabited for centuries by Bantu-speaking Kgalagadi, with whom they have interaction (Barnard 1992). Our knowledge of Taa social structure mainly relies on the work of H.J. Heinz, a parasitologist who started a study in parasitic diseases of the Bushmen and ended with a major interest in !Xoon culture. Their unique linguistic complexity (!Xoon seems to possess the vastest number of phonemes of any other language in the world — Traill 1994) made Taa more attractive for linguists than for anthropologists. !Xoon traditional society and mating system involves small scale exogamy and multilocal post-marital residence, with a tendency towards uxorilocality (Heinz 1994). The largest unit is the band cluster, which in many cases corresponds to the dialect unit (Barnard 1992). Each band cluster is divided into several bands, which have 30-45 members each (Heinz 1994).

4.2.3 Kalahari Khoe

Some Khoe speakers have inhabited the central Kalahari for a long time and are associated with a traditional foraging lifestyle. Naro, G|ui and G‖ana speak languages of the West Kalahari branch and live in strong interaction with other foragers. Naro are some of the most numerous Khoisan populations, estimated by Guenther (1986) as numbering 9,000 individuals. They live in western Botswana, in areas with a relatively good water supply, where ranchers have their cattle posts: the majority of Naro live in these cattle posts or in towns (Barnard 1992). Naro practice band-exogamous marriage and have fluid residence patterns (Barnard 1976).

G|ui and G‖ana used to inhabit the area of central Botswana now occupied by the Central Kalahari Game Reserve (CKGR), in isolation from the rest of Khoisan foragers (Barnard 1992). The CKGR was founded in 1961 by George B. Silberbauer, who is also the author of an important ethnography of the G|ui people: in the 60s, the Khoisan in CKGR numbered around 2,000 individuals

(Silberbauer 1965, Barnard 1992). In 1986 the government decided to turn the CKGR into a strict wildlife reserve, and suggested the relocation of the local hunter-gatherers. In 1997 the people from the main village of !Xade were moved to another settlement in the west of the CKGR, called New !Xade.

Traditionally, G|ui practiced exogamy in a way proportional to the size of the band (endogamy was common in large bands) and multilocal post-marital residence: uxorilocal first (for the time given to the bride service at the brides parents house) and virilocal after (Silberbauer 1981). The G‖ana practiced exogamy to provide alliances over the region: 74% of their partners were born in a different location (Cashdan 1984). The tight bonds between the Kalahari Khoe populations and the other foragers of the central Kalahari, reflected also in the linguistic record, suggested that they would have shifted from a Kx'a or Taa substrate to a Khoe language only in relatively recent times, when exposed to other Khoe languages that appeared in the area around 2,000 years ago (Güldemann 2008a — see following section).

4.3 Khoe pastoralists

Nowadays, Khoe speakers show a higher level of diversity in comparison to Tuu and Kx'a speakers. First, the Khoe languages are more diverse and scattered over a wider area (Figure 4.1.a). Second, they are very diverse in terms of their way of subsistence (Figure 4.1.b): while some populations are or were foragers (with a focus on fishing on the Okavango river) at least until historical times, some other Khoe speakers were attested to be pastoralists; although currently only the Nama of Namibia still practice a lifestyle based on herding. Today, most of these populations are integrated in the trade system of the Bantu-speaking societies, and practice herding as well as cultivating (Barnard 1992). Third, they are very diverse in physical appearance (Figure 4.1.c), while many individuals can be ascribed to the prototypical "Bushman" type, populations of the eastern Kalahari together with the Khwe of the Okavango delta and Caprivi strip are characterized by a darker skin phenotype, earning them the label of "Black bushmen" (e.g. Weiner 1964, Jenkins 1986, Gusinde 1966). Also Damara from Namibia possess a dark skin phenotype.

The majority of Khoe speakers live in more decentralized areas of the Kalahari; linguistic evidence suggests that they may represent the descendants of a migration of Khoe-Kwadi speakers with a herding economy (Güldemann 2008a). The putative origin of these Khoe-Kwadi populations is in East Africa, where livestock was first domesticated (Phillipson 2005, Deacon and Deacon 1999 — see figure 4.3). The arrival of pastoralist populations is supported by archaeological evidence: a complex of pottery styles and remains of domesticated animals appears in the coastal regions of what is now South Africa and Namibia and in northern Botswana 2,000 years ago almost simultaneously, suggesting a rapid spread over the territory (Deacon and Deacon 1999, Smith 1992, Mitchell 2002, Pleurdeau et al. 2012). Once the pastoralists reached the Kalahari, they came into contact with the local foragers, with whom they start an intensive exchange promoted by the complementarity of the two subsistence strategies (Deacon and Deacon 1999).

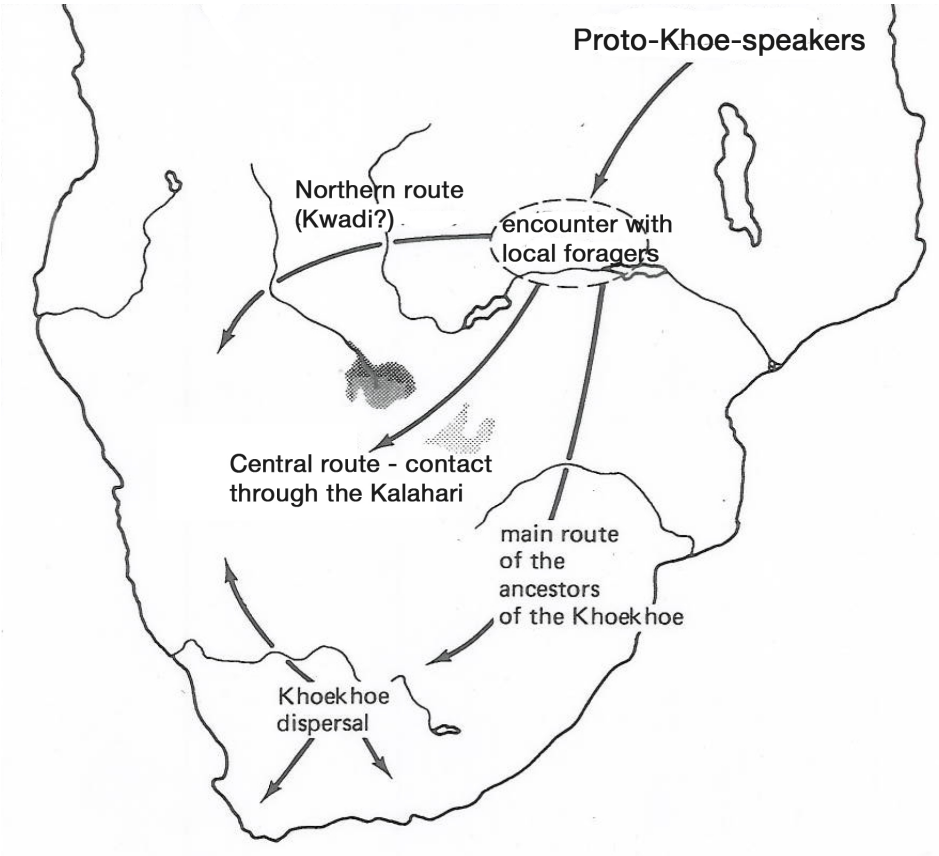


FIGURE 4.3: Probable migration routes of the Khoe speaking pastoralists (modified from Barnard 1992)

4.3.1 East Kalahari Khoe

The Khoe speakers of the East Kalahari inhabit a region around the salt pans of eastern Botswana. In their dialects, they use the term "shua" or "tschua" for person, rather than the term "khoe" (Barnard 1992). The population called Shua lives in the north of the salt pans, while the Tschua (or Tshwa) live in the south. Other dialects are also reported in the area, like Tcire Tcire or Danisi. These populations are in trade exchange with the local Bantu-speakers (primarily Tswana and Kalanga), with whom they establish characteristic contract works: they take care of the cattle of the Bantu-speaking group and have some benefits in return like milk, meat, or the right to use the cattle for some activities. This kind of contract between Bantu speakers and Khoisan is called "mafisa" (Barnard 1992). Due to these relationships they can be currently described as herders, even if they also currently practice foraging.

4.3.2 West Kalahari Kxoe

The populations speaking languages of the West Kalahari Kxoe branch are also called Khwe, and include speakers of the ǁAni, Buga and ǁXo dialects who live around the delta of the Okavango river, a swamp area affected by seasonal flooding in the north of Botswana and the Caprivi Strip. They share this territory with other local Bantu speakers like the Mbukushu, who are predominantly agriculturalist, and the Yeyi, who are the prototypical Okavango fishermen (Barnard 1992). Since herding is not practicable because of tse tse fly infestations, the economies of these Khwe populations are based on fishing in the wet season and hunting and gathering in the dry season (Barnard 1992). Post-marital residence is often virilocal from the beginning of the marriage, and polygamy is common (Heinz n.d., Barnard 1992).

4.3.3 Khoekhoe

Nama and Hai‖om are Khoe speakers of Namibia, who are linguistically distinct from the West Kalahari Khoe and East Kalahari Khoe. The Nama are very well described in the ethnographic record, which reports much of their history

of chiefs, battles and migrations, and are also a very numerous consisting of ∼90,000 individuals (Barnard 1992). They used to live a nomadic life, with their main protective concern being their cattle and women (not lands) and have a culture deeply influenced by cattle and pastoralism. They practice polygyny and have patrilineal society, with virilocal post-marital residence (after initial uxorilocality and bride service); they intermarry with neighboring tribes to keep connections between families (Hoernlé 1985, Barnard 1992). They probably originated in the southern cape, and subsequently moved to Namibia in two splits: the Great Nama, who settled in the great Namaqualand of Namibia prior to the European contact, and Little Nama, who moved in tribal groups during the XIX century (Westphal 1963, Barnard 1992). In 1904, the Nama together with their traditional enemies the Bantu-speaking Herero started a revolt against the German authority that culminated in a war: the consequences were a severe reduction in population size and a restructuring of the internal tribal organization (Barnard 1992). The Hai‖om speak a language almost identical to the one of the Nama and live in north Namibia, near the Etosha pans. Their origin and relationships within the Khoisan groups are rather mysterious, due to the absence of literature and their ambiguous characterization: they are quite numerous, physically they resemble their !Xoon neighbors, and possibly they are described as !Xoon shifters towards a Nama language in recent times (Barnard 1992).

4.3.4 The extinct Kwadi of Angola

Kwadi was a distinct isolated language related to the Khoe family. In the 1950s only a few families could speak the language, according to the description of the ethnographers (Estermann 1976). The Kwadi people, who inhabited a small region of the Angolan coast and were cattle herders, have mysterious origins and relationships with the other Khoisan people, and are poorly described in the literature. They were described as black people of southwestern Angola together with the hunter-gatherer Kwisi and Cimba, who currently speak a Bantu language. Estermann (1976) suggests that the Kwadi might be related to Kwisi and Damara, and Kwadi were described as the result of a mix of Khoisan and Kwisi characters.

4.4 South African Tuu

Tuu languages had a larger distribution in the past, covering most of South Africa (Figure 4.1, Figure 4.2). These former foraging populations have lost their identities and have been assimilated by other populations (mainly Nguni Bantu speakers), after experiencing contact with the various Bantu-speaking populations and European colonialists immigrants. The stereotype of the classic "Bushmen" for the colonialist time comes from the encounters of these Khoisan foragers of the Cape, not from Khoisan of the Kalahari (Barnard 1992).

Some of these Tuu extinct languages include the |Xam of the Karoo, for which an extensive linguistic and cultural survey is available from the work of Wilheilm Bleek from 1870. Descendants of the |Xam are the Karretjie people, small bands of donkey cart drivers who provide cattle herding for the colonists (who referred to them as the "tame Bushmen" — de Jongh 2002) and who nowadays speak Afrikaans.

The N‖ng or "mountain Bushmen" from Lesotho were already "dying out" in the end of the XIX century (Barnard 1992). The ‖Xegwi of the eastern Transvaal are one of the less known Khoisan people sand in 1955, only 66 living individuals were reported (Potgieter 1955). The |'Auni - ǂKhomani lived in the northern areas of the northern Cape and in southern Botswana and southern Namibia; in 1983 a remnant population of hunter-gatherers was still inhabiting the area, but was speaking mainly Nama (Steyn 1984).

4.5 The case of the Damara

The Damara speak a Khoe language, but their characterization as "Khoisan" people is controversial, because of their cultural connections to the Herero and Himba and because of their physical appearance which is remarkably similar to prototypical Bantu speakers. Damara cannot be grouped with any other population inhabiting the same area: their unique identity corresponds to the way the Damara see themselves and is confirmed by the description drawn by their neighbors and by physical anthropologists (cf. Nurse et al. 1976). They

have also been called Berg Damara, or "Hill Damara", in contrast to the Herero who were called "Cattle Damara". "Damara" is still the term for Herero in the Tswana language. Their origin is mysterious: it is believed that they arrived in Namibia before the Himba and Herero, and certainly before the Nama (Barnard 1992). For the past century they have worked as blacksmiths and servants for the Nama, from whom they adopted the language, and for the Herero to a lesser extent. Not much is known about their existence under the domination of the Herero and Nama in the XIX century, or before. Today they number at least 90,000 people (Barnard 1992), representing one of the largest ethnolinguistic groups of the country.

Chapter 5

GENETICS OF AFRICA

Paleoanthropological and archeological evidence suggests that early hominin species evolved in Africa: in fact, the earliest forms of Anatomically Modern Humans (AMH) appeared on this same continent 200 kya (Stringer 2002, Mc Dougall et al. 2005, Barham and Mitchell 2008). Genetic investigations on African populations have, therefore, generally contributed data useful for our understanding of the emergence of *Homo sapiens*. Various lines of evidence supporting the African origin of AMH have benefited from studies on the genetic variation of extant populations. Indeed, compared to non-African populations, Africans have overall the highest levels of genetic diversity, the widest population substructure, and the smallest linkage disequilibrium (LD) between loci (Garrigan et al. 2007, Li et al. 2008, Tishkoff et al. 1996, 2002, 2003, Jakobsson et al. 2008, Harding and McVean 2004). Furthermore, African populations display a higher number of private alleles and haplotypes in comparison to populations on other continents, where only a subset of African genetic diversity can be found (Conrad et al. 2006, Jakobsson et al. 2008, Ramachandran et al. 2005, Tishkoff et al. 2003).

Archeology and genetics are often used in conjunction for studies on human evolution, thereby compensating for their respective limitations and providing indispensable essential data to reconstruct the early stages of the evolution of AHM; for this time scale, other disciplines such as linguistics have inadequate resolution (See Chapter 2). Archeology provides morphological evidence from

fossils that can be directly traced in space and time, with the aid of stratigraphy and radiometric dating. However, the archeological record is incomplete and its resolution depends on the effort expended in finding and excavating sites, as well as on the effects of site formation processes and taphonomy on the remains left behind by early AMH. Genetic variability in modern humans, on the other hand, can provide a complete draft of the extant worldwide variability, which should reflect the prehistoric demographic events. This kind of evidence, nevertheless, is subject to other limitations: people who currently live in a given territory might not represent the people who inhabited the same territory in the past and dating their presence from the coalescence of their genetic profiles poses further problems. Citing an example initially proposed by Barbujani (1998), if European people colonize Mars tomorrow, their time of coalescence would date back to the Paleolithic, but dating the colonization of Mars to the Paleolithic would be erroneous.

At present, Africa displays a wide diversity of people, languages (see Chapter 3), cultures and environments. The climates range from those present in the largest desert of the world to those of the extensive equatorial rainforests, from those typical of savannah habitats to those of mountain highlands These climates have often shifted and changed in the past (Kuper and Kropelin 2006), offering the ideal background for the evolution of the extant diversity.

5.1 Where in the African continent did *Homo sapiens* originate?

The precise location of the first appearance of AMH has not been indicated unanimously by the different disciplines that deal with the study of human origins (see a review in Batini and Jobling 2011). Most archeological data support an East African origin. The earliest remains associated to the appearance of morphological traits specific of AMH were found in Ethiopia, and dated 150-190 kya: these are *Omo 1* and *Homo sapens idaltu* (White et al. 2003, Mc Dougall and Fleagel 2005). The way these modern traits emerged is debated

and the prevailing hypotheses suggest that they resulted from a gradual morphological transition from archaic to modern forms (Brauer 2008) or from a sudden change into a distinct species (Lieberman et al. 2002). The earliest species of hominins also evolved in East Africa: *Ardipithecus kadabba* (5.2-5.8 million years ago, Haile-Selassie 2001), *Ardipithecus ramidus* (4.4 mya, White et al. 2009), and the better-known *Australopithecus* species, *anamensis* (3.9-4.2 mya, Leakey et al. 1995) and *afarensis* (i.e. "Lucy", 3-3.9 mya, Johanson and White 1979). The origin of the genus *Homo* is linked in particular to the East African Rift, which is an active continental rift zone extending from the Afar region of Ethiopia to Mozambique. This rift valley started forming in the Early Pliocene and favored the appearance of lakes and wet habitats, as well as causing a biotic fragmentation of the rain forests; factors that probably played a crucial role in the evolution of the first hominins, as proposed by the "East Side Story" model (Coppens 1983). This system of rifts also turned out to be an extremely favorable environment for the preservation of fossils, probably introducing a bias in the archeological record; in fact other areas like Central and West Africa could be a potential source of ancient species (Brunet et al. 2005).

From a genetic point of view, different studies tried to establish the region of origin of modern humans within the African continent. While many studies are based on the assumption of an East African origin, placing the source of extant human variation in Ethiopia (Ramachandran et al. 2005, Prugnolle et al. 2005, Pagani et al. 2012), other studies have highlighted the fact that the most ancestral genetic components are found in the present-day hunter-gatherers of southern Africa (Tishkoff et al. 2009, Henn et al. 2010). A study on the Y chromosome conducted on more than 2,000 African samples, on the other hand, locates the most ancestral lineages in Central-West Africa (Cruciani et al. 2011); a more recent study confirms these findings with the discovery of the most ancient clade of the Y chromosome phylogeny (Mendez et al. 2013). Ramachandran et al. (2005) proposed the same region of origin in their study on autosomal microsatellites variation, correlating geographic distances and F_{ST} distances. Their conclusions, however, were cautious because of the scarcity of population data available within Africa (at least at the time the study had been conducted).

According to the "Out of Africa" theory, modern humans migrated away from Africa no earlier than 50-70 kya, and spread throughout the world (see a review in Henn et al. 2012), following either a southern coastal route through Arabia (Macaulay et al. 2005, Armitage et al. 2011) or a land route through the Levant (Forster and Matsumura 2005, Reed and Tishkoff 2006, Mellars 2006). The timing and the area from which AMH departed from the African continent is currently under debate. The oldest presence of AMH outside Africa is dated 100-130 kya and is found in the Levant (Grün et al. 2005, Grün 2006), attesting an earlier exit which probably took advantage of favorable climatic fluctuations (Bar-Yosef 2000). Sites in North Africa belonging to the Aterian culture, the beginning of which has recently been dated to as early as 140 kya (Richter et al. 2010), point to a possible different area of origin of modern humans (Garcea 2010, Hublin and McPherron 2012). Nevertheless, the time range of 50-70 kya proposed for the major exit from Africa finds confirmation in the paleontological record (Trinkaus 2005), as well as in the mitochondrial (Ingmann et al. 2000, Macaulay et al. 2005), X chromosome (Kaessmann et al. 2001), Y chromosome (Underhill et al. 2000) and autosomal data (Tishkoff et al. 2009). Paleoanthropological data, such as neurocranial morphometry, suggest that a single out of Africa would not suffice in explaining the extant variability, but rather support the existence of a more complex pattern of migrations — possibly more than one single population exit (Gunz et al. 2009).

5.2 Classical markers and autosomal markers

Genetic polymorphisms can be analyzed at the level of their gene products, such as blood groups and serum proteins. These were the first kinds of analyses developed in the field of human population genetics, and for this reason they are called "classical markers". Nowadays, some of these genes are still investigated in population comparisons, but at the DNA level (e.g., the major histocompatibility complex HLA — Histocompatibility Leukocyte Antigen). All the genetic markers analyzed at the DNA level for recombinant loci on nuclear chromosomes (excluding the sex chromosomes) are called autosomal markers: after the introduction of PCR techniques and the improvement of DNA based analysis,

these largely replaced the use of classical markers. The first studies of classical markers in African populations found a distinction between North Africans and sub-Saharan Africans (Cavalli-Sforza et al. 1994), confirming the role played by the Saharan desert as a barrier in preventing gene flow, as well as separating ecological zones. Within sub-Saharan Africa, a study on the gamma-globulin GM system and Rh haplotypes showed that the main signal of genetic differentiation is explained by affiliation to the four language phyla (Excoffier et al. 1991). As shown in the review from Sanchez-Mazas and Poloni (2008), Afroasiatic populations from East Africa exhibit a higher level of diversity, both among and within populations. They also share similar allelic or haplotypic frequencies with some Khoisan, supporting the link between them and East Africa, while Hadza and Sandawe do not resemble the Khoisan populations included in that study.

Large-scale autosomal studies with African populations appeared only in the last few years and confirmed the highest diversity in African populations, the high level of private alleles, high long-term population sizes on average (Tishkoff et al. 2009, Campbell and Tishkoff 2010), and found that the out of Africa migration resulted in a population bottleneck and reduced diversity (Liu et al. 2006, Ramachandran et al. 2005). The studies of Tishkoff et al. (2009) and Henn et al. (2011) are currently the most comprehensive because of the number of African populations sampled and the amount of autosomal data examined: their main findings confirmed the presence of genetic structure in the continent as well as the high differentiation of hunter-gatherer populations. In particular, Tishkoff et al. (2009) could identify 14 clusters of genetic variation that distinguish populations according to their cultural identity and linguistic affinity, and observed widespread elevated levels of mixed ancestry which are explained by prehistorical migrations over the continent. Finally, they suggested a shared ancestry for geographically separated Pygmies and Khoisan foragers. Other recent autosomal studies focused on Khoisan populations in particular (Pickrell et al. 2012, Schlebusch et al. 2012) and will be discussed in the last paragraph.

5.3 Uniparental markers

Non-recombinant genetic markers with uniparental inheritance have been widely employed in the first studies of molecular variation in humans. Their advantageous features include the possibility of reconstructing genealogies whose origin and diffusion can be traced in time and space (Underhill and Kivisild 2007). The phylogeographic component of Y chromosome and mtDNA studies is best represented in the use of haplogroups, which are lineages characterized by stable mutations shared from a common ancestor that are usually named with capital letters (Karafet et al. 2008, van Oven et al. 2009).

Genetic admixture can often be sex-biased when the patterns of contact involve differential mating preferences, or unbalanced relationships between different groups based on sociological factors. In this view, the comparison of mtDNA and Y chromosome can provide information on the directionality of the demographic processes, when the two markers describe non-overlapping patterns within the same geographical region (Oota et al. 2001, Bolnick et al. 2006, Destro-Bisol et al. 2004, Gunnarsdottir et al. 2011). Paper II (7) describes a sex-biased situation of contact between Bantu speakers and Khoisan, comparing mtDNA and Y chromosome data from several populations of southwest Zambia.

5.3.1 mtDNA

The maternally inherited mitochondrial DNA consists of a circular haploid molecule of 16569bp, with a small non-coding region at high variability (hypervariable region, including two segments HVS1 and HVS2, each approximately 350 bp long) and a large region of 37 contiguous genes coding for 24 RNAs and 13 proteins (Anderson et al. 1981). Until recently, the majority of studies focused only on the sequencing of the hypervariable region and the typing of specific sites of the coding region (in order to assign the corresponding haplogroup). At present, the trend is to sequence complete mtDNA genomes, to gain more fine-grained information and enough power to perform statistical analyses like simulations and phylogeny based demographic reconstructions (Torroni et al. 2006). Papers II (7), III (8) and IV (9) analyze the maternal genetic makeup of a set of

populations with data from full mtDNA genomes. The value of mtDNA in reconstructing human origins was first highlighted by the pioneering work of Cann et al. (1987), one of the first studies that supported the African origin of AMH from a genetic perspective.

The phylogenetic tree of modern mtDNA variation coalesces at around 200 kya (Behar et al. 2008, Soares et al. 2009 — see Figure 5.1). As stated in the introduction, the earliest splits of this tree are located in Africa, while only a subset of variation is present outside of Africa. African lineages are named after the letter L: most common (and almost exclusive of the continent) are L0, L1, L2 and L3, with L4, L5 and L6 at minor frequencies. The first split is between L0 and L1'6: L0 lineages coalesce at ∼150 kya. L0d and L0k haplogroups have been recognized as characteristic of the Khoisan populations (Tishkoff et al. 2007, Behar et al. 2008). The genealogy of L0d and L0k lineages is reviewed in Paper III (8), with the aid of what is currently the largest dataset of individuals belonging to these ancient lineages.

The distribution of haplogroup L0 was overlapped by the diffusion of L1, which coalesces 140 kya, and subsequently by the appearance of L2 and L3, which coalesce at ∼100 and 70 kya respectively (Behar et al. 2008). The most recent African haplogroup, L3, is the precursor of the haplogroups present in the rest of the world (Watson et al. 1997, Macaulay et al. 2005), possibly marking the out of Africa migration. The majority of the sub-lineages of African macrohaplogroup L are widespread over sub-Saharan Africa, but some have been tentatively associated with the Bantu expansion, such as L0a, L2a, L3b, and L3e (Salas et al. 2002). Finally, lineages of L1c are characteristic of Pygmies, in particular from the Western group (Batini et al. 2007; Quintana-Murci et al. 2008).

5.3.2 Y chromosome

The paternally inherited, non-recombinant portion of the Y chromosome (NRY) displays a lower level of diversity than mtDNA, but the pattern of variation is more structured (Underhill and Kivisild 2007). In general, genetic structure correlates with language affiliation more for Y chromosome than for mtDNA (Forster and Renfrew 2011), a result that is also explained by the diffused practice

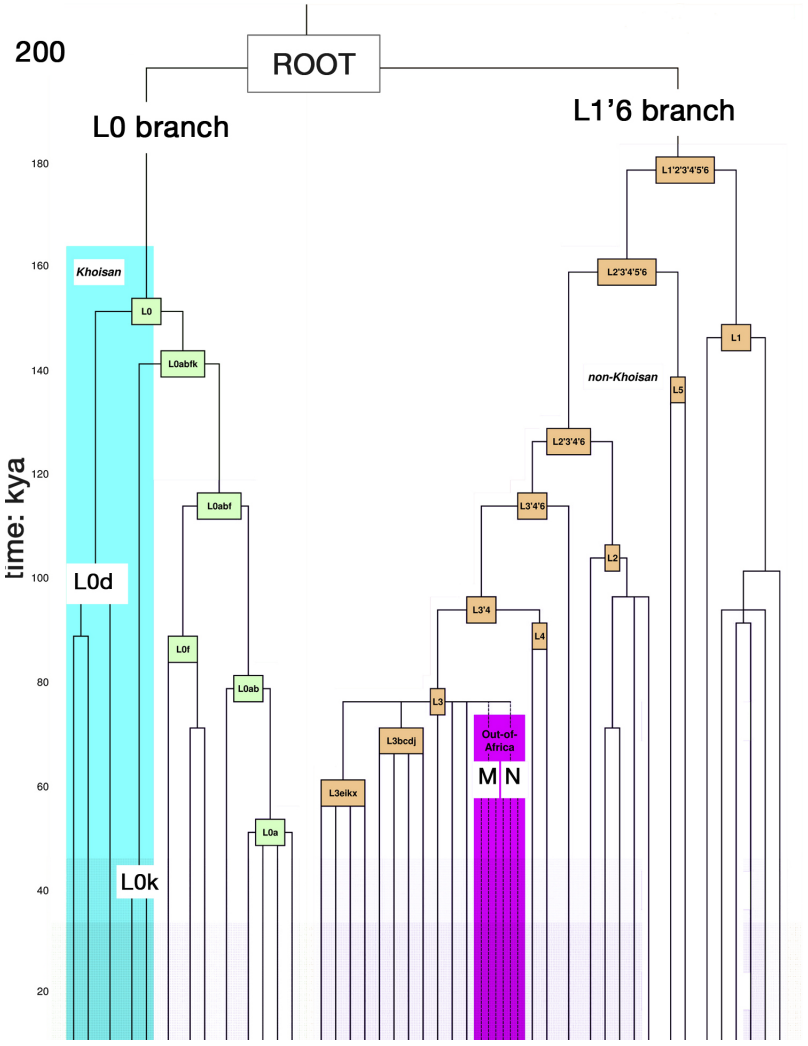


FIGURE 5.1: Schematic representation of the mtDNA phylogenetic tree, with major African haplogroups highlighted. Haplogroups associated with Khoisan populations in light blue, haplogroups associated with the out of Africa in purple. (modified from Behar et al. 2008)

of patrilocality (Heyer et al. 2012). For the Y chromosome, haplogroups are defined on the basis of stable SNPs, in a similar way to the mtDNA. Further definition is gained from the analysis of repetitive units or STRs (Short Tandem Repeats), which are usually typed with the aid of commercial kits in sets of 7, 12 or 17 loci. Papers I (6) and II (7) analyze the paternal genetic makeup of several african populations with data from Y chromosome haplogroups and a set of 12 STR loci.

The Y chromosome phylogeny also harbors deep and divergent splits, represented by haplogroups A and B, which are observed only in Africa. Haplogroup A and haplogroup B-M112 are characteristic of Pygmy and Khoisan populations (Wood

et al. 2005, Batini et al. 2011), while B-M150 is associated with the Bantu migration (Berniell-Lee et al. 2009). Haplogroup E is the most diffused in the continent, and probably originated in East Africa (Underhill et al. 2001). An alternative origin in Asia has been proposed (Hammer et al. 1997, Chandrasekar et al. 2007), based on the distribution of the sister clade D in Asia, and on the derived position of both clades in the phylogeny. Haplogroup E includes many well defined branches like E-M2, the most diffused one and characteristic of Niger-Congo populations (Wood et al. 2005); E-M35, characteristic of Afro-Asiatic populations (see review in Lancaster 2009); E-M33, found in West Africa (Wood et al. 2005); E-M75, found at low frequencies all over sub-Saharan Africa (Wood et al. 2005, Berniell-Lee et al. 2011). Paper I (6) discusses the distribution of the sublineages of haplogroup E across the African continent, focusing on the sublineages of E-M2, and finds markers that are selectively present in Bantu speakers and absent in other speakers of the Niger-Congo phylum. It is clear that the higher the resolution (allowing further sublineages to be defined), the more detailed descriptions could be drawn for the populations of interest. In fact, recent papers considered sequence data instead of single SNP typing, and proposed a revision of the phylogeny, especially of the early splits of the A and B branches (Cruciani et al. 2011, Scozzari et al. 2012, Wei et al. 2012, Mendez et al. 2013).

5.4 Ancestral population structure: *Ex Africa semper aliquid novi*

Ancestral populations in Africa were probably structured before the out of Africa exodus (Figure 5.2): the deep coalescent times of mtDNA and X chromosome genealogies suggest the presence of ancient lineages as remnants of a wider prehistorical diversity (Gonder et al. 2007, Behar et al. 2008, Shimada et al. 2007, Yotova et al. 2007). From an archeological perspective, craniometrical data in AMH fossils from 200 to 60 kya reported a high level of morphological divergence, which can be interpreted as a signal of ancient population structure dating back to the Pleistocene (Gunz et al. 2009). The term "population structure" refers to genetic heterogeneity resulting from non-random mating; the presence

of such distinct populations can be hypothesized, but is impossible to trace and associate to archeologically defined populations without directly linking to the related ancient DNA material (aDNA).

Admixture could have also occurred with archaic forms of Homo that had possibly coexisted with AMH in Africa. The only archaic hominin species for which DNA has been sequenced include the Neandertal and Denisovan. A variable proportion of admixture with these forms of Homo is found in non-African populations; in particular, 13% of the genome of all Eurasians and native Amerindians is of Neandertal origin (Green et al. 2010), and Papua New Guineans and Australians have another 3.5% of their genome consisting of Denisovan origin (Reich et al. 2011). Recently, admixture with Neandertals has been uncovered for North Africa as well: North African populations have a significant excess of derived alleles shared with Neandertals, when compared to sub-Saharan Africans, which are confirmed to be the only populations not affected by the admixture event with Neandertals (Snchez-Quinto et al. 2012).

Computational analysis performed without the direct comparison with archaic hominin DNA already suggested the presence of archaic introgression in Eurasian and African populations (Wall et al. 2009). Hammer et al. (2011) specifically searched for a signal of archaic admixture in Africa, using a set of non-coding autosomal loci typed in Mandenka, Biaka Pygmies and Khoisan (San) populations: their simulations supported a model of 2% introgression which happened ∼35 kya from an archaic population that split from AMH ∼700 kya. The patterns of LD in hunter-gatherers suggest that some portion of the genome would have been incorporated from a now-extinct taxon that might have lived in central Africa.

Archaic admixture in sub-Saharan Africa can be hypothesized from simulation data, but the direct comparison between the DNA of local extant populations and aDNA data is not feasible for the lack of well-preserved biological remains. In these conditions, the presence of ancient structure can be assumed from highly divergent lineages that survived until present time, which can be seen as relict of a prehistoric genetic variation landscape (see for example Behar et al. 2008). The time scale does not suggest the presence of distinct hominins, but rather a more

variegated scenario of structured populations that contributed to the evolution of AMH in different proportions. This scenario is tested with simulations in Paper II (7) and Paper III (8), where novel lineages within haplogroup L0k are discovered and linked with a prehistoric variation in southern African hunter-gatherers: some of these populations may be extinct at present, but their genetic component was partially absorbed by Bantu-speaking immigrants that had contact with them.

However, such ancestral and isolated lineages are not exclusive of mtDNA. The most basal clade of the Y chromosome phylogeny, named A00, has been recently discovered in one African American individual: its origin traces back to Cameroon, and its divergence from the rest of the trees is dated at ∼400 kya, posing new challenging questions on the level of ancestral structure and possible archaic introgression in early hominids (Mendez et al. 2013).

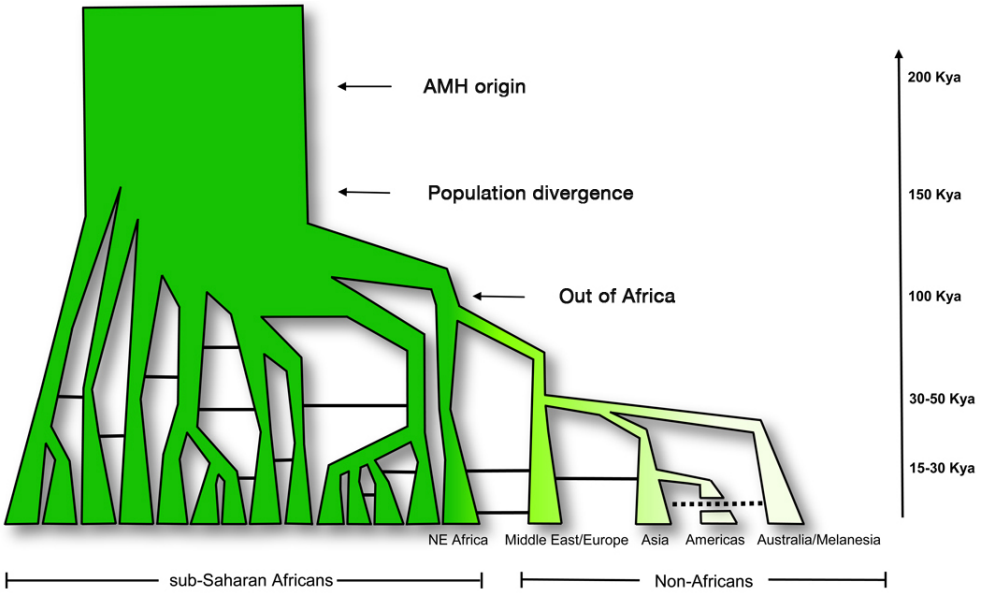


FIGURE 5.2: Model of population structure in African populations and gene flow. Decreasing intensity of color represents the loss of genetic diversity after the out of Africa. Solid horizontal lines indicate gene-flow between ancestral human populations. (adapted from Campbell and Tishkoff 2008)

5.5 Humans in recent times: major routes of migration and contact

The current genetic variation in Africa is deeply influenced by recent episodes of long-range and short-range migrations, and subsequent gene flow between local populations and immigrants. As discussed above, genetic structure in the African continent is mostly associated with ethnolinguistic identity: population movements can be traced looking at the genetic variation within each language phyla (see Chapter 2 for a description of the African linguistic phyla) and also considering the geographic background of the continent together with the major eco-zones and geographic barriers, like the Sahara (Figure 5.3).

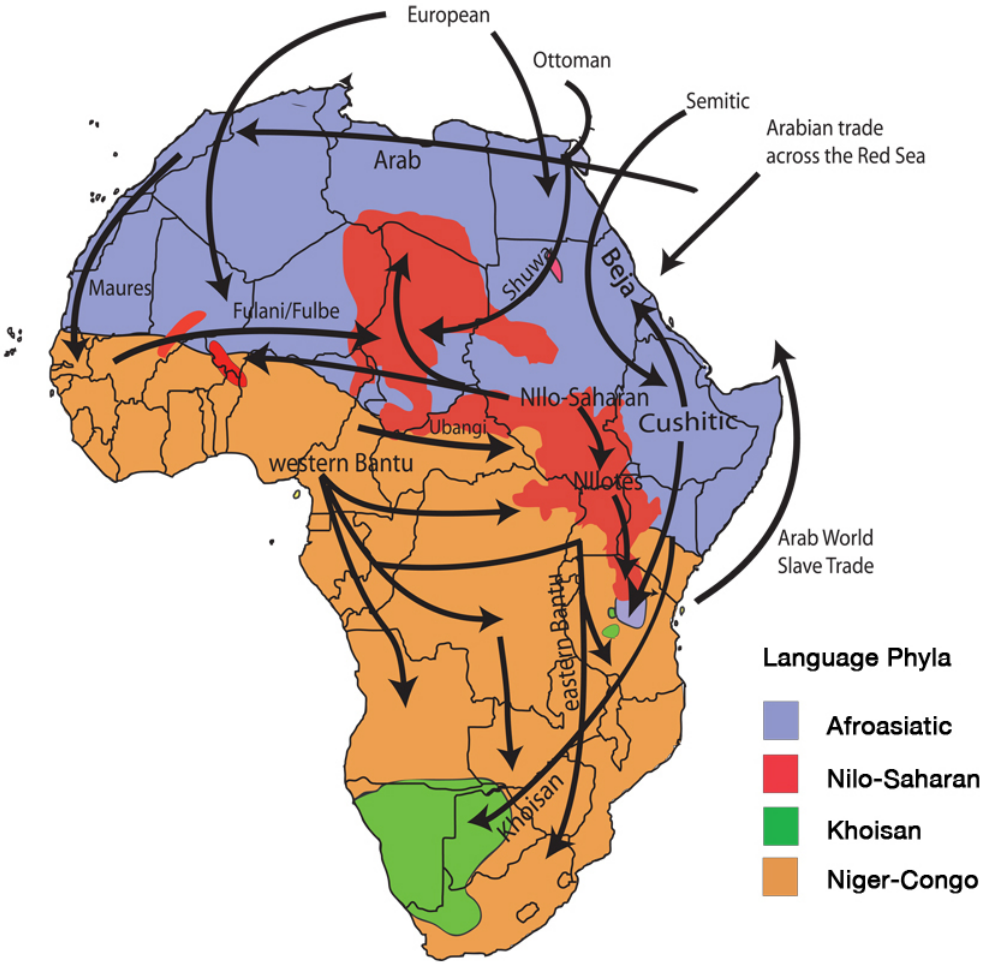


FIGURE 5.3: Map of major routes of migration within and out of Africa (adapted from Campbell and Tishkoff 2008)

Niger-Congo is the most widespread linguistic phylum. Its origin is located in Central-West Africa, where it possibly spread concomitantly with the Holocene climate change 10 kya (Dimmendaal 2008). We investigated the demographic trace of this expansion in a genetic study which considered different Niger-Congo populations from Burkina Faso, representative of the variability in western Africa with individuals sampled speaking languages from Mande and Gur families (Barbieri et al. 2012b). In this study we analyzed full mtDNA genomes of almost 300 individuals, realizing one of the first non-biased population-based datasets of mtDNA genomes in Africa. With this power of resolution we could generate Bayesian Skyline Plots (BSPs) with a narrow confidence interval. The BSP displays the variation of effective population sizes through time, and shows an unequivocal signal of increase compatible with the beginning of the Holocene Climate Optimum (Figure 5.4), which we suggest could also be correlated with the expansion of the first representative of the Niger-Congo phylum.

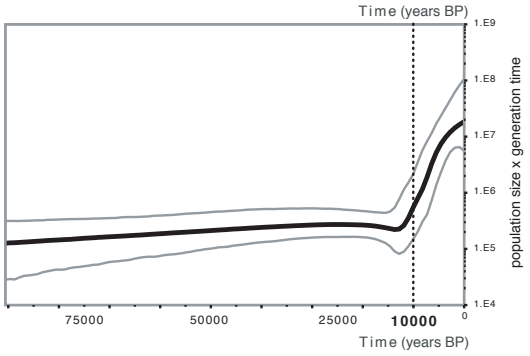


FIGURE 5.4: Bayesian Skyline Plot of mtDNA coding regions for all individuals from Burkina Faso belonging to African haplogroup L, performed with a relaxed clock model and the mutation rate employed in Atkinson et al. 2008 (Barbieri et al. 2012b)

This spread influenced the genetic variation in most of West Africa, but a subsequent expansion associated with a language family of the Niger-Congo phylum affected the genetic profile of a broader portion of the continent, from Central to Southern Africa. This was the Bantu expansion, which started 4-5 kya in the Grassfields of Cameron (Blench 2006, Vansina 1995) and rapidly spread until reaching the southernmost regions of South Africa. The presence of signals of Bantu genetic ancestry or of markers associated with the Bantu-speaking groups support a demic spread of Bantu-speaking populations (Tishkoff et al. 2009,

Berniell-Lee et al. 2009) who eventually intermarried with the locals in a typical sex-biased pattern (i.e. marrying their women). This happened when Bantu speakers met the Pygmies of Central Africa, who in the majority of cases abandoned their original language in favor of a Bantu language (Destro-Bisol et al. 2004, Batini et al. 2010, Verdu et al. 2013), and also when they met Khoisan populations in southern Africa, from whom, in some cases, they incorporated click sounds (Behar et al. 2008, Quintana-Murci et al. 2010, Batini et al. 2011). Paper II (7) and Paper IV (9) describe different scenarios of contact between Bantu speakers and Khoisan in southwest Zambia (where Bantu languages are currently the most widely-spoken) and in Botswana and Namibia (where both Khoisan languages and Bantu languages are spoken).

Within the African continent, a major corridor of migration is represented by the Sahel, connecting East and West Africa (Hassan et al. 2008, Černỳ et al. 2007, Tishkoff et al. 2009). In this territory most of the Nilo-Saharan languages are spoken: the highest proportion of Nilo-Saharan genetic ancestry is found in Central-Southern Sudan, indicating a possible place of origin of this phylum that subsequently had spread southeastward towards East Africa (Tishkoff et al. 2009). A study centered on the distribution of L3f mtDNA lineages suggests that Chadic pastoralists, who speak a language of the Afroasiatic phylum, originated in East Africa and spread ∼8 kya (Černỳ et al. 2009).

Several studies detected the greatest level of substructure in East Africa (Tishkoff et al. 2009), reflecting the presence of speakers of Khoisan (Hadza and Sandawe), Afroasiatic (Cushitic), Nilo-Saharan (Nilotic), and Bantu languages. This pattern is likely to be the result of subsequent waves of migrations within the past 5,000 years (Ehret 1974, 1983, Newmann 1997). A recent study explored the variation in Afroasiatic populations and revealed a similar ancestry component for the majority of Afroasiatic populations from Ethiopia, together with Semitic populations from Yemen and Egypt, and Chadic populations of Central Africa, while Afroasiatic from North Africa and the Levant (Berber and Semitic) clustered separately (Boattini et al. 2013). Many Nilo-Saharan speakers in East Africa have a high level of Afroasiatic Cushitic ancestry (Tishkoff et al. 2009), confirming a level of contact between these groups who also share common practices of cattle herding (Blench 2006).

The Sahara represents a geographical and environmental barrier in separating North African populations from the rest of the continent, at least since 5,000 years ago when desertification affected a larger region (Brooks et al. 2005). However, migrations through this desert are reported in historical times: a slave trade was carried by the Arab conquest of North Africa in the seventh century. The observation of mtDNA sub-Saharan lineages in North Africa supports the genetic impact of these recent migrations, while the settlement of this territory some 40,000 years ago did not involve Sub-Saharan people (Harich et al. 2010).

5.6 Genetic profile of Khoisan populations from available literature

During the last century, physical anthropologists dedicated increasing attention to the foraging San and pastoralist Khoe populations of southern Africa: their first genetic studies contributed to understanding the variability and the prehistory of the Khoisan populations within the African continent. The first study, based on classical markers, dates back to 1932 and was done by Pijper (1932), who analyzed ABO blood groups. More studies followed and included a variety of serogenetic markers in a relatively large number of Khoisan populations: a great contribution was put forth during the 70s and 80s by George T. Nurse and Trefor Jenkins. The first results of these studies based on classical markers suggested some distinction between the San and the Khoe populations. Analyses of the ABO and Rhesus blood group systems as well as the haptoglobins showed a strong signal of differentiation: specifically, the B allele has a very low frequency in the San groups, while in Khoe it was found at frequencies 4-8 times higher, similar to the Zimbabwean and Zambian Bantu-speakers and, to a lesser extent, the South African Bantu-speakers (Pijper 1932, Jenkins and Nurse 1972). Jenkins et al. (1971) analyzed the variation within several Khoisan populations using a combination of data from various serogenetic studies available (blood groups, serum protein and red cell enzyme systems). From their results, Khoe speaking populations clustered together: in particular, Khoe speaking Hai‖om are included in the Khoe cluster, contrasting the hypothesis that they could be a !Xuun group who shifted to a Khoe language. All the San populations except

the !Xuun are included in the same cluster, which is structured more according to the geographic proximity than to the language affiliation: for example, the Naro, who speak a Khoe language, cluster with ǂKx'ao-ǁ'ae and Ju|'hoan. The two central Kalahari Khoe speaking G|ui and G‖ana form a separate branch within the San cluster. The neighboring Bantu-speaking groups (Kgalagari and Tswana from Botswana, Herero from Namibia) form two separate clusters that also include the Khwe and Dama, who are classified as "Khoisan speaking Negros" by Jenkins (1986). These serological studies suggested that Khwe and Dama populations are genetically similar to Bantu-speakers (Nurse et al. 1976; Nurse and Jenkins 1977).

Studies on molecular markers are affected by a chronic scarcity of Khoisan populations included and by a superficial description of the chosen populations, which are often difficult to localize unequivocally in their appropriate ethnolinguistic and geographic context (cf. Mitchell 2010). Regarding uniparental markers, an emphasis has been placed on the presence of early divergent lineages, such as haplogroups L0d and L0k for mtDNA (Tishkoff et al. 2007, Behar et al. 2008) and haplogroups A-M51, A-M23 and B-M112 for the Y chromosome (Wood et al. 2005). In more detail, foraging San populations like Ju|'hoan, !Xuun, Nama, ǂKhomani and Karretjie People have values of approximately 90-100% L0d and L0k, while the Khwe and Damara have less percentages: 60% for the Khwe and only 20% for the Damara (Soodyall et al. 2008, Henn et al. 2011, Schlebusch et al. 2011). L0d is also found at a frequency of 5% in the click speaking Sandawe from Tanzania, but not in the click speaking Hadza (Tishkoff et al., 2007). Y chromosome haplogroup frequencies for Khoisan populations can be retrieved from Soodyall et al. (2008) and Wood et al. (2005). The Ju|'hoan have the highest frequency of typical Khoisan Y-chromosomal haplogroups, showing 90% or more, while other Khoisan populations included in literature have a variable percentage of typical Khoisan haplogroups: the !Xuun have 60%, the Nama lack haplogroup B and have only 35-64% haplogroup A (with 10-20% of Y-chromosomal haplogroups of Eurasian origin), the Damara and the Khwe have only ∼11-16% (with 10% Eurasian haplogroups in the Damara), the ǂKhomani from South Africa have 63% (with 10% Eurasian haplogroups — Henn et al. 2011), and the Karretjie People have only 10% (in spite of showing 100% L0d for the maternal

line — Schlebusch et al. 2011). These results support a variable level of sex-biased admixture for Khoisan populations, for which some of them (in particular Khoe speakers) have more non-Khoisan specific paternal lineages, which could come from Bantu neighbors or from European colonizers. Güldemann & Stoneking (2008) review the genetic data available for mtDNA and Y chromosome and display the variation within Khoisan and neighboring African populations with the help of MDS plots (Figure 5.5). For mtDNA, Ju|'hoan from Botswana (there denoted as Ju-B) are quite separated from all other African populations, while the !Xuun from Angola (there denoted as Ju-A) are intermediate between the Ju|'hoan and other Africans. The Khwe appear closer to other Africans than to the other Khoisan populations. For the Y chromosome, the !Xuun (there Ju-1 and Ju-3) are also intermediate between the Ju|'hoan (there Ju-2 and Ju-4) and the non-Khoisan groups, and the Khwe and Damara are located close to other Africans.

A small amount of studies addressed the variability of Khoisan populations with autosomal data. Henn et al. (2011) focused on the South African ǂKhomani and on the East African Hadza and Sandawe, showing a similar ancestral component between the ǂKhomani and the Ju|'hoan, while Hadza and Sandawe have their own characteristic ancestral components. Schuster et al. (2010) using 12 Ju|'hoan individuals, one !Xuun, and two Taa speakers, show a signal of differentiation between the three populations, who clearly diverge from other African groups. More significant is the contribution from two recent studies that, for the first time, included a large sample of different Khoisan ethnolinguistic groups and a fine-scaled analysis on autosomal data. The first is from Schlebusch et al. (2012): with 11 populations typed on a 2.3 million SNP array, they find a divergence between Khoisan from the North (Ju|'hoan and !Xuun) and from the South (Nama, ǂKhomani and Karretjie) dating to 35kya, and they connect the pastoralist Nama to other East African pastoralist populations by the presence of the same gene variant of lactose tolerance. The second study is from Pickrell et al. (2012) who include more Khoisan ethnolinguistic groups in a total of 23 populations examined with a genome-wide array designed for studies of population history. The authors find a split with an upper date of 30 kya similar to

FIGURE 5.5: MDS plots illustrating pairwise genetic distances in sub-Saharan African populations. Solid symbols denote groups with click languages. a) based on F_{ST} distances for mtDNA HV1 sequences; b) based on F_{ST} distances for eight Y-SNP haplogroups (adapted from Güldemann & Stoneking 2008)

the one in Schlebusch et al. (2012), but between Northwest and Southeast Kalahari Khoisan; they also observe variable levels of non-Khoisan admixture in all Khoisan populations that began 1,200 years ago, and finally they suggest a link between eastern Hadza and Sandawe in terms of a fraction of shared ancestry. Paper IV (9) includes all the populations typed in Pickrell et al. (2012), but with a higher number of individuals: complete mtDNA genomes are discussed and compared to the autosomal data of Pickrell et al. (2012), and analyzed to describe a detailed history of the various Khoisan populations on the maternal perspective.

Chapter 6

PAPER I: Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups

This chapter includes the paper "**Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups**" written by de Filippo, Cesare*, Chiara Barbieri*, Mark Whitten, Sununguko W. Mpoloka, Ellen D. Gunnarsdóttir, Koen Bostoen, Terry Nyambe, Klaus Beyer, Henning Schreiber, Peter de Knijff, Donata Luiselli, Mark Stoneking, and Brigitte Pakendorf. (*first authors equally contributed), as it appears in the published version on *Molecular Biology and Evolution*, 2011, 28(3): 1255-1269

Molecular Biology and Evolution

Research article

Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups

Cesare de Filippo,*†,[1] Chiara Barbieri,*†,[1] Mark Whitten,[1] Sununguko Wata Mpoloka,[2] Ellen Drofn Gunnarsdóttir,[3] Koen Bostoen,[4] Terry Nyambe,[5] Klaus Beyer,[6] Henning Schreiber,[7] Peter de Knijff,[8] Donata Luiselli,[9] Mark Stoneking,[3] and Brigitte Pakendorf[1]

[1]Max Planck Research Group on Comparative Population Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Department of Biological Sciences, University of Botswana, Gaborone, Botswana

[3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Royal Museum for Central Africa, Université libre de Bruxelles, Tervuren, Belgium

[5]Livingstone Museum, Livingstone, Zambia

[6]Department of Asian and African Studies, Humboldt University, Berlin, Germany

[7]Department of African Linguistics and Ethiopian Studies, University of Hamburg, Hamburg, Germany

[8]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

[9]Department of Experimental Evolutionary Biology, University of Bologna, Bologna, Italy

†These authors contributed equally to this work.

***Corresponding author:** E-mail: cesare_filippo@eva.mpg.de; chiara_barbieri@eva.mpg.de.

Associate editor: Sarah Tishkoff

Abstract

Technological and cultural innovations as well as climate changes are thought to have influenced the diffusion of major language phyla in sub-Saharan Africa. The most widespread and the richest in diversity is the Niger-Congo phylum, thought to have originated in West Africa ~10,000 years ago (ya). The expansion of Bantu languages (a family within the Niger-Congo phylum) ~5,000 ya represents a major event in the past demography of the continent. Many previous studies on Y chromosomal variation in Africa associated the Bantu expansion with haplogroup E1b1a (and sometimes its sublineage E1b1a7). However, the distribution of these two lineages extends far beyond the area occupied nowadays by Bantu-speaking people, raising questions on the actual genetic structure behind this expansion. To address these issues, we directly genotyped 31 biallelic markers and 12 microsatellites on the Y chromosome in 1,195 individuals of African ancestry focusing on areas that were previously poorly characterized (Botswana, Burkina Faso, Democratic Republic of Congo, and Zambia). With the inclusion of published data, we analyzed 2,736 individuals from 26 groups representing all linguistic phyla and covering a large portion of sub-Saharan Africa. Within the Niger-Congo phylum, we ascertain for the first time differences in haplogroup composition between Bantu and non-Bantu groups via two markers (U174 and U175) on the background of haplogroup E1b1a (and E1b1a7), which were directly genotyped in our samples and for which genotypes were inferred from published data using linear discriminant analysis on short tandem repeat (STR) haplotypes. No reduction in STR diversity levels was found across the Bantu groups, suggesting the absence of serial founder effects. In addition, the homogeneity of haplogroup composition and pattern of haplotype sharing between Western and Eastern Bantu groups suggests that their expansion throughout sub-Saharan Africa reflects a rapid spread followed by backward and forward migrations. Overall, we found that linguistic affiliations played a notable role in shaping sub-Saharan African Y chromosomal diversity, although the impact of geography is clearly discernible.

Key words: human, language, geography, migration, Y chromosome, Bantu.

Introduction

Modern humans originated ~200,000 years ago (ya) in Africa, subsequently colonizing the rest of the globe. Genetic studies indicate that the ancestral African populations could have been structured even before ~100,000 ya when modern humans first began migrating out of Africa (Campbell and Tishkoff 2008; Wall et al. 2009). Genetic diversity values are much higher in African populations than elsewhere (Campbell and Tishkoff 2008). Africa is also linguistically very diverse: More than 2,000 languages are reported for the whole continent, comprising 30% of the world's languages (Gordon and Grimes 2005). Disregarding some isolates, African languages have been classified into four major phyla (Greenberg 1948): Afro-Asiatic, Khoisan (which, however, is no longer considered a historical unit by several specialists, see Güldemann and Vossen 2000), Niger-Congo, and Nilo-Saharan. Of these, the largest linguistic phylum is Niger-Congo (Williamson and Blench 2000), comprising ~1,400 languages and containing many related language families and several distantly or questionably related language groups (Sands 2009). For instance,

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

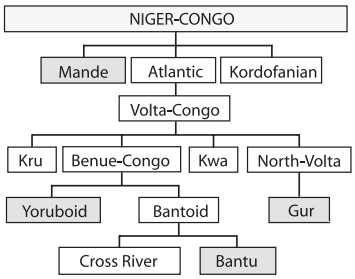


FIG. 1. Niger-Congo language tree. Schematic tree of the Niger-Congo language phylum that comprises three major branches: Mande, Kordofanian, and Atlantic-Congo (Williamson 1989). In gray boxes, linguistic families that are represented in our data set.

Mande and Kordofanian—two of the three major branches of Niger-Congo (fig. 1)—have been suggested as belonging to an earlier split, and some authors even doubt the affiliation of one or the other to the phylum (Williamson and Blench 2000; Dimmendaal 2008).

Since the migration of modern humans out of Africa, numerous population movements have played a role in shaping patterns of linguistic and genetic variation within the continent itself (Campbell and Tishkoff 2008). New forms of subsistence and technological improvements such as those derived from agriculture have driven population expansions even over long geographic distances. However, the major African linguistic phyla are assumed to have originated and spread much earlier than the advent of agriculture, which developed relatively late in sub-Saharan Africa: Cultivated plants did not appear before 4,000 ya (Neumann 2005). Indeed, it has been suggested that the expansion of Niger-Congo and Nilo-Saharan started ~12,000–10,000 ya with the improving climate at the beginning of the Holocene when speakers were still hunter gatherers (Blench 2006; Dimmendaal 2008). Nevertheless, it seems plausible that these expansions were triggered by technological innovations (e.g., bow, arrows, and domesticated dogs) and/or climatic changes (e.g., wetter conditions) in the Holocene approximately 11,000 ya (Blench 2006).

The most significant and well-known migration event in sub-Saharan Africa that has been associated—although not unanimously—with agricultural innovations, and at a later stage with iron technologies, is the expansion of the Bantu language family belonging to the Niger-Congo phylum (fig. 1). These languages are assumed to have originated in the Grassfields region between Cameroon and Nigeria not more than 5,000 ya and spread from this homeland throughout sub-Saharan Africa to Somalia in the East and as far as the Cape in the South (Nurse and Philippson 2003). The manner in which Bantu languages and speech communities spread throughout sub-Saharan Africa remains a matter of debate among specialists (Vansina 1979, 1995; Ehret 2001; Holden 2002; Eggert 2005; Holden and Gray 2006; Bostoen 2007). The general view of Diamond and Bellwood (2003) suggests that Bantu languages and agricultural techniques spread together with people throughout sub-Saharan Africa. However, this view is opposed by other investigators emphasizing the effect of cultural spread rather than movement of people (see Vansina 1995; Nichols 1997; Robertson and Bradley 2000). Several genetic studies that focused mainly on the uniparentally transmitted mitochondrial DNA (mtDNA) and Y chromosome are in favor of the first hypothesis, namely that the Bantu expansion was a joint linguistic and demographic event. As regards mtDNA, several haplogroups such as L0a, L2a, L3b, and L3e have been associated with the Bantu expansion (Salas et al. 2002), whereas for the Y chromosome, haplogroups E1b1a (defined by the single nucleotide polymorphism [SNP] M2) and B2b (defined by M150) have been connected to this event (cf. Thomas et al. 2000; Cruciani et al. 2002; Berniell-Lee et al. 2009). However, no differences have been detected in frequency and diversity levels of haplogroup E1b1a between Bantu and other Niger-Congo populations. In fact, not only does the geographic distribution of E1b1a extend far beyond the area settled by speakers of Bantu languages, but its frequency and the associated STR diversity are even higher in non–Bantu-speaking regions, such as Guinea Bissau (Rosa et al. 2007). In their extensive study of Y chromosomal variation in Africa, Wood et al. (2005) genotyped M191, which defines a sublineage of E1b1a called E1b1a7, which was also associated with the Bantu expansion (Zhivotovsky et al. 2004). They found a significant correlation between linguistic and Y chromosome variation, which is driven in large part by the correlation of Y chromosomal variation and the Bantu language family. They inferred that sex-biased migrations between expanding Bantu agriculturalists and hunter gatherers have notably affected the patterns of Y chromosomal variation in sub-Saharan Africa. However, this study was based on biallelic markers alone, and data from the entire south-central part of sub-Saharan Africa were lacking.

Although studies of autosomal polymorphisms are becoming more common as a result of technological advances (e.g., Hammer et al. 2008; Tishkoff et al. 2009; Bryc et al. 2010; Sikora et al. 2010), investigations of uniparental markers still offer valuable insights into human prehistory that cannot be obtained by autosomal markers alone. One advantage is the possibility to reconstruct phylogenies of mutations and to trace the origins of polymorphisms as well as their geographical spread, which is not possible with autosomal data due to recombination. Furthermore, uniparental markers greatly enable the detection of culturally determined sex-biased events, such as patrilocality or matrilocality or polygyny (cf. Kayser et al. 2006, 2008). Because patrilocality and/or polygyny are common social practices in sub-Saharan Africa (Pebley et al. 1988), the Y chromosome is expected to retain a clearer signal of demic migration events because the mtDNA and autosomes brought by marrying local women could with time dilute the original genetic composition.

The aim of this paper is to investigate in more detail the combined Y chromosomal variation of biallelic and

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

Table 1. Details of the 26 Populations Included in This Study With Approximate Geographic Coordinates.

Group	Code	Sample Size	Latitude	Longitude	Linguistic Affiliation[a]	Country[b]	References
Algeria	ALG-AA	20	32.0	3.0	Afro-Asiatic	Algeria	present study
Angola Bantu	ANG-B	230	−17.0	15.0	NC—Bantu	Angola	Coelho et al. (2009)
Botswana Bantu	BOT-B	40	−24.7	25.9	NC—Bantu	Botswana	present study
Burkina Faso Gur	BF-G	183	13.0	−1.5	NC—Gur	Burkina Faso	present study
Burkina Faso Mande	BF-M	152	12.6	-3.6	NC—Mande	Burkina Faso	present study
C.A.R. Pygmies	CAR-P	23	4.0	17.0	Various	C.A.R.	present study
Cameroon Bantu	CAM-B	28	5.0	11.0	NC—Bantu	Cameroon	Berniell-Lee et al. (2009)
Cameroon Pygmies	CAM-P	27	5.0	13.4	NC—various	Cameroon	Berniell-Lee et al. (2009)
D.R.C. Bantu	DRC-B	58	−5.0	18.8	NC—Bantu	D.R.C.	present study
D.R.C. Pygmies	DRC-P	11	1.0	29.0	Nilo-Saharan	D.R.C.	present study
Ethiopia	ETH-AA	98	9.0	38.7	Afro-Asiatic	Ethiopia	present study
Gabon Bantu	GAB-B	795	−0.7	12.0	NC—Bantu	Gabon	Berniell-Lee et al. (2009)
Gabon Pygmies	GAB-P	33	0.5	13.6	NC—Ubangi	Gabon	Berniell-Lee et al. (2009)
Kenya Bantu	KEN-B	10	−3.0	37.0	NC—Bantu	Kenya	present study
Kenya Nilo-Saharan	KEN-NS	79	0.5	36.0	Nilo-Saharan	Kenya	present study
Namibia	NAM-K	6	−21.0	20.0	Khoisan	Namibia	present study
Nigeria	NIG-Y	12	8.0	5.0	NC—Yoruboid	Nigeria	present study
Senegal	SEN-M	15	14.0	−14.0	NC—Mande	Senegal	present study
South Africa Bantu	SA-B	8	−29.0	26.0	NC—Bantu	South Africa	present study
Tanzania Afro-Asiatic	TZ-AA	25	−2.8	36.0	Afro-Asiatic	Tanzania	Tishkoff et al. (2007)
Tanzania Bantu	TZ-B	64	−4.0	33.0	NC—Bantu	Tanzania	Tishkoff et al. (2007)
Tanzania Khoisan	TZ-K	121	−3.1	34.4	Khoisan	Tanzania	Tishkoff et al. (2007)
Tanzania Nilotic	TZ-NS	31	−2.1	35.4	Nilo-Saharan	Tanzania	Tishkoff et al. (2007)
Uganda	UGA-NS	118	2.7	34.3	Nilo-Saharan	Uganda	Gomes et al. (2010)
Zambia East Bantu	ZAE-B	69	−15.5	23.0	NC—Bantu	Zambia	de Filippo et al. (2010)
Zambia West Bantu	ZAW-B	480	−12.0	31.0	NC—Bantu	Zambia	present study

NOTE.—[a]NC refers to Niger-Congo linguistic phyla.
[b] C.A.R. stands for Central African Republic and D.R.C. for Democratic Republic of Congo

microsatellite markers in sub-Saharan Africa to gain insights into (pre)historic population movements, in particular those associated with the spread of the Niger-Congo language phylum. In order to obtain a more fine-grained coverage of the Y chromosomal diversity in the continent, we analyze over 1,100 samples from several populations belonging to the major linguistic phyla in West, Central, and East Africa and combine these with published data. We analyze the distribution of subclades of the widespread E1b1a lineage to obtain a more detailed view of the genetic variation present in the Niger-Congo phylum and to investigate the potential genetic effects of the Bantu migration. Furthermore, we investigate the two main hypotheses about the spread of Bantu languages over sub-Saharan Africa: a mere cultural diffusion (so-called "language shift"; Nichols 1997 and Sikora et al. 2010) or an actual movement of people via a demic diffusion (Diamond and Bellwood 2003).

Materials and Methods

Samples

A total of 1,090 saliva samples or buccal swabs were collected from healthy male volunteers after obtaining informed consent. About 480 samples from Bantu speakers from the Western Province of Zambia were collected in 2007 by C.d.F., E.D.G., T.N., K.Bo., B.P., and M.S.; 58 samples from Bantu speakers from the Democratic Republic of Congo (D.R.C.) were collected by C.d.F., K.Bo., and Joseph Koni Muluwa in 2008; 335 samples from Burkina Faso (speaking either Niger-Congo Mande or Gur languages) were collected by M.W., H.S., and K.Be. in 2008; 40 samples from Bantu speakers from Botswana were collected by S.W.M. in 2010; 98 samples from Ethiopians speaking Afro-Asiatic languages and 79 samples of Nilo-Saharan speakers from Kenya were collected by collaborators of D.L. in 2003, 2007, and 2008. DNA was extracted from the saliva samples from Botswana, Burkina Faso, D.R.C., and Zambia following the method previously described by Quinque et al. (2006). DNA extraction from the buccal swab samples from Ethiopia and Kenya was performed following the procedure described in Miller et al. (1988).

In addition, 85 unrelated sub-Saharan African individuals from the Human Genome Diversity Cell Line Panel (Cann et al. 2002) as identified by Rosenberg (2006) were included in the analyses. These include the Biaka Pygmies from the Central African Republic (C.A.R.), Mbuti Pygmies from D.R.C., Bantu speakers from Kenya, Khoisan from Namibia, Niger-Congo Yoruba from Nigeria, Niger-Congo Mandenka from Senegal, and Bantu speakers from South Africa. Furthermore, to bolster the number of Afro-Asiatic groups included in this study, the Afro-Asiatic–speaking Mozabites from Algeria were also genotyped, even though they do not belong to the geographic region of sub-Saharan Africa as such.

For the purposes of this study, the data set has been divided into 26 major geographic and/or linguistic groups as summarized in table 1 (for details of the ethnolinguistic affiliation of the groups as determined by self-identification, see supplementary table 2, Supplementary Material online).

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

Markers

The Nilo-Saharan samples from Kenya and some of the Ethiopian samples were initially screened at the University of Bologna through restriction fragment length polymorphism analysis of the biallelic markers M42 and M60, which define the A and B lineages, respectively. The remaining 1,174 samples were genotyped for 24 SNPs (12f2, M106, M124, M145, M168, M170, M172, M174, M175, M20, M201, M207, M213, M214, M269, M45, M52, M69, M9, M91, M96, MEH2, SRY10831, and Tat) defining the major branches of the Y chromosome tree (Karafet et al. 2008). These sites were amplified in a multiplex polymerase chain reaction (PCR) and then typed by means of two SNaPshot assays consisting of 12 SNPs each following the manufacturer's specifications (Applied Biosystems, http://www3.appliedbiosystems.com). We further genotyped seven additional SNPs (M33, M35, M2, M191, M75, U174, and U175) on those individuals ascertained to be haplogroup E for a deeper characterization of this lineage (fig. 2) in an additional multiplex PCR and SNaPshot assay. Subhaplogroups of haplogroup E have been defined according to the nomenclature specified in Karafet et al. (2008): E1b1a* (xE1b1a8 and xE1b1a7), E1b1a8, E1b1a7* (xE1b1a7a), E1b1a7a, E* (xE1b1a, xE1a, xE1b, and xE2). Genotyping details are listed in supplementary table 1 (Supplementary Material online). The markers U174 and U175 were additionally typed for this study in the samples from Eastern Zambia that had previously been genotyped for the other markers (de Filippo et al. 2010). Finally, we genotyped 12 short tandem repeat (STR) loci (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) by means of the Promega Y-Powerplex kit (http://www.promega.com). When two peaks were detected in the duplicated STR locus DYS385, the smaller allele was arbitrarily assigned to DYS385a and the larger to DYS385b. Both SNP and STR genotyping were performed on the ABI 3130xl Genetic Analyzer and analyzed using the GeneMapperID v3.2 software (Applied Biosystem).

Comparative Data

In order to extend our study of Y chromosomal variation to a wider geographical coverage of sub-Saharan Africa, we included published data sets having a similar amount of SNP and STR genotype information as our data. The published data were classified on geographic and linguistic grounds as follows: Khoisan, Afro-Asiatic, Nilo-Saharan, and Bantu speakers from Tanzania (Tishkoff et al. 2007); non-Pygmy Bantu speakers and Pygmies (Bantu and non-Bantu speakers) from both Cameroon and Gabon (Berniell-Lee et al. 2009); Bantu speakers from Angola (Coelho et al. 2009); and a Nilo-Saharan group from Uganda (Gomes et al. 2010). However, these studies genotyped individuals belonging to haplogroup E only to the level of E1b1a, with the exception of Gomes et al. (2010) who additionally genotyped M191. We therefore inferred the frequency of the haplogroup E sublineages studied here—namely E1b1a8, E1b1a7a, and E1b1a7*—from the STR haplotypes using Linear Discriminant Analysis (LDA) with the R statistical software by means of the function "lda" from the package MASS (Venables and Ripley 2002). Because Tishkoff et al. (2007) and Coelho et al. (2009) subtyped only M2 and M35 on the haplogroup E samples, we also applied LDA to those individuals who were E*(xE1b1a and xE1b1b1). Of these, the individuals from Tishkoff et al. (2007) being possibly haplogroup D or E (i.e., carrying the YAP mutation) were considered as belonging to haplogroup E under the assumption that haplogroup D is virtually absent in the African continent (Jobling and Tyler-Smith 2003; Wood et al. 2005). We tested the power of LDA to reliably infer haplogroups from STR haplotype data as described in the supplementary text (Supplementary Material online) before applying it to the above-mentioned data sets. However, it should be kept in mind that the comparative data inferred by LDA may not be as reliable as our genotyped data.

Data Analyses

Standard measures of genetic diversity, pairwise genetic distances between groups expressed as R_{ST} and proportion of haplotypes not shared were calculated in *R*. Correspondence analysis (CA) of haplogroup frequencies in all populations was performed using the function "ca" from the R package ca (Nenadic and Greenacre 2007). Analysis of molecular variance (AMOVA) and pairwise F_{ST} between groups were carried out with Arlequin software v3.1 (Excoffier et al. 2005) based on haplogroup frequencies. A matrix of geographic great circle distances between all groups (with the exclusion of populations with less than ten individuals) was generated. We performed a Mantel test (Z value) to investigate whether the geographic distances are correlated with genetic distances. Individuals who had STR missing values were excluded from some analyses.

Patterns of haplotype sharing among groups were explored as follows. STR haplotypes that were shared among at least three groups were ranked based on their frequency in the entire combined data set. We explored the distribution of shared haplotypes among groups that were merged (here called metagroups) according to their geographic location as well as their linguistic affiliation (and ethnicity in the case of the Pygmies, who are known to have acquired their language from their agriculturalist neighbors). With regard to linguistic affiliation, individuals from Western Zambia who speak a language belonging to the Eastern Bantu branch (Fortune 1970; Bostoen 2009) were classified with the Bantu speech communities from Eastern Zambia. To test if the observed patterns simply reflect sample size differences among the various metagroups, we randomly assigned the shared haplotypes to groups and subsequently merged the groups into the various metagroups. We repeated this process 1,000 times and recorded the number of haplotypes shared between each pairwise comparison of metagroups to estimate the significance level.

The average squared distance (ASD) statistic (Goldstein and Pollock 1997) was calculated to estimate the time since

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

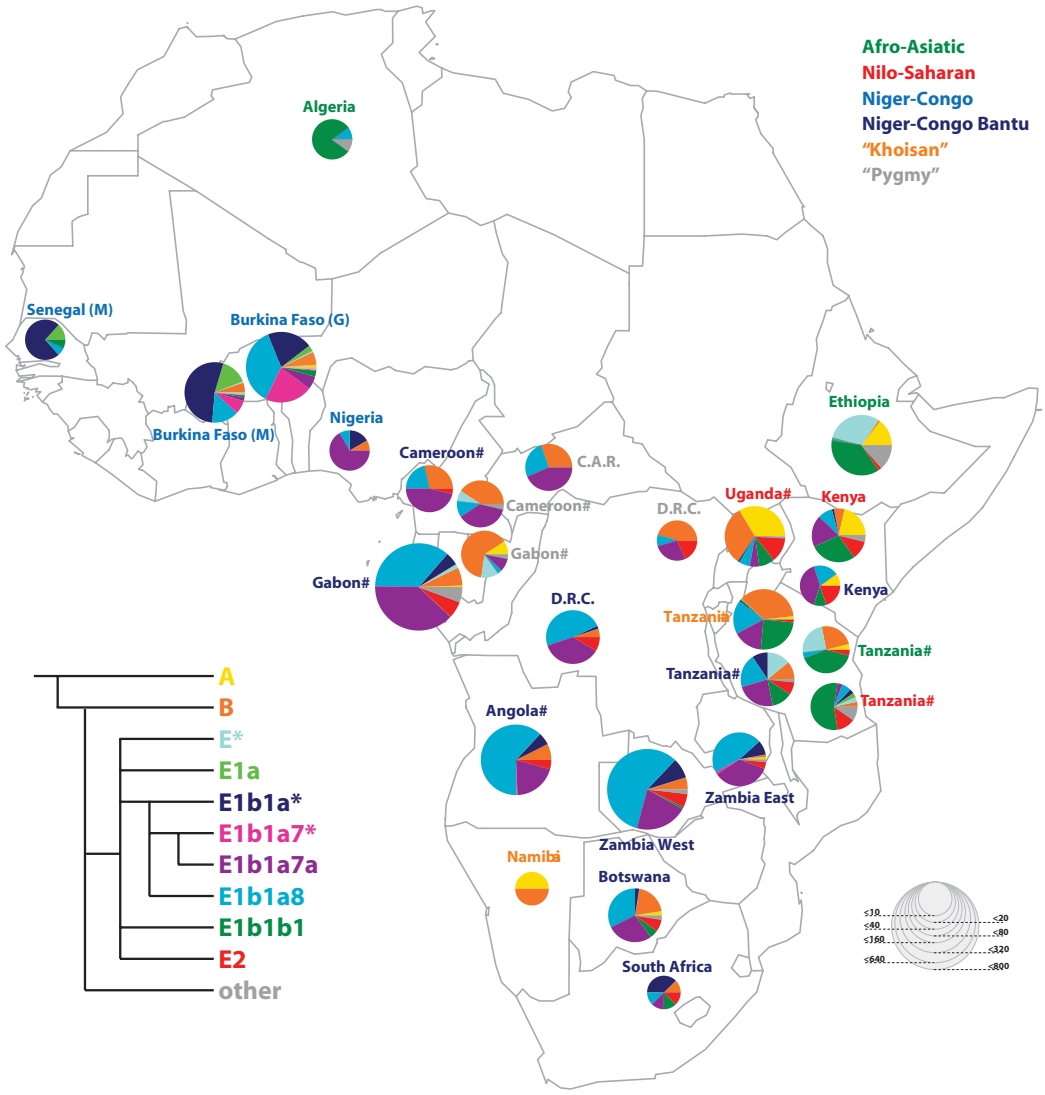


Fig. 2. Haplogroup composition of the combined data set. The size of the pie charts is proportional to the sample size as shown in the bottom right. Groups marked with # indicate that the subhaplogroup composition of E1b1a was inferred by LDA. Only the major African haplogroups (A, B, and subhaplogroups of E) are displayed; the remaining haplogroups are lumped under the label "other." Population labels are color coded according to linguistic phyla as indicated in the upper right, with Pygmy groups (gray) indicated separately from other groups.

the most recent common ancestor (tMRCA) for ten microsatellites (excluding DYS385a/b). Under the Stepwise Mutation Model, the tMRCA is expected to be ASD/2μ, where μ is the mutation rate per generation per locus, averaged across loci. Therefore, to calculate the tMRCA and associated confidence intervals (CI), the mutation rates reported in the Y-STR haplotype reference database (http://www.yhrd.org) were used, and a generation time of 25 years was considered.

Because Pygmy groups are commonly believed to have shifted from their original language to that of their agricultural neighbors, which makes their current linguistic affiliation misleading, they were considered as a separate ethnic unit, regardless of the language they speak, and excluded from the AMOVA analysis.

Results

Y Chromosome Haplogroups in Sub-Saharan Africa

Figure 2 shows the haplogroup composition for 2,736 samples belonging to 26 groups (see references in table 1). STR haplotypes and SNP haplogroups genotyped here as well as those inferred by LDA (with associated relative posterior probabilities) are reported in supplementary table 2 (Supplementary Material online) and the phylogenetic relationships of the SNPs typed are in supplementary figure

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

3 (Supplementary Material online). Overall, the haplogroup composition in all the groups reflects what has been previously observed in the African continent, with A (mainly present in Khoisan speakers and Eastern groups), B (mainly found in hunter-gatherer Pygmies and Khoisan as well as their neighbors), and E (in almost all groups) representing the majority (87%) of the haplogroups.

Haplogroup E1b1a (including all its sublineages typed in this study) is present in all groups (excluding the Namibian Khoisan) and was found at a frequency of ~68.5% in the entire data set. This is in agreement with previous studies of African Y chromosomal variation (Wood et al. 2005; Tishkoff et al. 2007; Berniell-Lee et al. 2009). With respect to the sublineages of E1b1a typed here, the most frequent haplogroup in the combined data set was E1b1a8 (~35%), which was found in all groups except in the Namibian Khoisan (which are, however, represented by only six individuals). All Bantu-speaking groups showed relatively high frequencies of this haplogroup, ranging from 18% to 62%, with the exception of the South African Bantu where the frequency was only 12.5%; however, this is due to the small sample size and not significantly different from the other groups (95% CI of sampling error = 3–53%). The second most common haplogroup, E1b1a7a, is present in African populations with an average frequency of 23% and shows moderately high frequencies in all Bantu and Pygmy groups. The highest frequencies are found in Nigeria (67%) and Bantu speakers from Cameroon (46%), which are both regions that are close to the putative homeland of the Bantu languages.

Another common haplogroup within haplogroup E is E1b1a* (xE1b1a8 and xE1b1a7) with an average frequency of 8.9%, which is a characteristic of all West African groups included here, with the highest frequencies in Mande speakers from Senegal (75%) and Burkina Faso (53%). Haplogroup B is also widespread, being found on average in 10.3% of the African groups included here.

Patterns of Y-STRs Diversity

Y-STR diversity values within specific haplogroups can be informative for discerning origins and migrations of haplogroups: In general, the highest diversity should be found in the population where the haplogroup originated, and lower diversity (due to successive founder events) may be associated with migrations. However, because STRs have a high mutation rate, these signals might be erased over time, and it can be insightful to examine the variance in repeat units. The STR variance has been described as evolutionary more stable and is correlated with the time that has elapsed since a haplogroup-defining mutation arose, thus serving as a rough estimator of tMRCA as well (Goldstein and Pollock 1997; Bosch et al. 1999). Yet, because the results of such estimates depend to a large extent on the mutation rates used, which are very variable and subject to considerable debate (Zhivotovsky et al. 2004), age estimations should be considered with due caution.

Values of diversity for 11 STR loci for all individuals as well as those carrying the E1b1a*, E1b1a7a, and E1b1a8 clades are reported in table 2. In general, regardless of the haplogroup composition and excluding populations with sample size less than ten individuals, Niger-Congo–speaking groups have slightly higher haplotype diversity than Nilo-Saharan–speaking groups (Mann–Whitney U test: $W = 27$, P value < 0.017), but these together have higher diversity values than Afro-Asiatic, Khoisan, and Pygmy groups ($W = 123.5$, P value < 0.005).

The STR haplotype diversity associated with E1b1a8 was found to be higher ($W = 45$, P value < 0.004) in all Bantu-speaking groups (except in Cameroon with a low sample size = 6) than all the other groups after removing groups with less than five individuals. However, the STR variance showed a different pattern with the highest values in Pygmies from C.A.R. (with a sample size of six) and Burkina Mande, whereas reduced values were found in all the other groups. Moreover, STR variances did not differ significantly among groups ($W = 24$, P value $= 0.526$).

For haplogroup E1b1a7a, the STR haplotype diversity levels were high (>0.90) in all groups, with the lowest values observed in Pygmies from C.A.R., Tanzanian "Khoisan," and the two Nilo-Saharan groups. Similar to E1b1a8, the highest STR variance for E1b1a7a was found in the C.A.R. Pygmies (0.49); however, the Bantu speakers from West Zambia and the Burkina Faso Gur speakers also had high STR variances (0.47 and 0.43, respectively).

With regard to the diversity associated with haplogroup E1b1a*, Niger-Congo non-Bantu have higher haplotype diversity and STR variance than the Bantu-speaking groups. Overall, there is some support for an association of E1b1a8 with higher diversity in Bantu-speaking groups and of E1b1a* with higher diversity in Niger-Congo non–Bantu-speaking groups. However, none of these patterns reach statistical significance: for E1b1a8 $W = 54$, P value $= 0.125$ and for E1b1a7a $W = 20$, P value $= 0.057$.

The tMRCA estimates for haplogroups E1b1a7 and E1b1a8 were calculated by means of the ASD statistic for the major ethno-linguistic groups (table 3). The highest tMRCA (~4,200 ya) for E1b1a7a was ascertained in the Yoruba from Nigeria, whereas the lowest (~2,000 ya) was in the Nilo-Saharans. With regard to E1b1a8, the highest tMRCA (~ 5,000 ya) was found in Mande speakers from both Burkina Faso and Senegal, whereas the lowest (~3,400 ya) was in the Bantu. The 95% CIs all overlap; overall, all these estimates are consistent with the time of the Bantu expansion (5,000–3,000 ya) and with an origin of both haplogroups in an area between West and Central Africa a few thousand years before the beginning of the expansion as indicated by the upper limits of the CIs.

Genetic Structure Within and Between Groups in Sub-Saharan Africa

To visualize the relationships among the different groups within sub-Saharan Africa, a CA was performed on the haplogroup frequencies (fig. 3). The first two dimensions together accounted for 59.2% of the total inertia and reflect both geographic and linguistic groupings. In the first dimension, the Niger-Congo–speaking groups and

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

Table 2. Diversity Values Based on 11 Y-STR Loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b), Where *N* is the Sample Size, HD is the Haplotype Diversity with its Standard Deviation (SD), and STR Var is the Variance of Repeat Units Averaged Across All 11 STR loci.

	ALL			E1b1a8			E1b1a7a			E1b1a*		
Group[a]	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var	*N*	HD (SD)	STR var
Bantu speakers												
ANG-B	230	0.992 (0.002)	1.35	143	0.982 (0.005)	0.32	46	0.987 (0.009)	0.36	13	0.962 (0.041)	0.38
BOT-B	39	0.993 (0.007)	2.57	13	1.000 (0.030)	0.40	10	0.933 (0.62)	0.19	1	—	—
CAM-B	28	0.992 (0.012)	3.63	6	0.933 (0.122)	0.31	13	0.987 (0.035)	0.27	0	—	—
DRC-B	43	0.992 (0.007)	0.97	21	0.990 (0.018)	0.39	16	0.967 (0.036)	0.26	0	—	—
GAB-B	795	0.997 (0.000)	1.67	289	0.993 (0.001)	0.39	303	0.992 (0.002)	0.39	39	0.966 (0.014)	0.40
KEN-B	10	1.000 (0.045)	1.56	2	1.000 (0.500)	0.41	4	1.000 (0.177)	0.11	0	—	—
SAB	8	1.000 (0.063)	2.72	1	—	—	1	—	—	3	1.000 (0.272)	0.33
TZ-B	64	0.999 (0.003)	2.67	13	0.987 (0.035)	0.32	15	0.990 (0.028)	0.35	6	0.933 (0.122)	5.65
ZAW-B	473	0.995 (0.001)	1.12	277	0.987 (0.002)	0.30	100	0.995 (0.002)	0.47	37	0.964 (0.018)	0.25
ZAE-B	69	0.997 (0.003)	0.83	32	0.992 (0.011)	0.44	24	0.989 (0.017)	0.30	6	1.000 (0.096)	0.35
Niger-Congo non-Bantu speakers												
BF-G	173	0.994 (0.002)	1.32	65	0.973 (0.010)	0.46	11	1.000 (0.039)	0.43	36	0.992 (0.009)	0.70
BF-M	148	0.988 (0.004)	1.28	21	0.981 (0.023)	0.74	2	1.000 (0.500)	0.50	81	0.972 (0.012)	0.57
NIG-Y	12	1.000 (0.034)	1.34	1	—	—	8	1.000 (0.063)	0.34	2	1.000 (0.500)	0.55
SEN-M	15	0.990 (0.028)	0.81	1	—	—	0	—	—	11	0.982 (0.046)	0.55
Hunter gatherers												
CAM-P	27	0.980 (0.016)	4.08	3	1.000 (0.222)	0.14	10	0.956 (0.059)	0.23	0	—	—
CAR-P	23	0.964 (0.022)	4.41	6	0.800 (0.237)	0.84	10	0.911 (0.077)	0.49	0	—	—
DRC-P	11	0.964 (0.051)	4.36	1	—	—	3	1.000 (0.272)	0.45	0	—	—
GAB-P	33	0.936 (0.026)	4.00	1	—	—	3	0.667 (0.314)	0.12	0	—	—
NAM-K	4	1.000 (0.177)	2.89	0	—	—	0	—	—	0	—	—
TZ-K	121	0.982 (0.004)	2.51	22	0.970 (0.024)	0.27	19	0.936 (0.037)	0.33	1	—	—
Nilo-Saharan												
KEN-NS	45	0.990 (0.007)	1.31	6	0.800 (0.172)	0.21	10	0.933 (0.062)	0.24	0	—	—
TZ-NS	31	0.991 (0.012)	5.79	2	1.000 (0.500)	0.27	1	—	—	1	—	—
UGA-NS	118	0.988 (0.003)	2.24	7	0.905 (0.103)	0.27	6	0.933 (0.122)	0.18	1	—	—
Afro-Asiatic												
ALG-AA	20	0.963 (0.033)	0.93	2	1.000 (0.500)	0.05	0	—	—	0	—	—
ETH-AA	64	0.980 (0.007)	1.02	0	—	—	0	—	—	0	—	—
TZ-AA	25	0.963 (0.021)	6.53	1	—	—	0	—	—	0	—	—

NOTE.—[a] The group codes correspond to those reported in table 1.

Pygmies (except those from Gabon) all have values less than 0.5, and all other groups have values greater than 0.5. The Afro-Asiatic groups cluster together and the Nilo-Saharan groups from Kenya, Uganda, and Tanzania are also located close to each other along the first dimension. The eastern Bantu speakers from Tanzania (and to a minor extent from Kenya) are closer to the other East African populations than are the other Bantu–speaking groups as a result of their modest frequencies of haplogroups A and E*, respectively. Dimension 2 largely divides the Niger-Congo populations into Bantu and non-Bantu, with the Western samples (Senegal and Burkina Faso) with highest values, driven by haplogroups E1a, E1b1a7*, and E1b1a*.

Table 3. Estimates of tMRCA (in years ago) of the Two Major Haplogroups (E1b1a7a and E1b1a8) Using ASD Statistic With 10 STRs (excluding DYS385a/b) and a Generation Time of 25 Years.

	E1b1a7a			E1b1a8		
Groups	*N*[a]	Mean	95% CI	*N*[a]	Mean	95% CI
NC—Bantu	532	3,238	2,022–6,792	798	3,396	1,933–8,951
NC—Gur	11	2,583	1,806–3,917	65	3,458	2,444–5,543
NC—Mande	2	—	—	22	4,987	3,164–10,281
NC—Yoruba	8	4,249	2,498–10,181	1	—	—
Pygmies	26	3,707	2,629–5,468	11	3,889	2,298–10,205
Khoisan	19	2,396	1,608–3,831	22	3,484	1,771–11,263
Nilo-Saharan	17	2,049	1,326–3,595	15	4,066	2,068–12,288

NOTE.—[a]Number of STR-haplotypes used.

To test whether the genetic structure was in better accordance with linguistic or geographic groupings, AMOVA analyses were performed (table 4). As mentioned in the Methods section, the four Pygmy populations were excluded from these analyses because of their assumed recent language shift. Both linguistic affiliation and geographic location are in good agreement with the Y chromosomal variation because the variance between groups is always higher than that between populations within a group. The variance among all the populations included in the study accounts for 15.4% of the total. When these are grouped according to their classification in one of the four major linguistic phyla, the between-group variability reaches 14.8%, whereas the variance within the linguistically defined groups is 8.7%. Grouping populations by geography into North, West, East, Central, and South Africa decreased the between-group variability to 9.96% and the variance within groups to 6.75%. When only

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

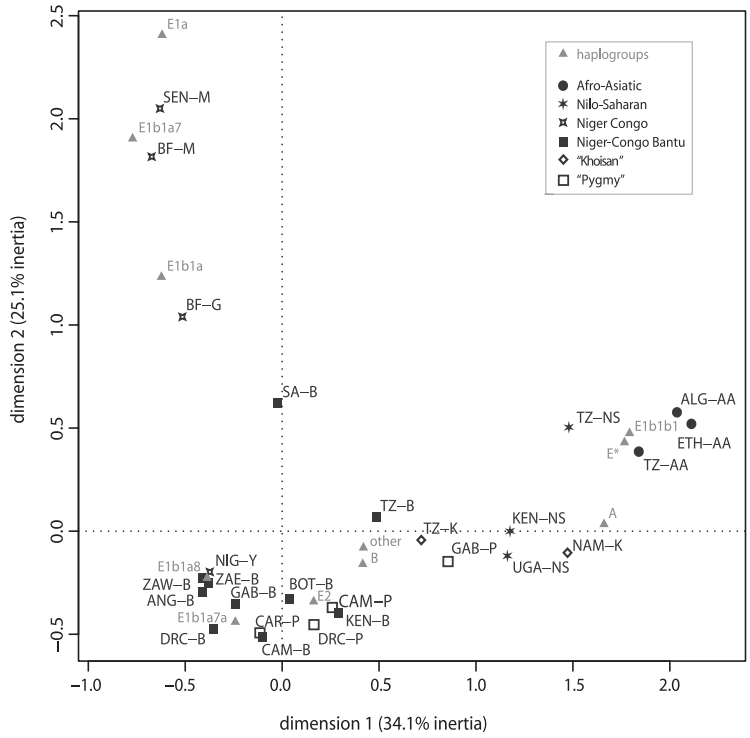


Fig. 3. Correspondence analysis performed on haplogroup frequencies. The population labels correspond to those reported in table 1.

Bantu-speaking populations were compared, the proportion of variance explained by differences between populations is much lower but still significant (4.7%, *P* value = 0).

We performed another AMOVA to quantify the differences between Niger-Congo, non-Bantu, and Bantu populations (see fig. 1). This highlighted a large amount of variation (11.6%, *P* value < 0.018) due to differences among groups and only 5.31% within groups. When performing this AMOVA with the lower haplogroup resolution used in previous studies (e.g., Wood et al. 2005)—that is, only E1b1a*(xE1b1a7) and E1b1a7 without their subhaplogroups E1b1a8 and E1b1a7a—the proportion of variation observed between Bantu and non-Bantu became nonsignificant (0.28%, *P* value = 0.35). This is a strong indication that the more fine-grained haplogroup genotyping used here adds considerably to our power to detect genetic substructure in Africa.

Mantel tests of correlation between geographic and genetic distances further confirmed that geography has had an important influence on Y chromosomal diversity

Table 4. AMOVA Based on Haplogroup Frequencies.

			Proportion of variation (%)		
Number of Groups	Grouping[a]	Total Number of Populations	Among Groups	Among Populations Within Group	Within Populations
1	**All populations**	**22**	**—**	**15.39****	**84.61****
1	**Bantu**	**10**	**—**	**4.69****	**95.31****
5	**Geography[b]**	**22**	**9.96****	**6.75****	**83.29****
4	**Language[c]**	**22**	**14.08****	**8.68****	**77.24****
2	**Niger-Congo[d]**	**14**	**11.58***	**5.31****	**83.10****
2	**Niger-Congo (low)[e]**	**14**	**0.28**	**5.67****	**94.06****

NOTE.—All values are significant with *P* value < 0.05* and *P* value < 0.01**, except for that in boldface.

[a] Pygmy groups were excluded because they are known to have undergone language shift.

[b] Geographic subdivision as follows: North (Algeria), West (Senegal, Burkina Faso, and Nigeria), Central (Cameroon, D.R.C., and Gabon), East (Ethiopia, Kenya, Tanzania, and Uganda), and South (Angola, Zambia, Botswana, Namibia, and South Africa).

[c] Linguistic grouping with the four major African phyla: Afro-Asiatic, "Khoisan," Niger-Congo, and Nilo-Saharan.

[d] Niger-Congo Bantu vs. non-Bantu.

[e] Niger-Congo Bantu vs. non-Bantu with a lower haplogroup resolution: E1b1a*(xE1b1a7) and E1b1a7. See main text for details.

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

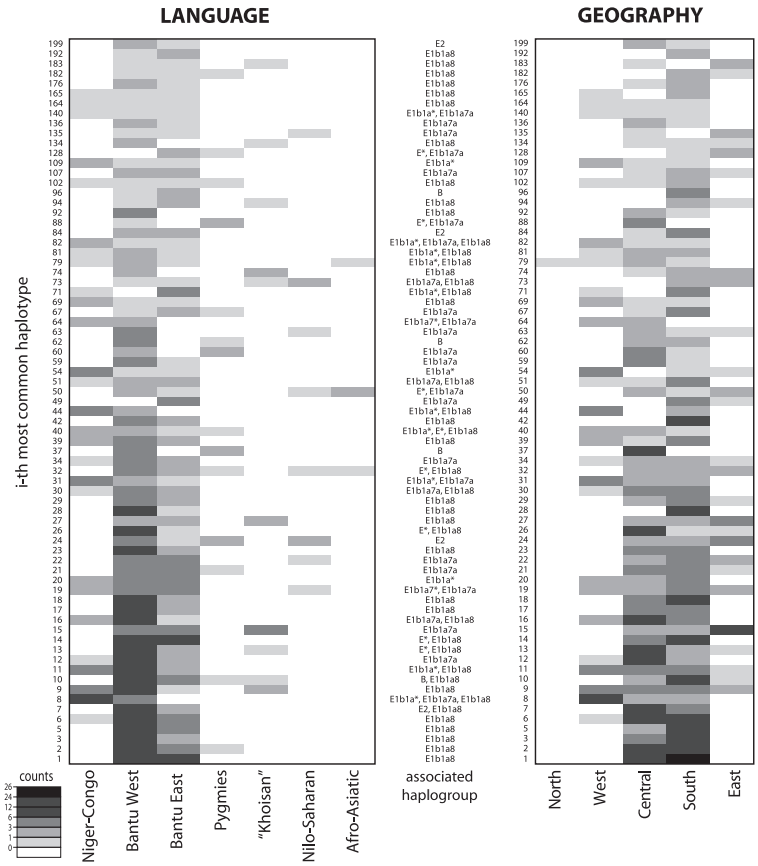


Fig. 4. Patterns of haplotype sharing. Heat plots showing the count of the most common haplotypes from 11 STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b) shared among at least three individual groups. Individual groups are combined into metagroups according to their linguistic affiliation (left) and geographic location (right); the same heat plot, but for single groups, is reported in supplementary figure 5 (Supplementary Material online).

in Africa. Indeed, both pairwise F_{ST} and R_{ST} matrices were correlated with the matrix of great circle geographic distances: $Z = 0.47$ (one-tail P value < 0.001) and 0.26 (one-tail p value < 0.015), respectively. When only Niger-Congo groups were considered, F_{ST} values were correlated with geography ($Z = 0.50$, one-tail P value < 0.001), but R_{ST} values were not ($Z = -0.02$, one-tail P value $= 0.51$). In contrast, the correlation of R_{ST} and geographic distances was still present when all the other groups (excluding Niger-Congo) were considered. In addition, pairwise R_{ST} values between groups were calculated for haplogroups E1b1a7a, E1b1a8, and E1b1a* and compared with the geographic distances between them. Only R_{ST} values associated with haplogroups E1b1a8 and E1b1a* exhibited a correlation with geographic distances, with $Z = 0.36$ (one-tail P value < 0.03) and 0.67 (one-tail P value $= 0.034$), respectively. However, because the dimension of the matrices might have an effect on the significance of the Mantel test, we controlled for the number of groups by redoing the test using only those groups that have both E1b1a7a and E1b1a8. In this test, no correlation was observed between geographic distances and pairwise R_{ST} for either haplogroups E1b1a7a or E1b1a8.

Distribution of Shared Haplotypes

Contrary to the geographical and linguistic structure apparent in the haplogroup data, a network based on 11 STR loci showed no structure at all; rather, haplotypes from East African and Central African Bantu groups are found clustered together. The extensive reticulation made it difficult to observe any patterns of overall haplotype sharing (supplementary fig. 4, Supplementary Material online). Therefore, in order to elucidate the relationships among groups from different geographic areas that may be due to common origin and/or recent migration, the combined data set was screened for widespread and shared haplotypes. Figure 4 shows the distribution of shared haplotypes among groups that were merged (here called metagroups as described in the Material and Methods), whereas the haplotype-sharing patterns for individual populations are shown in supplementary figure 5 (Supplementary Material online). The total number of haplotypes shared by at least

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

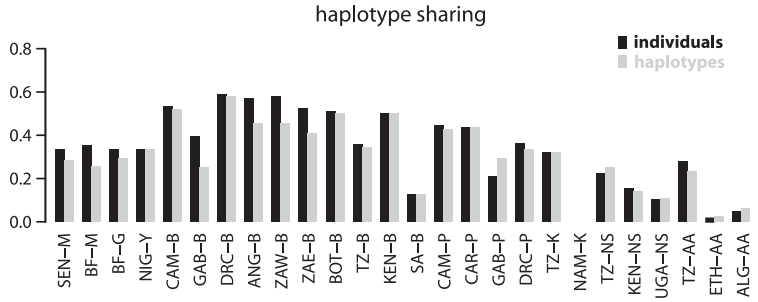


FIG. 5. Proportion of shared haplotypes. Histogram of the proportion of shared haplotypes between one group and all other groups based on 11 STRs. Black bars represent the proportion of all individuals sharing their haplotype (with any of the other groups) over the total number of individuals in a group; gray bars represent the proportion of unique shared haplotypes over the total number of haplotypes detected in a group.

three groups was 73, which is significantly less than expected if individuals are assigned to groups at random (mean = 166, range = 152−183; P value < 0.001 based on 1,000 permutations). This analysis indicates that there is a significant effect of population structure on the shared haplotypes and also indicates that the observed pattern was not caused by differences in group sample sizes. None of the 73 shared haplotypes was shared across all the metagroups. Also, no haplotype was found in all the groups within each metagroup (supplementary fig. 5, Supplementary Material online).

When grouped according to linguistic/ethnic affiliation, the West Bantu metagroup, which includes samples from Cameroon, Gabon, D.R.C., Angola, and Western Zambia and corresponds to the majority of the data set, shares 69 of 73 haplotypes with at least one of the other metagroups. Nilo-Saharan and Afro-Asiatic groups shared a low proportion of haplotypes with all other groups, ranging from 1 to 8 and from 0 to 3, respectively.

When grouped according to geography, the Southern and Central African metagroups share the most haplotypes (55), with fewer haplotypes shared between Central and Western Africa (23), Central and Eastern Africa (21), or Western and Southern Africa (26). The presence of significant structure detectable in this analysis in the STR data (which are subject to different patterns of mutation and variation as compared with the more stable haplogroup data) contrasts with the lack of structure in the network but is in good accordance with the results seen in the CA and AMOVA. This provides further indication that the inferred haplogroup frequencies are fairly accurate because the STR data were all genotyped.

To what extent do these haplotype-sharing patterns (fig. 4) simply reflect sample size differences among the various metagroups? The results of our permutation test (described in the Material and Methods and shown in supplementary table 3, Supplementary Material online) indicate that for the linguistic metagroups, the Western and Eastern Bantu do share more haplotypes than expected by chance, whereas the Niger-Congo (non-Bantu) shares significantly fewer haplotypes than expected by chance with the Pygmy, Nilo-Saharan, and Afro-Asiatic metagroups. Similarly, for the geographic metagroups, there is significantly more sharing between Central and Southern Africa and significantly less sharing between Eastern Africa and all other groups (except Southern Africa). Overall, this test demonstrates that the haplotype-sharing patterns in figure 4 do indicate population relationships and not just overall sample size differences between metagroups. In particular, there is more haplotype sharing than expected by chance involving groups toward the center of Africa (i.e., Western and Eastern Bantu and Central and Southern Africa). Moreover, the Bantu from D.R.C.—who are located in the center of the geographic area studied herein (fig. 2) and who are on average closest geographically (≈2,022 km) to all other African populations—shows the highest proportion of shared haplotypes with other groups (fig. 5).

Discussion

Haplogroup Variation Within Niger-Congo Speech Communities and Sub-Saharan Africa

The Niger-Congo phylum is one of the major language groups in the world and is the largest in the African continent in terms of number of languages, number of speakers, and geographical area it covers. To a certain extent, the linguistic branching pattern displayed in figure 1 is paralleled by Y chromosomal markers characteristic of the different subgroups of the Niger-Congo phylum included here: Mande, Gur, and Bantu. Indeed, haplogroups E1b1a* and its derivative E1b1a8 are characteristic of the Mande, which belong to the earliest split of the linguistic tree. The derived haplogroup E1b1a7* is characteristic of Gur speakers, and the most derived haplogroup analyzed here, E1b1a7a, is characteristic of Bantu-speaking groups, who represent one of the most derived branches of the Niger-Congo linguistic tree.

Although previous genetic studies on Y chromosome variation have linked haplogroup E1b1a and its sublineage E1b1a7 (when genotyped) specifically to the Bantu

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

expansion (Thomas et al. 2000; Cruciani et al. 2002; Zhivotovsky et al. 2004; Wood et al. 2005; Berniell-Lee et al. 2009), our results demonstrate that this association extends to all of Niger-Congo, not just Bantu. Indeed, E1b1a does not differ in frequency between Niger-Congo non-Bantu and Bantu, and this is also true if E1b1a7 is taken into account. In fact, an AMOVA with the haplogroup resolution used previously (Wood et al. 2005), that is, only E1b1a*(xE1b1a7) and E1b1a7—for Bantu versus Niger-Congo non-Bantu results in nonsignificant variation (0.28%, P value = 0.35) between these two groups. Therefore, to increase resolution, we for the first time analyzed two additional markers (U174 and U175) in a large number of African populations, resulting in a total of four E1b1a sublineages. Notably, the AMOVA carried out with this increased haplogroup resolution now finds significant variation between Bantu and Niger-Congo non-Bantu (11.58%, P value < 0.018). In addition, with these new markers, we were able to detect the presence of substructure even within the Niger-Congo non−Bantu-speaking groups as described below.

Niger-Congo non−Bantu-speaking groups in West Africa are distinct from Bantu speakers and groups belonging to the other African phyla as shown in the CA plot (fig. 3). This distinct position is mainly driven by haplogroup E1b1a* (almost absent in all non−Niger-Congo groups), which has high frequencies in Mande speakers and exhibits a clinal reduction from western toward eastern and southern Africa. A strong positive correlation was ascertained between the haplotype diversity levels and STR variance associated with E1b1a*. These results suggest that this haplogroup was present for a longer time in Western Africa—which is the presumed place of origin of the defining M2 mutation (Rosa et al. 2007)—and that two of the derived mutations considered here (e.g., M191 and U174) did not occur in the ancestors of the Mande; the low frequencies of E1b1a7a found in these groups could be due to later admixture. On the other hand, only Gur speakers are characterized by the presence of haplogroup E1b1a7*, which was previously associated with the Bantu expansion with a probable origin in western Central Africa (Underhill et al. 2000; Cruciani et al. 2002; Zhivotovsky et al. 2004; Wood et al. 2005) and that here we found practically restricted to Burkina Faso. Instead, a new sublineage of E1b1a7, namely E1b1a7a, which may also have originated in western Central Africa, is associated with the Bantu expansion. Indeed, we found that this marker has its highest frequencies in Nigerian Yoruba (where this haplogroup also appears to be oldest, with an estimated tMRCA of ~4,200 ya, cf. table 3) and Cameroonian Bantu speakers, both of whom are located close to the homeland of the Bantu languages. Furthermore, for other studies reporting high frequencies of M191 in Bantu-speaking groups, we suggest that those individuals are likely to harbor the derived mutation U174 (see, e.g., Appendix A in Wood et al. 2005). This is confirmed by the results of the LDA for the Ugandan data set where all individuals who had been genotyped as E1b1a7 were inferred to belong to E1b1a7a.

Bantu and non−Bantu-speaking groups can be distinguished by a second haplogroup, namely E1b1a8. However, we could not associate it unambiguously with the Bantu populations because the highest tMRCA estimate (~5,000 ya, table 3) was found in the Mande-speaking group and it also is found at high frequency in the Burkina Faso Gur speakers and in other western Central African populations (cf. table 1 in Veeramah et al. 2010). Nevertheless, we believe that further subtyping of markers on the background of U175 might reveal new insights concerning its association with Bantu-speaking groups (as we found with U174). Likewise, the discovery of further subclades within E1b1a7 and E1b1a8 might add more structure to the data and erase this apparent homogeneity of the Bantu groups.

The presence of both E1b1a7a and E1b1a8 in all Pygmy groups—directly genotyped in the C.A.R. and D.R.C. Pygmies and inferred from STR data for the Cameroon and Gabon Pygmies—may be the result of sex-biased migrations between agriculturalist and hunter-gatherer societies, where paternal lineages move from the former into the latter (Destro-Bisol et al. 2004; Tishkoff et al. 2007; Quintana-Murci et al. 2008). However, judging from the networks for both haplogroups (supplementary fig. 4, Supplementary Material online), recent admixture with Bantu-speaking neighbors may not account for the origin of all of these haplotypes. Although some haplotypes are shared with, or differ by only a few mutational steps from, Bantu speakers and hence may indeed reflect recent admixture, other haplotypes found at the periphery of the network are unique to Pygmies. The Pygmy groups tend to exhibit high levels of STR variance along with low levels of haplotype diversity, indicating the presence of a few very divergent (and therefore probably old) lineages. The older age of E1b1a8 in Pygmies than in Bantu, in contrast to the similar age of E1b1a7a in both Pygmies and Bantu (table 3), suggests the possibility that a few individuals belonging to haplogroup E1b1a8 were present in Pygmies prior to their contact with Bantu-speaking groups; individuals belonging to E1b1a7a were introduced at an early stage of the expansion (for instance, when the Bantu agriculturalist started to explore the rain forest), with later introgression of new haplotypes of both haplogroups after contact. Furthermore, this scenario of E1b1a7a introgression may have been mirrored on the Western side of sub-Saharan Africa as indicated by the young tMRCA estimate in Gur from Burkina Faso (table 3).

Overall, the distribution of the four E1b1a sublineages reflects what has been suggested from historical linguistic studies about the prehistory of Niger-Congo languages that had "[. . .] a long standing epicenter of spread in West Africa, with spreads through the forest and well to the south" (Nichols 1997).

Eastern Africa exhibits distinct patterns of Y chromosome haplogroups compared with Western and Central Africa. Eastern African Nilo-Saharan and Afro-Asiatic groups are characterized in general by high frequencies of lineages A and B as well as E* and E1b1b1, leading to

de Filippo et al. · doi:10.1093/molbev/msq312 MBE

their clustering in the CA plot (fig. 3). The inclusion of Algeria as an additional Afro-Asiatic−speaking group, even though it is located outside sub-Saharan Africa, confirms that E1b1b1 is characteristic of Afro-Asiatic−speaking populations. It has been suggested that this marker may have spread with agropastoralist migrations from their putative origin in East Africa toward Northern Africa (Cruciani et al. 2002; Arredi et al. 2004) and Southern-Central Africa (Henn et al. 2008). In this study, E1b1b1 is absent in Angola and present at only very low frequency (<1%) in our Zambian sample but is found in appreciable frequency in Botswana (5%). This raises the question whether the demic diffusion of pastoralism from Eastern to Southern Africa followed an eastern route that circumvented Angola and Zambia or whether the later arrival of Bantu-speaking groups replaced the former pastoralist populations in Angola and Zambia but not Botswana. Investigations of samples from southeastern Africa (e.g., Mozambique and Zimbabwe) are needed to disentangle these questions.

The Nilo-Saharan samples also have relatively high frequencies of haplogroup E2. Both E2 and E1b1b1 are also common in eastern Bantu speakers, and E2 is additionally found in the D.R.C. Pygmies, possibly introduced by contact with neighboring populations. Finally, another haplogroup found in relatively high frequencies in some of the East African groups (but also present in Cameroon and Gabon Pygmies) is E*. However, because this haplogroup is defined not by a shared derived allele but by the absence of derived alleles, we cannot exclude that these individuals belong to sublineages of M96 not tested here.

In general, a similar pattern of haplogroup composition is characteristic of all neighboring groups of Eastern Africa. This appears to suggest gene flow between the groups regardless of their language; however, the low number of shared haplotypes (fig. 4) in the area (especially between eastern Bantu from Kenya and Tanzania and the Nilo-Saharan and Afro-Asiatic groups) indicates little recent contact. Possibly, the similarities in haplogroup composition are an indication of more ancient contact.

Pattern of Diversity and the Bantu Expansion(s)

In contrast to the structure observable at the level of Y chromosomal haplogroups, there is a notable absence of structure at the resolution of STR markers. There is no obvious geographic patterning to the networks (supplementary fig. 4, Supplementary Material online); in particular, haplotypes are widely shared, especially between Eastern and Western Bantu-speaking groups. There are also no clear patterns of clinal reduction in haplotype diversity and STR variance for both haplogroups E1b1a7a and E1b1a8 in the Bantu speakers (contrary to other studies, e.g., Pereira et al. 2002) as would be expected with a serial founder event of male lineages expanding from their homeland throughout sub-Saharan Africa. These data might seem to contradict the most widely cited model of the Bantu expansion, which involves the joint movement of people and language together with the diffusion of agriculture (Diamond and Bellwood 2003). However, this model has been called into question not only by linguists (Nichols 1997) and historians (Vansina 1995) but also in a recent genetic study on ~2,800 autosomal SNPs (Sikora et al. 2010). Although Nichols (1997) and Sikora et al. (2010) assert that the Bantu expansion could rather have taken place by cultural diffusion alone (i.e., "language shift" where the original inhabitants of sub-Saharan Africa would have adopted a Bantu language without major immigration of Bantu peoples), Vansina (1995) calls into question the overly simplistic assumptions of either population replacement or language shift. However, although our data do not provide evidence for the serial founder effect expected by a migration of peoples over long geographical distances—with levels of diversity (e.g., haplotype diversity and STR variance; see table 3) reduced proportionally to the distance from the homeland—the overall genetic homogeneity of the Bantu-speaking groups included here and the widespread sharing of haplotypes on the background of E1b1a7a and E1b1a8 reject the hypothesis of mere cultural diffusion. Under this assumption, one would expect greater differences between geographically distant groups because they would have developed in situ for a long time. The overall genetic homogeneity of Bantu-speaking groups was also detected in a recent study of a large number of autosomal STR loci in a large number of African populations (Tishkoff et al. 2009), although the most widespread ancestry component derived from STRUCTURE analysis extended beyond Bantu-speaking groups to include all Niger-Congo groups. Another factor to be considered is the recent time of this expansion suggested to be 3,000−5,000 ya (Blench 2006), which would reduce the accumulation of variability and structure among populations. The tMRCAs estimated for the sublineages E1b1a7a and E1b1a8 are in accordance with a recent expansion. We suggest that a more plausible scenario is one in which there was continuous backward and forward migration after an initially rapid spread as indicated by the significant amount of haplotype sharing between Western and Eastern Bantu-speaking groups (fig. 4 and supplementary fig. 5 and supplementary table 3, Supplementary Material online). Thus, our Y chromosome evidence suggests recent expansion and ongoing contacts over the large geographic area occupied by Bantu speakers. This is in good accordance with linguistic evidence showing that the Bantu languages as we know them today have been shaped over the last four millennia through successive stages of "punctuation" and "equilibrium" (Dixon 1997). Punctuational bursts of change at the time of language splitting can account for only 31% of the total divergence in the basic vocabulary of Bantu languages (Atkinson et al. 2008), whereas convergence effects due to multilingualism and intensive and protracted contacts between speech communities certainly played an equally important role in shaping the current Bantu language area (Schadeberg 2003). For instance, the emergence of a relatively homogenous group of so-called "Savannah Bantu" languages, sometimes seen as a Bantu

"subclade" (e.g., Ehret 2001), is most likely the result of intensive contact between languages originally belonging to distinct Eastern and Western Bantu branches (Möhlig 1981; Nurse and Philippson 2003; Bostoen and Grégoire 2007). Phenomena such as political centralization and economic integration involving communities separated over long distances is equally reflected in the archaeological record of several regions of Central, Eastern, and Southern Africa, certainly from the last millennium onward but even earlier (Fagan 1977; Denbow 1990; Chami 1999; De Maret 2005; Phillipson 2005).

Our conclusion contradicts the conclusion of Sikora et al. (2010), who suggest language shift in southeastern Bantu from Mozambique as an explanation for their distinctiveness from three other Bantu populations in the data set (the Luhya from Kenya as well as the Kenyan and South African Bantu groups included in our study). These discrepancies may be explained by the differences in the populations included (southeastern Bantu from Mozambique being unfortunately absent in our data set) or in the type of markers used because autosomal and Y chromosomal markers underlie different demographic trajectories. In summary, our interpretation of the spread of Bantu as a major migratory phenomenon provides a better explanation for the present-day distribution of the paternal lineages in Africa than the alternative scenario of cultural diffusion of the Bantu languages but need not necessarily hold true for the maternal lineages or autosomal markers.

Conclusions

The pattern of Y chromosomal variation in sub-Saharan Africa appears to be driven by the joint effect of both linguistic affiliation and geographical distribution, which to some extent are also correlated. These results were quantified by means of an AMOVA where the percentage of variance explained by differences between groups is larger for the grouping based on linguistic affiliation (~14%) than for that based on geographical criteria (~10%). This somewhat larger effect of language over geography was also found in other studies (Tishkoff et al. 2009 and Bryc et al. 2010). However, there is still a strong effect of geographical proximity (i.e., isolation by distance) on the patterns of Y chromosomal variation as demonstrated by the significant correlation observed between geographic and genetic distances calculated as F_{ST} or R_{ST} values (for haplogroups and STRs, respectively). When considering only Niger-Congo groups, the correlation between R_{ST} and geographic distances is no longer significant, probably because of the recent expansion of the language phylum.

The data presented here make it clear that there is considerable structure within haplogroup E1b1a in Africa. Analyzing the four sublineages E1b1a*(xE1b1a8), E1b1a8, E1b1a7 (xE1b1a7a), and E1b1a7a together with STRs allowed deeper insights into the Y chromosomal variation in this continent and one of the events that shaped it, namely the Bantu expansion. We suggest that the M2 mutation was present in the ancestors of the Niger-Congo populations at an early stage and was subsequently involved in the spread of the language phylum; furthermore, mainly the E1b1a subhaplogroups E1b1a7a and E1b1a8 are implicated in the Bantu expansion. However, some portions of Africa remain understudied; only when these lacunae have been filled will it be possible to come to more definitive insights into the prehistory of this area.

Supplementary Material

Supplementary figs S1—S5, supplementary text, and supplementary tables S1—S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to all the donors of the samples genotyped here; to Vicent Katanekwa, Dudu Musway, Joseph Koni Muluwa, Manuela Cioffi, Gianluca Frinchillucci, and Francesca Lipeti for invaluable assistance with sample collection; to Michael Cysouw, Michael Dannemann, Roger Mundry, and Marc Bauchet for assistance with the statistical analyses and R programming, as well as to Antje Müller for help with DNA extractions and genotyping. This study was supported by the Max Planck Society.

References

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet*. 75:338—345.

Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008. Languages evolve in punctuational bursts. *Science* 319:588.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol Biol Evol*. 26:1581—1589.

Blench R. 2006. Archaeology, language, and the African past. Lanham (MD): Alta Mira Press.

Bosch E, Calafell F, Santos FR, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J. 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet*. 65: 1623—1638.

Bostoen K. 2007. Pots, words and the Bantu problem: on lexical reconstruction and early African history. *J Afr Hist*. 48: 173—199.

Bostoen K, Grégoire C. 2007. 'La question bantoue: bilan et perspectives'. Mémoires de la Société de Linguistique de Paris (NS) 15, special issue: tradition et rupture dans les grammaires comparées de différentes familles de langues. Leuven (Belgium): Peeters p. 73—91.

Bostoen K. 2009. Shanjo and Fwe as part of Bantu Botatwe: a diachronic phonological approach. In: Ojo A, Moshi L, editors. Selected Proceedings of the 39th Annual Conference on African Linguistics. Sommerville (MA): Cascadilla Proceedings Project. p. 110—130.

Bryc K, Auton A, Nelson MR, et al. (11 co-authors). 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A*. 107:786–791.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet*. 9:403–433.

Cann HM, de Toma C, Cazes L, et al. (38 co-authors). 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Chami FA. 1999. Roman beads from the Rufiji Delta, Tanzania: first incontrovertible archaeological link with the Periplus. *Curr Anthropol*. 40(2):237–241.

Cann HM, Sequeira F, de Toma C, Cazes L. (38 co-authors). 2002. A human genome diversity cell line panel. *Science*. 296:261–262.

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 Evol Biol*. 9:80.

Cruciani F, Santolamazza P, Shen P, et al. (16 co-authors). 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*. 70:1197–1214.

de Filippo C, Heyn P, Barham L, Stoneking M, Pakendorf B. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *Am J Phys Anthropol*. 141:382–394.

De Maret P. 2005. From pottery groups to ethnic groups in Central Africa. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 420–440.

Denbow JR. 1990. Congo to Kalahari: data and hypotheses about the political economy of the Western stream of the early iron age. *Afr Archaeol Rev*. 8:139–176.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, Spedini G, Capelli C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol*. 21:1673–1682.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.

Dimmendaal GJ. 2008. Language ecology and linguistic diversity on the African continent. *Lang Linguist Compass*. 2:840–858.

Dixon RMW. 1997. The rise and fall of languages. Cambridge: Cambridge University Press.

Eggert M. 2005. The Bantu problem and African archaeology. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 301–326.

Ehret C. 2001. Bantu expansions: re-envisioning a central problem of early African history. *Int J Afr Hist Stud*. 34:5–27.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online*. 1:47–50.

Fagan BM. 1977. Early trade and raw materials in South Central Africa. In: Konczacki ZA, Konczacki JM, editors. An economic history of tropical Africa. Volume 1: the pre-colonial period. London: Frank Cass. p. 179–192.

Fortune G. 1970. The languages of the western province of Zambia. *J Lang Assoc East Afr*. 1:31–38.

Goldstein DB, Pollock DD. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered*. 88:335–342.

Gomes V, Sanchez-Diz P, Amorim A, Carracedo A, Gusmao L. 2010. Digging deeper into East African human Y chromosome lineages. *Hum Genet*. 127:603–613.

Gordon RG, Grimes BF. 2005. Ethnologue: languages of the world. Dallas (TX): SIL International. p. 1272.

Greenberg JH. 1948. The classification of African languages. *Am Anthropol*. 50:24–30.

Güldemann T, Vossen R. 2000. Khoisan. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 99–122.

Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet*. 4:e1000202.

Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A*. 105:10693–10698.

Holden CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc R Soc Lond B Biol Sci*. 269:793–799.

Holden CJ, Gray RD. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In: Forster P, Renfrew C, editors. Phylogenetic methods and the prehistory of languages. Cambridge: The MacDonald Institute for Archaeological Research. p. 19–31.

Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*. 4:598–612.

Kayser M, Brauer S, Cordaux R, et al. (12 co-authors). 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol*. 23:2234–2244.

Kayser M, Lao O, Saar K, Brauer S, Wang X, Nurnberg P, Trent RJ, Stoneking M. 2008. Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am J Hum Genet*. 82:194–198.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*. 18:830–838.

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 16:1215.

Möhlig WJG. 1981. Stratification in the history of the Bantu languages. *Sprach Gesch Afr*. 3:251–316.

Nenadic O, Greenacre M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *J Stat Softw*. 20:1–13.

Neumann K. 2005. The romance of farming: plant cultivation and domestication in Africa. In: Stahl AB, editor. African archaeology: a critical introduction. Malden (MA): Blackwell Pub. p. 249–275.

Nichols J. 1997. Modeling ancient population structures and movement in linguistics. *Annu Rev Anthropol*. 26:359–384.

Nurse D, Philippson G. 2003. The Bantu languages. London and New York: Routledge. p. 708.

Pereira L, Gusmao L, Alves C, Amorim A, Prata MJ. 2002. Bantu and European Y-lineages in sub-Saharan Africa. *Ann Hum Genet*. 66:369–378.

Pebley A, Mbugua W, Goldman N. 1988. Polygyny and fertility in sub-Saharan Africa. *Fertil Determ Res Notes*. 21:6–10.

Phillipson D. 2005. African archaeology. Cambridge: Cambridge University Press.

Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. 2006. Evaluation of saliva as a source of human DNA for population and association studies. *Anal Biochem*. 353:272–277.

Quintana-Murci L, Quach H, Harmant C, et al. (23 co-authors). 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A*. 105:1596–1601.

Robertson JH, Bradley R. 2000. A new paradigm: the African early iron age without Bantu migrations. *Hist Afr*. 27:287–323.

Rosa A, Ornelas C, Jobling MA, Brehm A, Villems R. 2007. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol Biol*. 7:124.

Y-chromosomal Variation in Sub-Saharan Africa · doi:10.1093/molbev/msq312 MBE

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet*. 70:841–847.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A. 2002. The making of the African mtDNA landscape. *Am J Hum Genet*. 71:1082–1111.

Sands B. 2009. Africa's linguistic diversity. *Lang Linguist Compass*. 3:559–580.

Schadeberg T. 2003. Historical linguistics. In: Nurse D, Philippson G, editors. The Bantu languages. London and New York: Routledge. p. 143–163.

Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. 2010. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet*. (online)

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB. 2000. Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba—the "Black Jews of Southern Africa". *Am J Hum Genet*. 66:674–686.

Tishkoff SA, Gonder MK, Henn BM, et al. (12 co-authors). 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol*. 24:2180–2195.

Tishkoff SA, Reed FA, Friedlaender FR, et al. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Underhill PA, Shen P, Lin AA, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet*. 26:358–361.

Vansina J. 1979. Bantu in the crystal ball .1. *Hist Afr*. 6:287–333.

Vansina J. 1995. New linguistic evidence and the Bantu expansion. *J Afr Hist*. 36:173–195.

Veeramah KR, Connell BA, Pour NA, Powell A, Plaster CA, Zeitlyn D, Mendell NR, Weale ME, Bradman N, Thomas MG. 2010. Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. *BMC Evol Biol*. 10:92.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. New York: Springer. p. 495.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*. 26:1823–1827.

Williamson K. 1989. Niger-Congo overview. In: Bendor-Samuel JT, Rhonda LH, editors. The Niger-Congo languages—a classification and description of Africa's largest language family. Lanham (MD): University Press of America. p. 3–45.

Williamson K, Blench R. 2000. Niger-Congo. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 11–42.

Wood ET, Stover DA, Ehret C, et al. (11 co-authors). 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet*. 13:867–876.

Zhivotovsky LA, Underhill PA, Cinnioglu C, et al. (17 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet*. 74:50–61.

SUPPLEMENTARY MATERIAL

Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups

Inferring haplogroup frequencies from STR haplotypes

Linear Discriminant Analysis (LDA) was performed in order to infer the most probable haplogroup from STRs haplotypes. LDA was carried out with the R statistical software by means of the function "lda" from the package MASS (Venables and Ripley 2002).

First, we tested the power of the LDA in our dataset of 877 individuals carrying the M2 mutation and representing the four haplogroups E1b1a* (xE1b1a7, xE1b1a8), E1b1a7* (xE1b1a7a), E1b1a7a, and E1b1a8. Four tests with 7, 10, 11 and 12 STRs were performed to verify whether the number of STR loci considered can influence the statistical power of the LDA; individuals with missing data were omitted from the analysis. In each of these tests we bootstrapped 1000 times over our dataset to assess the probability of matching the inferred with the genotyped haplogroup. The agreement between the inferred and the genotyped haplogroups was used to evaluate the performance of the LDA. When using all 12 STRs for all 877 individuals, the average performance across 1000 bootstraps was only 83.9%. The performance increased to 93.7% when we considered a subset of 597 individuals from those groups close to or in the geographic area occupied by Bantu speaking groups, here termed the 'Sub-Saharan Bantu' (SSB) dataset (see Supplementary fig. 1 for more details, Supplementary Material online). Similar results were observed when LDA was tested with the other E haplogroups (xE1b1a) by excluding all the individuals carrying the M2 mutation (data not shown).

Second, we tested whether the inaccuracies inherent in the inference process might result in significantly different frequencies of the four sub-lineages of E1b1a in the inferred datasets. A jack-knife procedure was applied over the entire dataset where one individual was considered as unknown and all the others as the reference dataset, and the inferred frequencies were compared to the observed frequencies of the entire and SSB datasets. Furthermore, different thresholds (≥0.6, 0.8, and 0.9) of the posterior probabilities (i.e. the likelihood of a STR-haplotype belonging to a certain haplogroup) were considered. In the entire dataset there were only a few significant differences between the observed and the inferred haplogroup frequencies, and there were no significant differences within the SSB dataset with and without thresholds (Supplementary fig. 2, Supplementary Material online).

Given the LDA results described above, and because our goal was to retrieve the E1b1a sub-haplogroup composition in 1367 samples from various groups that live close to or within the area inhabited by Bantu speakers, we used the SSB dataset as reference without applying any threshold for our final inference. For the Tanzanian sample set of Tishkoff et al. (2007) we used 11 STRs because only the sum of DYS385a/b was available, but this did not affect the results substantially, as shown in Supplementary fig. 1 (Supplementary Material online).

Network analysis

The patterns of haplotype variation within haplogroups E1b1a7a and E1b1a8 were investigated with the help of Median Joining networks (Bandelt et al. 1999) constructed with Network 4.11 (www.fluxus-engineering.com). Weights were assigned to each individual STR locus as inversely proportional to the variance observed in our dataset (Bosch et al. 2006). Individuals that had STR missing values were excluded from the analysis. Furthermore, because 31 individuals had an "irregular" STR value (e.g. with a decimal portion, indicating a partial repetition of one unit of repeat) at the loci DYS439 and/or DYS385a/b, these STRs loci have been split in two: one with the integer value and the other with presence or absence of a decimal value. The weights for these two artificial loci were assigned to 100.

Literature Cited

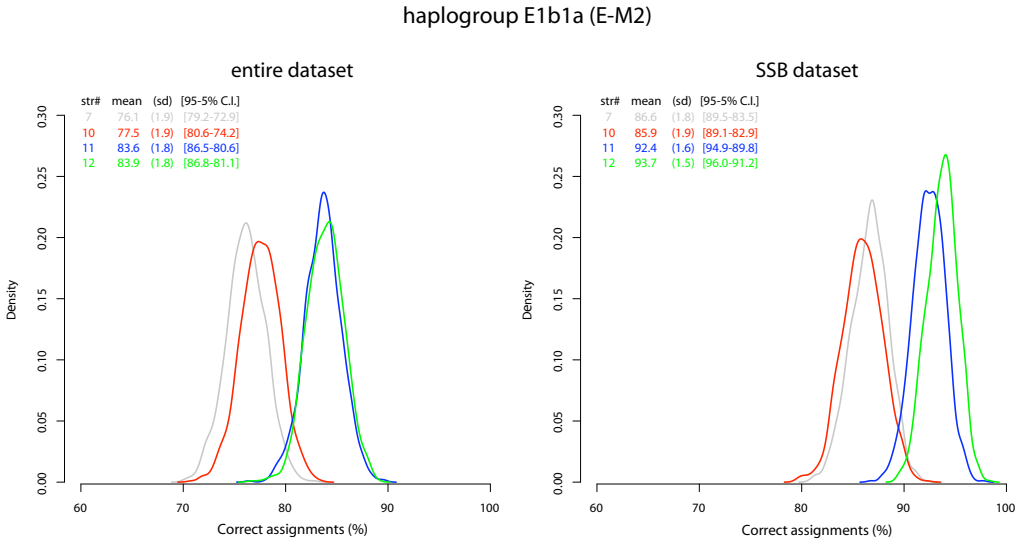
Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48.

Bosch E, Calafell F, Gonzalez-Neira A, et al. (12 co-authors) 2006. Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann Hum Genet 70:459-487.

Tishkoff, SA, Gonder MK, Henn BM, et al. (12 co-authors) 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180-2195.

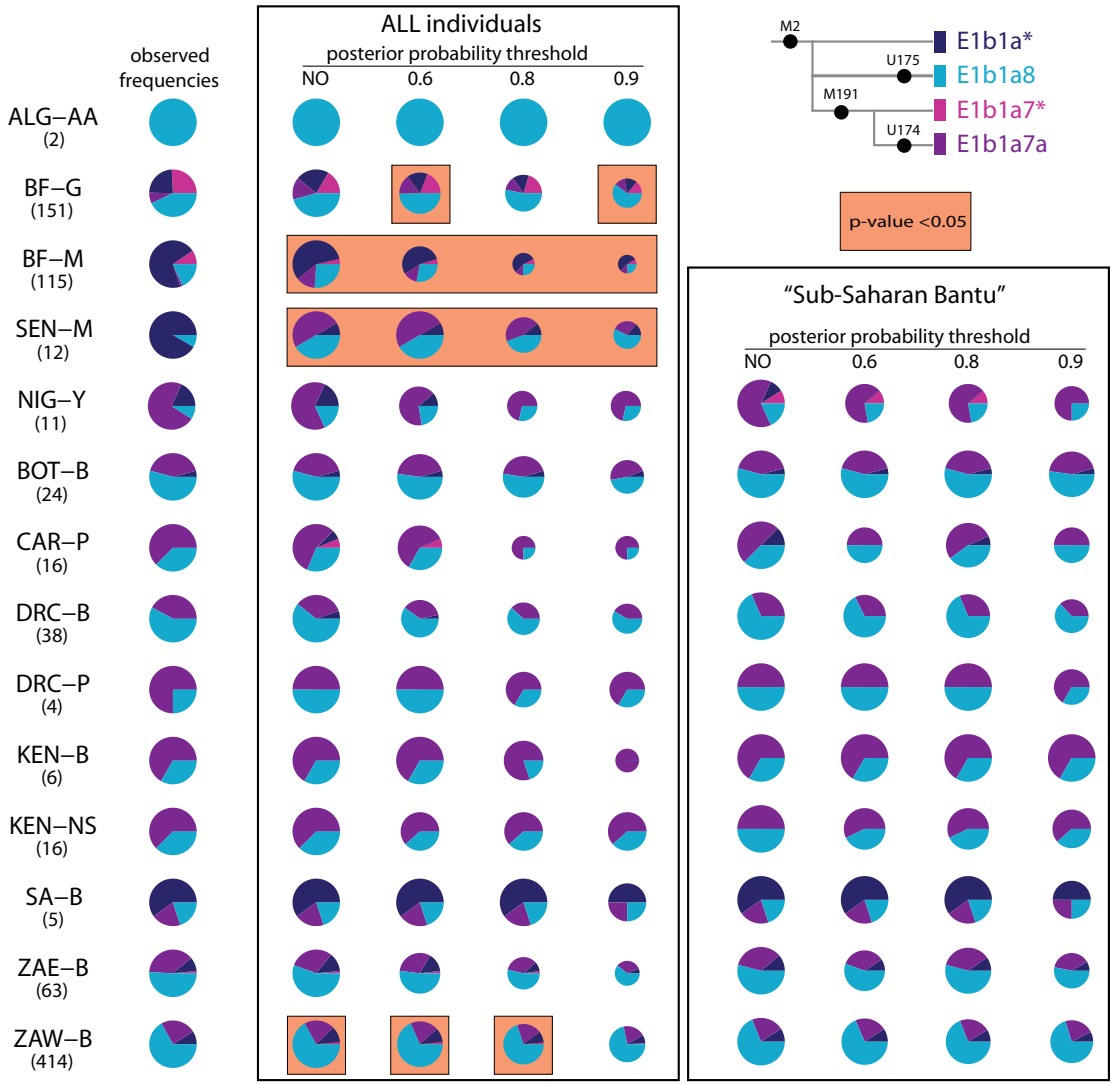
Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. In: Springer, editor. p. 495.

Supplemental Fig. S1: Test of the performance of Linear Discriminant Analysis (LDA)



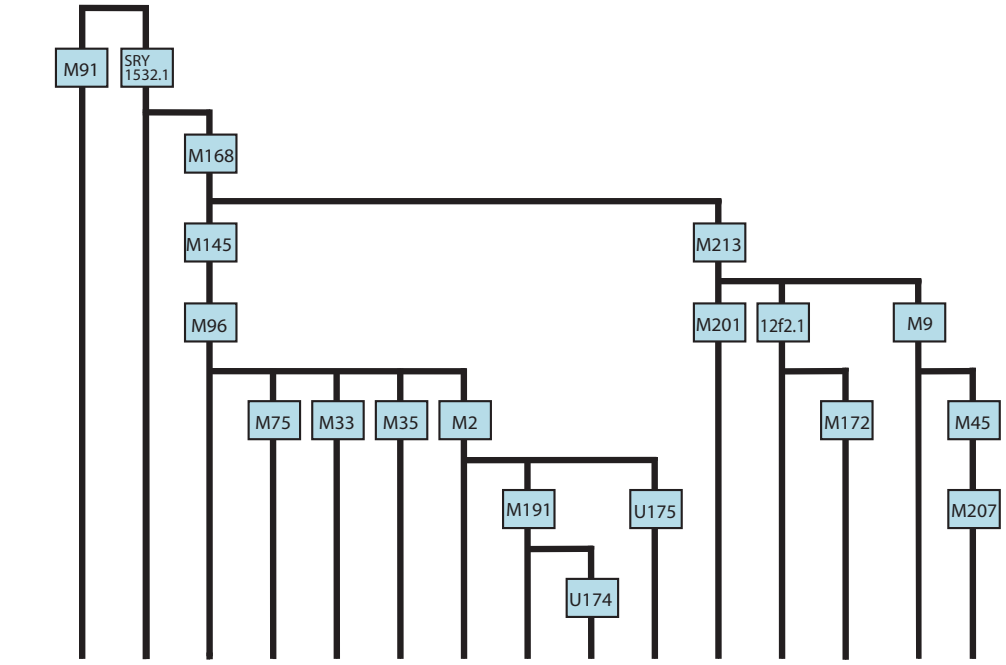
The two plots show the distribution of the percentage of correct assignments of LDA over 1000 bootstraps using all 877 individuals (left), or a subset of individuals (right) belonging to Bantu groups or to groups living in close proximity to Bantu groups (SSB). The SSB dataset included 597 individuals from the following groups: Pygmies from D.R.C. and C.A.R.; Bantu from Botswana, D.R.C., Kenya, South Africa and Zambia; Yoruba from Nigeria; Nilo-Saharan from Kenya. Line colors indicate number of STR loci used: gray, 7 (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393); red, 10 (DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393,DYS437, DYS438, DYS439); blue, 11 (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b); green, 12 (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS385a, and DYS385b).

Supplemental Fig. S2: Test of LDA for inferring haplogroup frequencies of E1b1a (E-M2) sub-lineages



The haplogroup of each individual was inferred by LDA using all other individuals as a reference dataset, with different thresholds for the posterior probabilities (i.e. 0.6, 0.8, and 0.9) as well as no threshold. Pie charts indicate haplogroup frequencies, colored according to the phylogeny at the top right of the figure. "ALL individuals" and "Sub-Saharan Bantu" are as defined in the legend to Supplemental Figure 1.

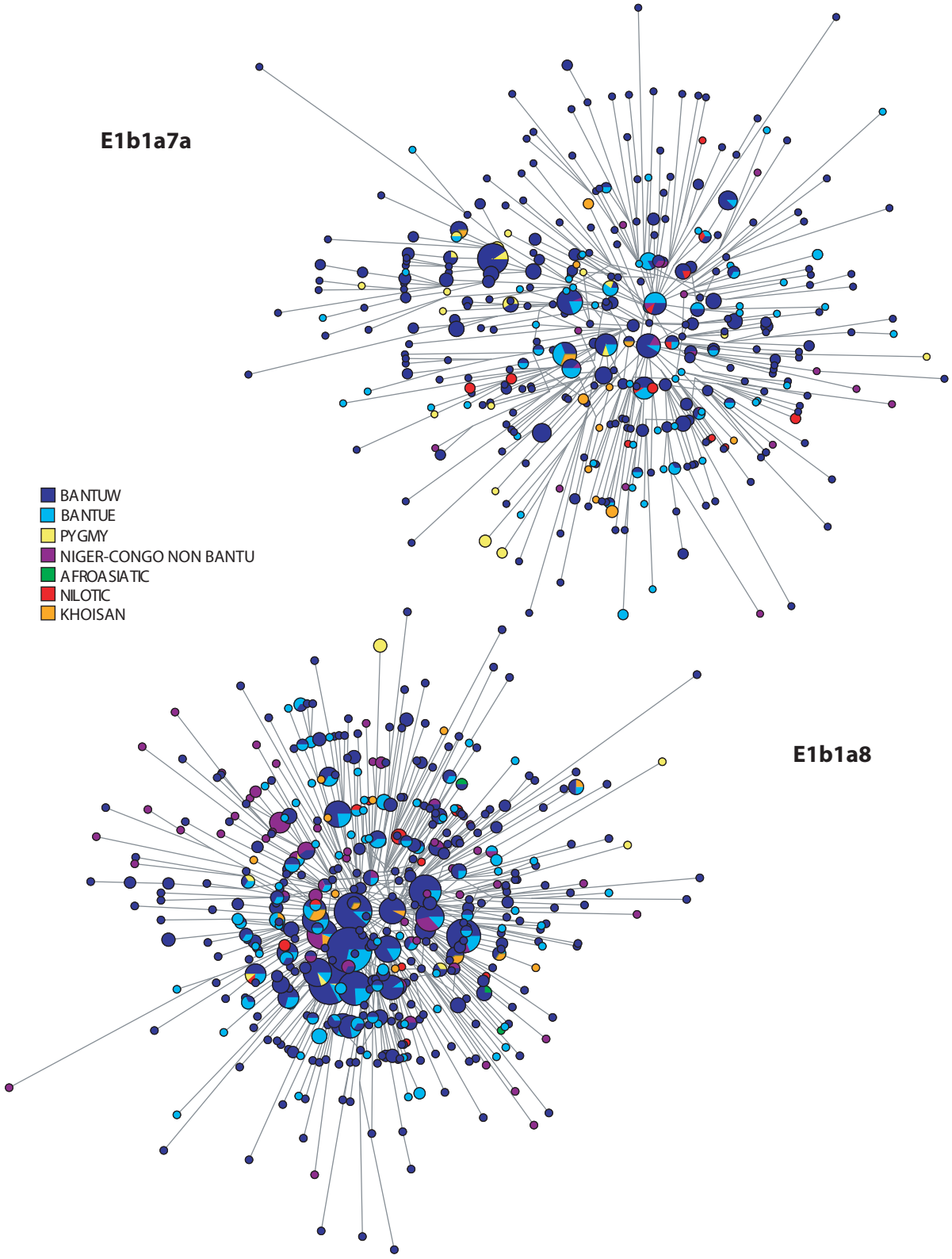
Supplemental Fig. S3: Haplogroup tree and frequency counts



	n	A	B	E*	E2	E1a	E1b1b1	E1b1a	E1b1a7	E1b1a7a	E1b1a8	G	J	J2	K,T	R
Algeria	20						16				2					2
BurkinaFaso Gur	183	2	11	1	1	5	5	38	40	11	67					2
BurkinaFaso Mande	152		8	1		22	2	81	11	3	22					2
Botswana Bantu	40	1	8		3		2	1		11	13					1
C.A.R. Pygmies	23		7							10	6					
D.R.C. Bantu	58		3		1			5		21	28					
D.R.C. Pygmies	11		5		2					3	1					
Ethiopia	98	14	1	29	2	1	37				1		5	2	6	
Kenya Nilo-Saharan	79	17	5		9		22	1		15	7				3	
Kenya Bantu	10	1	0		2		1			4	2					
Namibia	6	3	3													
Nigeria	12		1					2		8	1					
Senegal	15					2	1	11			1					
South Africa Bantu	8		1		1		1	3		1	1					
Zambia West Bantu	480		23		26		4	39		101	278	1			2	6
Zambia East Bantu	69	1	1		3			6	1	24	32					1

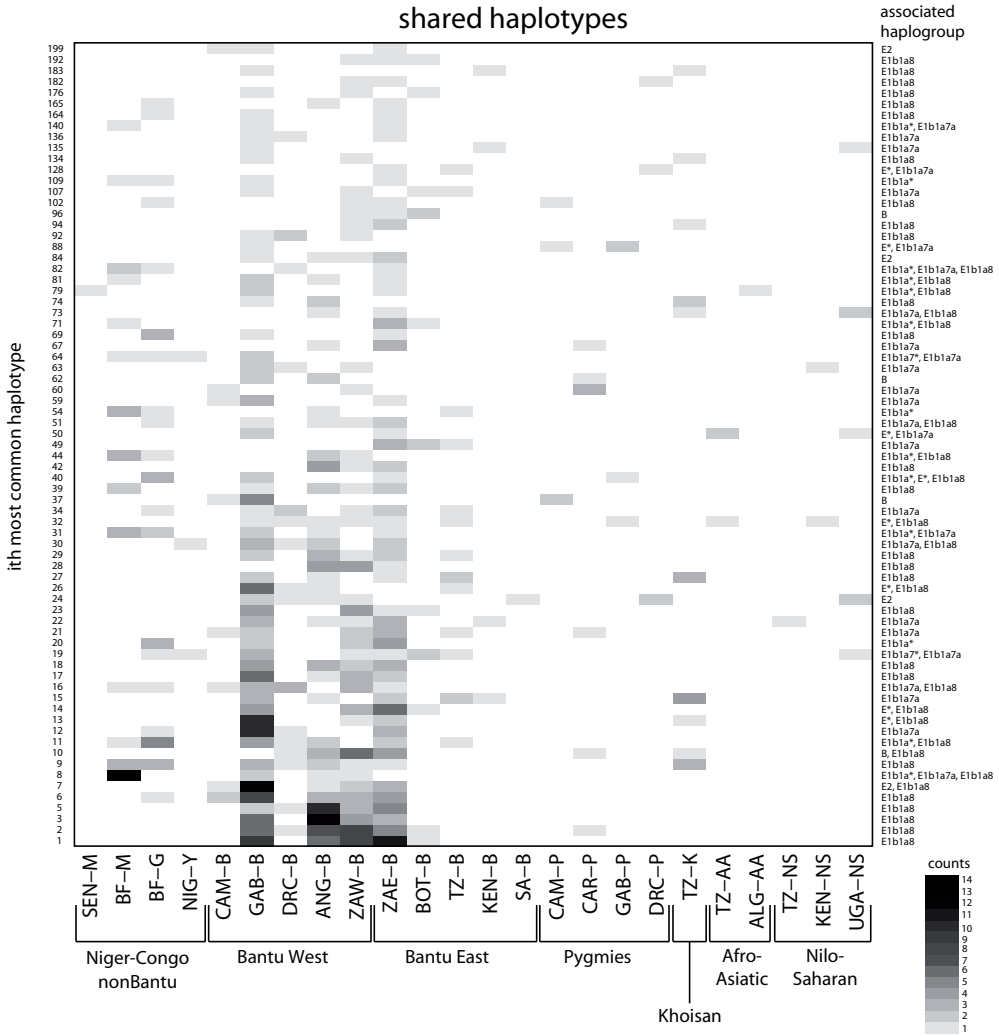
The tree shows the phylogenetic relationship of the markers genotyped here. The table below the figure gives the count of each haplogroup in each of the 16 populations genotyped in the current study.

Supplemental Fig. S4: Networks for STR haplotypes belonging to haplogroups E1b1a7a and E1b1a8



The analyses are based on 11 STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b). Because 31 individuals had an "irregular" STR value (e.g. with a decimal portion, indicating a partial repetition of one unit of repeat) at the loci DYS439 and DYS385a/b, these STRs loci have been split in two: one with the integer value and the other with presence or absence of a decimal value.

Supplemental Fig. S5: Patterns of haplotype sharing for individual groups



Heat plots showing the count of the most common haplotypes from 11 STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, and the sum of DYS385a/b) shared among at least three individual groups.

Supplementary Table S1. Data relative to the primers employed in the SNPs typing.

haplogroup	UEP	dbSNP accesion number	PCR forward primer seq. (5-3')	PCR reverse primer seq. (5-3')	SNaPshot primer sequence (5-3')	SNaPshot primer orientation	Ancestral / Derived allele	SNaPshot-plex	PCR-plex
D2, J	12f2	AC005820	cactgactgatcaaaatgcttacagat	ggatcccttccttacacttataca	gtgccacgtcgtgaaagtctgacaaaacatgtaagtctttaatccatctc	forward	A/-	all-1	24-plex
M1	M106	AC010889	tgtacttggacaggtgaagca	tcgcttttccacctactcct	gtcgtgaaagtctgacaatagttcccttatgacagtatc	forward	A/G	all-1	24-plex
R2	M124	AC010889	tcaaagtcacagtatctgaactagca	tcatattgagattttgctttcct	ccccccccccccaggtgccacgtcgtgaaagtctgacaaggggaacagggaagt	reverse	C/T	all-2	24-plex
DE	M145	AC010137	cctcccactccttttttgat	gcatacttgcctccacgact	gactaaactaggtgccacgtcgtgaaagtctgacaatagacaccagaaagaaaggc	forward	G/A	all-1	24-plex
CR	M168	AC002531	tgtttttgcagagacttgga	ctgcccctctatcagaccat	ggtgccacgtcgtgaaagtctgacaagttttaattcttcagctagc	reverse	G/A	all-2	24-plex
I	M170	Rs2032597	cagctcttattaagttatgttttcatattctgtg	gtcctcatttttacagtgagacacaac	caacccacactgaaaaaaa	reverse	T/G	all-2	24-plex
J2	M172	Rs2032604	tgagccctctccatcagaag	gccaggtacagagaaagtttgg	ccaactgactaaactaggtgccacgtcgtgaaagtctgacaacaaacccatttttgatgctt	forward	T/G	all-1	24-plex
D	M174	Rs2032602	tctccgtcacagcaaaaatg	gacccatcttgcaaggaaaa	cgtcgtgaaagtctgacaaccttctggagtgccc	forward	T/C	all-1	24-plex
O	M175	Rs2032678	gatttaaactctctgaatcaggcacat	ttctactgataccttttttctgttcattc	acgtcgtgaaagtctgacaacacatgccttctcacttctc	forward	T/A	all-1	24-plex
E1b1a7	M191	rs2032590	aggagcaagtacagcgagca	taccacagcgccaggataat	ccccccccccgtctgacaacatttttttctttacaacttgacta	forward	T/G	hg-E	7-plex
E1b1a	M2	rs3893	aggcactggtcagaatgaag	aatggaaaatacagctcccc	ttatcctccacagatctca	reverse	T/C	hg-E	7-plex
L	M20	AC009977	agttggccctttgtgtctgt	catgttcagtgcaaatgcaac	cgtgaaagtctgacaacacatttgtaggttcaaccaactgtggattgaaaat	forward	A/G	all-2	24-plex
G	M201	AC004474	gatctaataatccagtatcaactgagg	ccagcatcctatcagcttca	caactgactaaactaggtgccacgtcgtgaaagtctgacaactaagtacctattacgaaaa	reverse	C/A	all-2	24-plex
R	M207	AC006376	ggggcaaatgtaagtcaagc	tgttcgctgctacgaatcttt	gtctgacaaaagtcaagcaagaaatta	forward	A/G	all-1	24-plex
F-R	M213	Rs2032665	ccatataaaaacgcagcattctgtt	tggagagaacttgagaaaaagtagagaa	tgacaatcagaacttaaaacatctcgttac	reverse	A/G	all-2	24-plex
NO	M214	Rs2032674	ccatggtccaattgtacagc	gaggtcaagggtggtgag	ctaaactaggtgccacgtcgtgaaagtctgacaaagacactgtctgaaaacaac	reverse	A/G	all-2	24-plex
R1b1b2	M269	AC007678	aaggggaatgatcagggttt	ccaaggtgctgggatacac	tgccacgtcgtgaaagtctgacaaggaatgatcagggtttggttaat	forward	T/C	all-2	24-plex
E1a	M33		cacaacttcattggctacgg	tatttgttgaagccccaag	ccccccctatctcataagttactagtta	forward	A/C	hg-E	7-plex
E1b1b1	M35		agggcatggtccctttctat	tgggttcaagtttccctgtc	ccccccccccccaactttcggagtctctgcctgtgtc	reverse	C/G	hg-E	7-plex
P	M45	Rs2032631	gagagaggatatcaaaaattggcagt	tgacagtggcaccaaaggtc	aacaactcagaaggagctttttgc	reverse	C/T	all-2	24-plex
H1	M52	AC009977	cctcaacttcccagagtgttg	gacgaagcaaacatttcaagagag	cgtcgtgaaagtctgacaaaatatcaagaaacctatcaaacatcc	reverse	T/G	all-2	24-plex
H	M69	AC010889	tgggtagcctgttcaaatcc	ttccctttgtcttgctgaaa	tgccacgtcgtgaaagtctgacaaggctgtttacactcctgaaa	forward	T/C	all-1	24-plex
E2	M75	rs2032639	tccacacatcaagaaaacttgc	ttgaacagaggcattgtga	ccccccccccccgtgaaaagacaattatcaaaccacatcc	forward	G/A	hg-E	7-plex
K-R	M9	Rs3900	aggaccctgaaatacagaactg	aaatatttcaacatttcacaaaggaa	actgcaaagaaacggcctaagatggttgaat	forward	C/G	all-1	24-plex
A	M91	AC010889	caaaaatccccctacattgc	gcagtgcccttccaaataaa	ccccccaactgactaaactaggtgccacgtcgtgaaagtctgacaattgctattctgttttttttt	forward	T/A	all-1	24-plex
E	M96	AC010889	gccagccaagaatgaagaga	tgagctgtgatgtgtaacttgg	ggaaaacaggtctctcataata	reverse	G/C	all-1	24-plex
Q1a	MEH2	AC010722 / AC004388 (Y/X)	tttgagtaagccatcacccc	tgcaaaaactgcattgatga	cccaactgactaaactaggtgccacgtcgtgaaagtctgacaaatgtaatttaaagcatagtg	forward	GG/GT*	all-2	24-plex
BR, R1a	SRY10831	Rs2534636	tcatccagtccttagcaaccatta	ccacataggtgaacctgaaaatg	tctggcctcttgtatctgacttttttcacacagt	forward	A/G	all-2	24-plex
N1c	Tat	AC002531	gactctgagtgtagacttgtga	gaaggtgccgtaaagtgtgaa	cgtcgtgaaagtctgacaacttctctcttgctgtgctctgaaatattaaattaaaacaac	reverse	A/G	all-1	24-plex
E1b1a7a	U174	rs16980586	tccctgcagtgaaatagttttg	ctcagactttaggtgagatttgc	ccctgacaagtgttgcataccagattaacccat	forward	G/A	hg-E	7-plex
E1b1a8	U175	rs16980588	ctggtcacactaaggcacca	tggcagaggaacttgaaaaga	ccccccccccccccccccgtctgacaaggccacaggtgctaatgaaacc	forward	GA/AA*	hg-E	7-plex

*:two np because the homologous region on the X chromosome is also amplified and therefore genotyped

Supplementary Table S2. Dataset of all samples considered in the analysis.
This table is available online at
http://mbe.oxfordjournals.org/content/28/3/1255/suppl/DC1

Supplementary Table S3. Pattern of haplotype sharing

Each of the two matrices represents pairwise comparisons between meta-groups according to linguistic/ethnic affiliation and geography: the lower triangular portions show the number of observed shared haplotypes between meta-groups; the upper triangular portions show the mean number of shared haplotypes over 1000 permutations.
The number in bold are significant (p-value < 0.05) after Bonferroni's correction for multiple test hypotheses, and the sign in parentheses indicated whether the observed value is higher (+) or lower (-) than the expected one (see also main text for further details).

obs\sim	Niger-Congo	Bantu West	Bantu East	Pygmies	Nilo-Saharan	"Khoisan"	Afro-Asiatic
Niger-Congo	-	25	25	21	17	7	12
Bantu West	26	-	28	24	20	8	14
Bantu East	24	**59(+)**	-	24	20	8	14
Pygmies	**2(-)**	**13(-)**	**10(-)**	-	17	6	12
Nilo-Saharan	**1(-)**	**8(-)**	**7(-)**	**2(-)**	-	5	10
"Khoisan"	1	10	8	1	1	-	4
Afro-Asiatic	**1(-)**	**3(-)**	**3(-)**	**1(-)**	2	0	-

GEOGRAPHY

obs\sim	North	West	Central	South	East
North	-	7	9	8	10
West	1	-	27	25	37
Central	**1(-)**	23	-	32	41
South	1	26	**55(+)**	-	32
East	**0(-)**	**5(-)**	**21(-)**	25	-

Chapter 7

PAPER II: Genetic Perspectives on the Origin of Clicks in Bantu Languages from Southwestern Zambia

This chapter includes the paper "**Genetic Perspectives on the Origin of Clicks in Bantu Languages from Southwestern Zambia**" written by Barbieri, Chiara, Anne Butthof, Koen Bostoen, and Brigitte Pakendorf, as it appears in the published version on *European Journal of Human Genetics* (2012, August 29). doi:10.1038/ejhg.2012.192.

European Journal of Human Genetics (2012), 1–7

www.nature.com/ejhg

ARTICLE

Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia

Chiara Barbieri[1,4], Anne Butthof[1,4], Koen Bostoen[2,3] and Brigitte Pakendorf[*,1]

Some Bantu languages spoken in southwestern Zambia and neighboring regions of Botswana, Namibia, and Angola are characterized by the presence of click consonants, whereas their closest linguistic relatives lack such clicks. As clicks are a typical feature not of the Bantu language family, but of Khoisan languages, it is highly probable that the Bantu languages in question borrowed the clicks from Khoisan languages. In this paper, we combine complete mitochondrial genome sequences from a representative sample of populations from the Western Province of Zambia speaking Bantu languages with and without clicks, with fine-scaled analyses of Y-chromosomal single nucleotide polymorphisms and short tandem repeats to investigate the prehistoric contact that led to this borrowing of click consonants. Our results reveal complex population-specific histories, with female-biased admixture from Khoisan-speaking groups associated with the incorporation of click sounds in one Bantu-speaking population, while concomitant levels of potential Khoisan admixture did not result in sound change in another. Furthermore, the lack of sequence sharing between the Bantu-speaking groups from southwestern Zambia investigated here and extant Khoisan populations provides an indication that there must have been genetic substructure in the Khoisan-speaking indigenous groups of southern Africa that did not survive until the present or has been substantially reduced.
European Journal of Human Genetics advance online publication, 29 August 2012; doi:10.1038/ejhg.2012.192

Keywords: Zambia; Bantu; Khoisan; mtDNA; Y chromosome; clicks

INTRODUCTION

Although clicks are generally considered the hallmark of the so-called 'Khoisan' languages, they have also been borrowed into some Bantu languages of southern Africa;[1] best known among these are the South African Bantu languages Zulu and Xhosa. Less well known is the fact that some Bantu languages further north also have click consonants, though to a far lesser degree.[2] These are spoken in a small contiguous area encompassing southeastern Angola, southwestern Zambia, northwestern Botswana, and northeastern Namibia (Figure 1), and belong to different subgroups of the Bantu family.[3,4] In the Botatwe subgroup, clicks are found only in Fwe, being absent from the closely related languages Shanjo, Totela, and Subiya and the more distantly related Tonga; in the Luyana subgroup, clicks are found in Mbukushu, but are absent from its close relative Kwamashi (cf. Figure 1).[2]

From a genetic perspective, Khoisan-speaking populations are characterized by specific haplogroups both on the Y chromosome and the mtDNA, which are found in considerable frequencies only in these populations or in groups with a known history of contact with such populations.[5,6] Among Bantu-speaking populations of southern Africa, the amount of detectable intermarriage with Khoisan peoples varies between regions and populations and is not always correlated with the presence of click sounds in the languages they speak. For example, so-called 'southeastern Bantu' populations from South Africa show ~29% of Khoisan-specific mtDNA haplogroups L0d and L0k[7] and ~5% of Y-chromosomal haplogroup A-M51,[8] while only some of their languages have clicks. Bantu-speaking groups from southern Angola also carry varying proportions of characteristic Khoisan haplogroups,[9] with the pastoralist Herero-speaking Kuvale showing surprisingly high levels of intermarriage (~22% of mtDNA haplogroup L0d and ~12% of Y-chromosomal haplogroup B-M112), but none of them has clicks.

The presence of clicks in certain Bantu languages of southwestern Zambia, and their absence in close relatives, raises the question of the origin of these consonants. Apart from their independent innovation in the Bantu languages, which is highly unlikely, there are three probable pathways by which clicks might have entered the Southwest Bantu languages that have them: (1) through superficial 'culture contact' in which Bantu speakers borrowed words containing clicks from Khoisan languages without further intimate contact; (2) through language shift, in which entire groups of Khoisan speakers, both men and women, gave up their original language in favor of a Bantu language, transferring some words and sounds to the new language in the process; or (3) through intermarriage between Bantu speakers and Khoisan speakers. If the sociocultural situation in prehistoric times was similar to that of the present,[10] this intermarriage is likely to have been sex-biased, with Khoisan-speaking women marrying Bantu-speaking men, but not the opposite.

During the migration of Bantu speakers to southwestern Zambia, there would have been several opportunities for contact with local Khoisan speakers. The oldest Early Iron Age archeological sites in the

[1]Max Planck Research Group on Comparative Population Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; [2]Royal Museum for Central Africa, Tervuren, Belgium
*Correspondence: Dr B Pakendorf, Laboratoire Dynamique du Langage, UMR 5596–CNRS & Université Lyon Lumière 2, 14 avenue Berthelot, 69363 Lyon cedex 07, France. Tel: +04 72 72 64 10; Fax: +04 72 72 65 90. E-mail: brigitte.pakendorf@ish-lyon.cnrs.fr
[3]Current address: Department of African Languages and Cultures, Ghent University, KongoKing Research Group, B-9000 Ghent, Belgium and Université libre de Bruxelles, Faculté de Philosophie et Lettres, B-1050 Brussels, Belgium
[4]Current address: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Received 5 April 2012; revised 12 July 2012; accepted 27 July 2012

Upper Zambezi valley, which are generally associated with the settlement of the first Bantu speech communities, date back ~2200 years.[11] These presumably Bantu-speaking communities reached areas already inhabited by hunter-gatherers who probably spoke languages related to modern-day Khoisan languages. As to the Western Bantu-Botatwe peoples (the Fwe, Shanjo, Totela, and Subiya), their ancestors were initially settled further to the east, in the Kafue plains, as indicated by linguistic and archeological data. From there, they spread to the southwest several hundred years ago, with a further migration south to the Zambezi river and beyond during the eighteenth and nineteenth centuries, to escape the pressures of the expanding Luyi/Lozi kingdom.[12] Contact between the ancestors of the Fwe with groups speaking click languages could thus have taken place at different points in time: shortly after the arrival of Bantu speakers in southern Africa, after the split-off of the Western Botatwe languages from the ancestral nucleus in the east, or after the southward migration in the eighteenth/nineteenth century.

In this paper, we attempt to solve the puzzle of the origin of clicks in some of the languages of southwestern Zambia with the help of fine-scaled Y-chromosomal analyses and sequences of complete mtDNA genomes from Fwe- and Mbukushu-speakers as well as their closest linguistic relatives, the Shanjo, Totela, Subiya, and Tonga, and the Kwamashi, respectively (cf. Figure 1). We aim at investigating which of the three possible contact scenarios is the most likely, with the following hypotheses: culture contact is expected to take place in the absence of a significant influx of Khoisan lineages, language shift is expected to lead to an influx of both paternal and maternal lineages, while sex-biased intermarriage is expected to lead to an influx of mtDNA lineages, but not Y-chromosomal ones.

MATERIALS AND METHODS

Materials and DNA analysis

Saliva samples from various populations settled over the entire Western Province of Zambia were collected in August–September 2007.[13] As reported in de Filippo *et al*,[13] after DNA extraction the Y chromosomes were analyzed for 31 single nucleotide polymorphisms, plus 12 short tandem repeat (STR) loci by means of the Promega Y-Powerplex kit (http://www.promega.com). From the total West Zambian data set, only those 132 individuals whose father's father was affiliated with one of the seven populations included in this study were chosen: Fwe, Shanjo, Subiya, Totela, Tonga, Mbukushu, and Kwamashi (see Figure 1 for the approximate location of collection sites for these samples and Supplementary Table 1 for details). The subset of the data analyzed for this study is given in Supplementary Table 2.

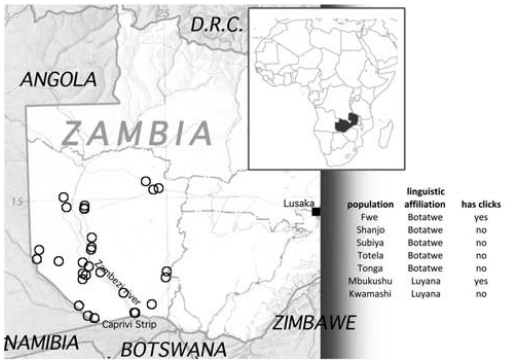


Figure 1 Map showing the position of Zambia within the African continent and the location of the villages sampled. Thirty-one circles are plotted according to their registered latitude and longitude; populations sampled are listed with information on language affiliation and presence/absence of clicks.

mtDNA full genome sequences were generated for 169 individuals whose mother's mother was affiliated with one of the seven populations listed above. Genomic libraries were hybridized with the protocol described in Maricic *et al*,[14] with in-solution capture on target mtDNA. Sequencing was performed on an Illumina GAIIx (Solexa) sequencer. Coverage ranged from an average minimum of 19 × to an average maximum of 438 ×. The number of bases with missing data (gaps, sites with coverage $<2\times$ or where the major base was not present at $>70\%$) is $<1\%$. The two poly-C regions (np 303–315, 16183–16194), which are prone to sequencing errors, were not considered in any of the analyses. All sequences were submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank) and given accession numbers JX303745 - JX303913.

Data analysis

Analysis of Molecular Variance (AMOVA) and standard diversity indices for the Y-chromosome haplogroups and Y-STR haplotypes, plus Φst and RST matrices of distances for the complete mtDNA sequences and the Y-STR haplotypes, respectively, were computed in Arlequin ver. 3.11.[15] For the STR analyses in Arlequin and Network one Tonga sample was not considered because of non-integer numbers of repeats at two loci. Nucleotide diversity and variance for the mtDNA sequence data in single populations was calculated in R with the function 'nuc.div' of the Pegas package.[16] Y-chromosomal haplotype and mtDNA sequence sharing were estimated and plotted with in-house scripts for R. The patterns of mtDNA sequence variation and STR haplotype variation were further investigated with the help of Median Joining networks[17] constructed with Network 4.11 (www.fluxus-engineering.com). For the STR networks, weights were assigned to each individual STR locus as inversely proportional to the variance observed in our data set.[18]

Multi-dimensional scaling analyses of matrices of genetic distances based on Y-chromosomal haplogroup frequencies and complete mtDNA sequences were plotted in Statistica ver. 10.[19]

Simulations were performed in R to assess the levels of migration rates compatible (at $P=0.05$) with the observed proportion of Khoisan-specific haplogroups in extant Bantu populations. Two possible Khoisan source populations were considered: the Ju as the prototypical 'San' group (as an average of haplogroup frequencies of the !Kung and !Xun from Soodyall *et al*[20] and the Tsumkwe San and Sekele/!Kung from Wood *et al*[21]), resulting in 90 and 75% Khoisan-specific mtDNA and Y-chromosome haplogroups, respectively; and the Khwe, with 60% and 16% Khoisan-specific mtDNA and Y-chromosome haplogroups, respectively (Soodyall *et al* 2008). Two of the Zambian populations were considered as recipient populations: the Fwe with 24 and 0%, and the Tonga with 0 and 3% Khoisan-specific maternal and paternal lineages, respectively. Contact was assumed to have taken place from 800 years ago (or 29 generations with a generation time of 28 years) until present, with constant effective population size of 10 000 for both source and recipient populations and a constant migration rate. The probability of seeing a proportion of Khoisan haplogroups within the 95% confidence intervals of the observed values (adjusted for the sample size of the recipient population) was calculated over 10 000 iterations and repeated for a range of migration rates, with the significant thresholds of migration taken from the final distribution of probabilities for each of the eight scenarios.

RESULTS

Y chromosome

The seven populations included here show a fairly homogenous Y-chromosomal haplogroup composition that is very similar to surrounding groups from east Zambia, Angola, DRC, and Gabon[13] (Table 1). Y-chromosomal haplogroups characteristic of Khoisan-speaking populations are found in only very low frequency and not at all in the Fwe and the Mbukushu, the two groups with clicks in their language: haplogroup A is entirely absent from the data set, and haplogroup B-M112 is found in only one individual each of the Subiya, Tonga, and Totela. The homogeneity of all the groups

Table 1 Y-chromosomal diversity and haplogroup composition

		STR Data					*Haplogroup data*								
	n	*Used for STR*	n *Htypes*	*Var*	*G Div*	*SD*	*HG Div*	*SD*	*B-M152*	*B-M112*	*E-M2*	*E-U174*	*E-U175*	*E-M75*	*R*
Fwe	26	26	21	0.40	0.98	0.02	0.44	0.10			0.04	0.19	0.73	0.04	
Shanjo	13	13	13	0.50	1.00	0.03	0.56	0.11			0.08	0.31	0.62		
Subiya	11	11	11	0.44	1.00	0.04	0.47	0.16		0.09	0.00	0.18	0.73		
Totela	13	13	11	1.37	0.97	0.04	0.76	0.10	0.15	0.08	0.23	0.08	0.46		
Tonga	32	31	28	0.94	0.99	0.01	0.65	0.06	0.03	0.03	0.06	0.38	0.47		0.03
Mbukushu	11	11	9	0.59	0.96	0.05	0.69	0.12			0.09	0.18	0.55	0.18	
Kwamashi	26	26	21	0.60	0.98	0.02	0.55	0.10			0.08	0.15	0.65	0.12	

Abbreviations: G Div, gene diversity; Htypes, haplotypes; HG Div, haplogroup diversity; *n*, number of individuals; SD, standard deviation; Var, variance.

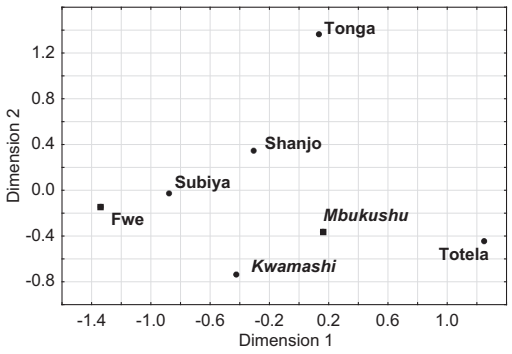


Figure 2 Multi-dimensional scaling based on F_{ST} distances calculated from haplogroup frequencies. Stress value: 0.011. Bantu Luyana populations are indicated in italics, Bantu Botatwe in regular bold. Populations that speak languages with clicks are indicated with a square, the remaining populations with a dot.

included in the analysis is apparent in the multi-dimensional scaling analysis (Figure 2), where no clear clusters emerge; only the Tonga, who are geographically the most distant population in the data set, are separated slightly from the other groups.

This homogeneity is further confirmed by the Y-STR analyses, which demonstrate extensive haplotype sharing among the populations (Supplementary Figures 1 and 2), and by the non-significant pairwise RST values between the populations (Supplementary Table 3). Furthermore, an AMOVA analysis (Table 2a) shows that the seven populations cannot be differentiated at all on the basis of Y-chromosomal haplogroup frequencies: the variance among populations (1%) is not statistically significant. Although there is significant differentiation between the groups at the STR level, this can be shown to be due entirely to the distinctiveness of the Totela, as evidenced by the complete lack of differentiation between groups when this population is removed from the analysis (Table 2a). Grouping the populations by presence *vs* absence of clicks or by linguistic subgroup (Botatwe *vs* Luyana) does not lead to any significant proportion of the variation being apportioned to the between-group component (Table 2a).

Y-chromosomal haplogroup diversity (Table 1) is fairly low overall (0.44–0.76), and especially in the Fwe and Subiya, consistent with the restricted complement of haplogroups present. All populations have relatively high Y-STR gene diversity values; in contrast, the Fwe and the Subiya show reduced Y-STR variance, with the Fwe having the lowest value.

mtDNA

With regard to their mtDNA haplogroup composition, the groups included here are characterized by relatively high frequencies of haplogroups that are widespread in sub-Saharan Africa: L0a, L1b, L1c, L2a, and L3e (Table 3). L1c is typically associated with pygmy populations from Central Africa, but the sublineages to which the southwestern Zambian sequences belong (L1c2a and L1c2b) are characteristic of Bantu speakers rather than pygmies.[5,22] In contrast to the near absence of characteristic Khoisan Y-chromosomal lineages in the southwestern Bantu groups, mtDNA haplogroups L0d and L0k are found in several populations. The Fwe stand out with a very high frequency (24.3%) of these Khoisan haplogroups, in particular a high frequency of L0k (18.2%); in the linguistically closely related Shanjo L0d and L0k reach 16.7%. In the other populations, the Khoisan haplogroups are present in at most low frequency. Although the sequence diversity values are in general fairly high (Table 3), the Fwe stand out as having the lowest value (0.93 ± 0.03) but relatively high mean pairwise differences and nucleotide diversity, demonstrating that whereas several sequences are shared between individuals, these are quite diverse.

An AMOVA performed on the mtDNA sequence data (Table 2b) shows only a low, though statistically significant, differentiation. This is probably due to the distinctiveness of the Fwe, who are significantly different from all populations except for the Shanjo and Mbukushu (though not after Bonferroni correction), while none of the other populations differ significantly from each other, as demonstrated by pairwise Φ_{ST} values (Supplementary Table 4). Grouping the populations by linguistic subgroup or according to the presence *vs* absence of clicks again has no significant effect on the apportionment of variation (Table 2b). The distinctiveness of the Fwe and the Shanjo, the two groups with the highest amount of Khoisan lineages, becomes apparent in the multi-dimensional scaling plot based on Φ_{ST} distances (Figure 3), where the first dimension separates these two populations from all the others.

Haplotype sharing patterns among the populations show an overall fairly high level of sharing (Supplementary Figure 3), and a more fine-scaled analysis of the shared sequences in a network (Supplementary Figure 4) highlights some interesting points. The Fwe and the Shanjo, who are both separated from the other populations in the multi-dimensional scaling plot and who are united in their high frequencies of Khoisan lineages, share only two haplotypes, both on the background of the Khoisan-specific haplogroups: one belonging to L0d and the other belonging to L0k, with another Shanjo L0k sequence only one mutational step away from a Fwe sequence. The results of a resampling test computed in R, where we drew two subsets of 20 and 24 individuals, respectively, from the total number of non-Khoisan sequences in the data set demonstrate that this complete lack of

Table 2 AMOVA analysis

(a) On Y-chromosome data *Criteria 1 group*	*Data set*		*Percentage of variance* *Between populations*	*Within populations*
All seven populations	Y-haplogroup		1.03	98.97
All seven populations	Y-STR		6.65**	93.35
Six populations (excluding Totela)	Y-STR		−1.47	101.47
Criteria 2 groups	*Data set*	*Between groups*	*Within groups*	*Within populations*
Presence clicks vs absence clicks	Y-STR	−2.5	7.95**	94.55**
Presence clicks vs absence clicks	Y-haplogroup	−0.75	1.4	99.35
Botatwe vs Luyana	Y-STR	−2.42	7.91**	99.59**
Botatwe vs Luyana	Y-haplogroup	0.01	1.03	98.96
(b) On mtDNA sequence data *Criteria 1 group*			*Between populations*	*Within populations*
All seven populations			1.91*	98.09
Criteria 2 groups		*Between groups*	*Within groups*	*Within populations*
Presence clicks vs absence clicks		1.68	1.12	97.2*
Botatwe vs Luyana		−0.22	2.01*	98.21*

*Denotes values of significance <0.05. **Denotes values of significance <0.01.

Table 3 mtDNA diversity and haplogroup composition

	Sequence data								*Haplogroup data*												
	n	n *Htypes*	*MPD*	*SD*	*π*	*Var*	*HG Div*	*SD*	*L0a*	*L0d*	*L0k*	*L1b*	*L1c*	*L2*	*L2a*	*L2b*	*L2c*	*L2d*	*L3b*	*L3d*	*L3e*	*L3f*	*L5a*
Fwe	33	20	65.8	29.0	0.0039	0.0009	0.93	0.03	0.12	0.06	0.18	0.27	0.12		0.09		0.03				0.12		
Shanjo	24	17	65.6	29.0	0.0039	0.0009	0.97	0.02	0.04	0.08	0.08	0.25	0.17			0.04					0.21	0.13	
Subiya	17	15	62.7	29.0	0.0037	0.0009	0.99	0.03	0.06	0.06		0.12	0.29		0.24						0.24		
Totela	29	27	62.7	28.0	0.0037	0.0009	0.99	0.01	0.14			0.07	0.17		0.17		0.07	0.03		0.03	0.24	0.07	
Tonga	22	22	66.5	30.0	0.0040	0.0010	1.00	0.01	0.14			0.09	0.23		0.27				0.09		0.14		0.05
Mbukushu	12	11	64.3	28.0	0.0038	0.0010	0.98	0.04			0.08	0.08	0.25		0.33					0.17	0.08		
Kwamashi	32	26	64.4	29.0	0.0038	0.0009	0.98	0.02	0.16	0.03		0.09	0.16	0.03	0.06	0.06				0.19	0.19	0.03	

Abbreviations: Htypes, haplotypes; HG Div, haplogroup diversity; MPD, mean pairwise differences; *n*, number of individuals; π, nucleotide diversity; SD, standard deviation; Var, variance.

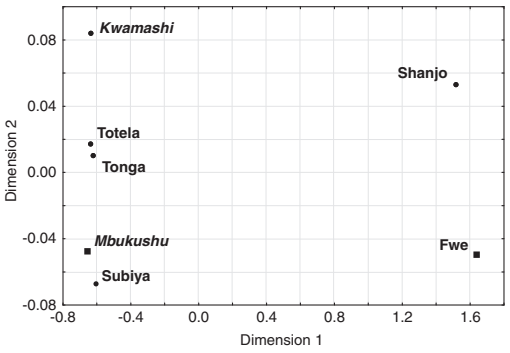


Figure 3 Multi-dimensional scaling based on Φ_{ST} distances between mtDNA sequences. Stress value: 0.007. Bantu Luyana populations are indicated in italics, Bantu Botatwe in regular bold. Populations that speak languages with clicks are indicated with a square, the remaining populations with a dot.

sharing of sequences belonging to non-Khoisan haplogroups is significant ($P = 0.04$, calculated over 10 000 repeats).

A comparison of the southwestern Zambian individuals belonging to haplogroups L0d and L0k with published sequences from southern Africa belonging to these haplogroups[5,23–25] shows a surprising lack of sequence sharing between the Zambians and others, be they Khoisan- or Bantu-speaking (Figure 4). Even though this is certainly due at least in part to a lack of available comparative data, it is still noticeable that the Zambian sequences are not even located close to any published Khoisan sequences, but at the end of very long branches. While any two non-Zambian sequences belonging to haplogroup L0d or L0k are on average separated by 10 mutations, the Zambian L0k sequences are separated from the closest non-Zambian sequences by 27, 26, and 15 mutations; the distance between the Zambian L0d sequences and the closest non-Zambian L0d sequences is 25, 14, 12, and 8 mutations.

Simulations

As can be seen from Table 4, the observed proportions of Khoisan-specific haplogroup frequencies in the Fwe are compatible with

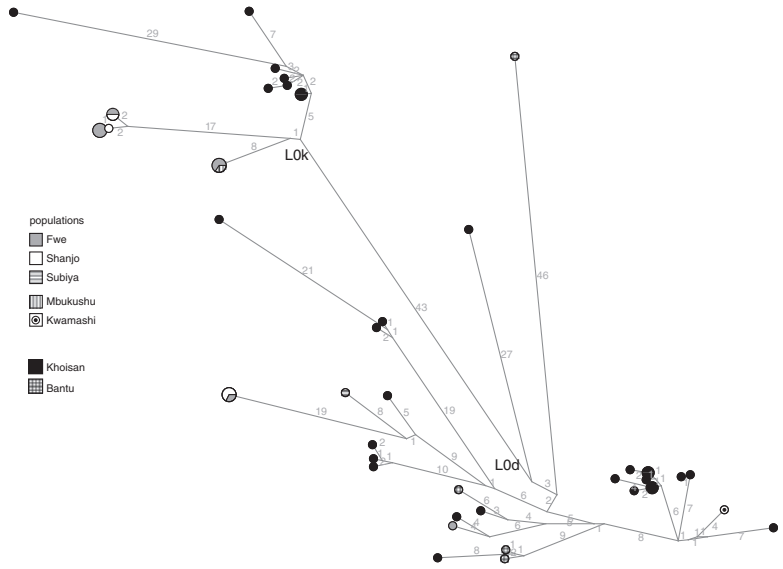


Figure 4 Median joining network of mtDNA sequences belonging to haplogroup L0k and L0d, including individuals from Zambia and sequences retrieved from the literature; the latter are here lumped as 'Khoisan' and 'Bantu', respectively. Numbers on the branches indicate the number of mutations having taken place along that branch.

female-biased gene flow from a Ju-like source population, or unbiased female and male gene flow from a Khwe-like source population. The observed frequencies in the Tonga (and other populations similar to them) are compatible with at most low levels of female gene flow, but potentially high amounts of male gene flow from a Khwe-like population.

DISCUSSION

Clicks in the Bantu languages Fwe and Mbukushu, spoken in southwest Zambia and adjacent areas, may have arisen in three different ways: through mere culture contact without intensive physical interaction, hypothesized to correlate with the absence of large amounts of Khoisan genetic admixture; through language shift of entire groups of Khoisan speakers to a Bantu language, hypothesized to lead to an influx of both paternal and maternal Khoisan lineages to the Bantu gene pool; or through (presumably sex-biased) intermarriage hypothesized to lead to admixture only in the maternal line. At first glance, the results of the Y chromosome and mtDNA analyses appear to indicate sex-biased interactions between the Bantu-speaking populations and Khoisan groups, with a noticeable influx of mtDNA haplogroups L0d and L0k without corresponding levels of introgression of characteristic Y-chromosomal haplogroups. Especially for the Fwe, who speak a Bantu language with clicks, the results appear to indicate that borrowing of click consonants was associated with the incorporation of Khoisan women, as has also been argued for the southeast Bantu Xhosa and Zulu:[26] nearly one quarter (24%) of the Fwe mitochondrial gene pool is of Khoisan origins, whereas no characteristic Khoisan Y-chromosome haplogroups were found in this population (Table 1 and Table 3). However, the simulations show that the observed haplogroup frequencies are compatible with two different scenarios, depending on whether the source population had a haplogroup composition more similar to Ju or to Khwe. Thus, should the Fwe have interacted with a Khwe-like source population, even large amounts of gene flow in the paternal line could have gone undetected in our approach, making it impossible to exclude the hypothesis of a language shift from Khwe-speakers to Fwe. On the other hand, should the source population have had a haplogroup composition similar to the Ju, our results are more compatible with sex-biased gene flow in the maternal line, with at most low levels of paternal gene flow. Two factors are in favor of the latter scenario of female-biased gene flow: first of all, the Fwe have a Y-STR variance of only 0.4 in combination with a haplotype diversity of 0.98 ($\pm$0.02; Table 1). This clearly shows that no very divergent Y-chromosomal haplotypes, such as one would expect to be present in a hunter-gatherer population long separated from the Bantu-speaking immigrants, have entered the Fwe gene pool, and argues against a large proportion of undetected male gene flow. Furthermore, the linguistic data show stronger affinities with a Ju language rather than with Khwe,[2] lending greater weight to our estimates of plausible migration rates based on a Ju source population.

Yet a further hypothesis is that the Fwe and Shanjo shared a common ancestor with Khoisan groups before shifting to their current Bantu language in a process of intermarriage with Bantu, thereby incorporating Bantu genetic lineages and, in the case of the Fwe, carrying over some of the Khoisan click consonants. However, as this scenario would involve the replacement with Bantu lineages of up to 100% of the original Khoisan Y chromosomes and up to 90% of the original Khoisan mtDNA lineages (cf. Table 4), it appears less plausible than the scenario proposed above, namely intermarriage of a Bantu-speaking group with a Khoisan-speaking group restricted to the maternal line.

In agreement with the estimates of migration rates (Table 4), the contact between the Khoisan and the ancestors of the Fwe appears to have been intense, as at least four of the five L0d and L0k mtDNA

Table 4 Migration rates suggested to explain the proportion of 'Khoisan' haplogroups, with significance >0.05

	%HG source population	*%HG receiver population*	n *receiver*	*CI for receiver population*	*Minimum migration rate*	*Migration rate associated with maximum probability*	*Maximum migration rate*
Ju into Fwe, mtDNA	90	24	33	0.1–0.39	0.002	0.01	0.03
Ju into Fwe, Y chromosome	75	0	26	0–0.11		0	0.012
Khwe into Fwe, mtDNA	60	24	33	0.1–0.39	0.0025	0.017	0.07
Khwe into Fwe, Y chromosome	16	0	26	0–0.11		0	>0.5 (always $P>0.05$)
Ju into Tonga, mtDNA	90	0	22	0–0.13		0	0.012
Ju into Tonga, Y chromosome	75	3	33	0–0.09		0	0.009
Khwe into Tonga, mtDNA	60	0	22	0–0.13		0	0.02
Khwe into Tonga, Y chromosome	16	3	33	0–0.09		0	>0.5 (always $P>0.05$)

haplotypes found in the Fwe are so divergent that it is unlikely that they could have evolved from only a couple of ancestral sequences of Khoisan origins. Thus, two of the L0k sequences are separated by 27 mutations; the two L0d sequences are separated by 56 mutations. To accumulate this amount of divergence from a single shared ancestor per haplogroup would take more than a thousand generations,[27] whereas Bantu speakers arrived in Zambia only around 40 generations ago.

These data are therefore compatible with a scenario of intense contact with relatively high levels of intermarriage in the maternal line leading to the borrowing of click phonemes into these languages. However, this apparently straightforward conclusion is complicated by the puzzling lack of haplotype sharing between the Fwe and Khoisan populations, and the long branches, which lead to the Zambian Bantu L0k and L0d haplotypes (Figure 4). This is clearly due in part to the lack of comparative data, as the few complete mtDNA genomes available from Khoisan populations were sampled in a non-random fashion and stem from a highly restricted number of populations. Indeed, a comparison with preliminary data from a more representative range of Khoisan populations shows fewer numbers of mutations separating the Zambian L0d sequences and those from Khoisan populations (Barbieri *et al*, unpublished data); nevertheless, there is still no direct haplotype sharing between extant Khoisan populations and the southwest Zambian groups. Furthermore, the Zambian L0k sequences remain completely distinct, even when more data are included in the analysis (data not shown).

One possible explanation for the lack of sequence sharing between the Bantu and extant Khoisan populations might be that subsequent drift has erased lineages in the Khoisan groups that were retained in the Bantu populations through admixture. An alternative explanation might be that the ancestral hunter-gatherer groups living in the area at the time of the Bantu immigration have since been replaced by the immigrants. A third possibility would be that there was genetic structure among the ancient Khoisan-speaking hunter-gatherer groups, and that the Fwe intermarried with a Khoisan group whose genetic composition differed from that of the populations included in molecular anthropological investigations to date. This assumption is supported to a certain degree by the presence of higher frequencies of L0k than L0d in the Fwe. This differs from what is found in Khoisan-speaking populations[7,28,29] and in populations that have experienced admixture with Khoisan groups,[7,9,26,30] where the proportion of L0d far outweighs that of L0k. More data on both Khoisan and Bantu-speaking groups of southwestern Africa are needed to shed light on this puzzle. Of course, these different explanations are not mutually exclusive, and it is plausible that the Fwe ancestors interacted with a Khoisan community that differed genetically from those still settled in southern Africa today, which was ultimately replaced by the newcomers.

Although the large proportion of Khoisan maternal lineages in the Fwe is in good accordance with the click consonants they have incorporated into their language, the high frequency of haplogroups L0d and L0k in the Shanjo is unexpected from a linguistic perspective, as their language did not incorporate clicks. It is of course quite possible that the Shanjo intermarried to the same extent with Khoisan-speaking women as the Fwe, but for sociocultural reasons did not borrow clicks. If the Khoisan mtDNA lineages in the Fwe and the Shanjo should indeed be the result of independent admixture events, the admixture would arguably have taken place with the same Khoisan-speaking population, as the Khoisan lineages found in the Fwe and the Shanjo are shared. Another possibility, however, is that the Khoisan mtDNA haplogroups found in the Shanjo did not originate directly from admixture with Khoisan communities, but through intermarriage with Fwe. This appears all the more plausible as two of the Shanjo L0d/L0k haplotypes are shared with the Fwe, with the third one being only one mutational step distant from a Fwe sequence type. What is extremely puzzling, however, is the significant lack of sharing of non-Khoisan lineages between the Fwe and Shanjo. This would appear to indicate that the possible intermarriage was biased specifically toward Fwe women with Khoisan maternal ancestry – a bias that is very hard to explain, though some form of social or physical preference may have been at play.

Similar to the Shanjo, it is possible that the Mbukushu did not interact directly with Khoisan communities, as their single L0k haplotype is shared with the Fwe. Unfortunately, however, due to the small sample size available for the Mbukushu, it is not possible to come to any definitive conclusions concerning their prehistory. Nevertheless, the linguistic data, too, is compatible with a possible influx of the click words in this language not through direct interaction with Khoisan speakers, but through borrowing from a Bantu language with clicks belonging to a different subgroup.[2]

In summary, although we cannot exclude substantial amounts of paternal gene flow from a Khwe-like source population and thus language shift, the genetic and linguistic data are in favor of admixture in the maternal line between some of the Bantu groups from Zambia and Khoisan-speaking populations. The amount of this intermarriage does not correlate with the presence of clicks in the languages of the groups concerned, as the Shanjo show a high frequency of Khoisan mtDNA haplotypes in the absence of clicks. The precise modality of the contact between the ancestors of the Fwe and Khoisan-speaking populations is hard to elucidate, but ultimate replacement of the Khoisan group by the Bantu-speaking community coupled with some female-biased admixture is the most plausible scenario. Furthermore, our results show that the mtDNA composition

of the ancestral Khoisan population is most likely to have been distinct from that found in Khoisan groups investigated to date, pointing to the existence of deep genetic structure in the ancestral Khoisan groups of southern Africa. This demonstrates that it will be possible to gain insights into the genetic structure of pre-Bantu Khoisan groups that may no longer exist by looking for their genetic traces in Bantu groups that they admixed with. However, our conclusions are hampered by a lack of comparative data, and in order to obtain further insights into the history of interactions between the immigrating Bantu speech communities and the autochthonous Khoisan groups, more data from populations speaking Khoisan languages as well as Bantu-speaking groups of southern and central Africa are needed.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful to all sample donors for their participation in this study. We thank Mark Stoneking for helpful comments, the Livingstone Museum, Cesare de Filippo, Terry Nyambe, Ellen Gunnarsdóttir, and Mark Stoneking for help with sample collection, Antje Müller for help in the lab, Serena Tucci for help with sequence alignment, and Cesare de Filippo for organizing sample information, extracting DNA, writing R scripts and commenting on the analysis. This research was supported by the Max Planck Society and the Belgian Science Policy; CB was supported by a grant from the Deutsche Forschungsgemeinschaft (to BP).

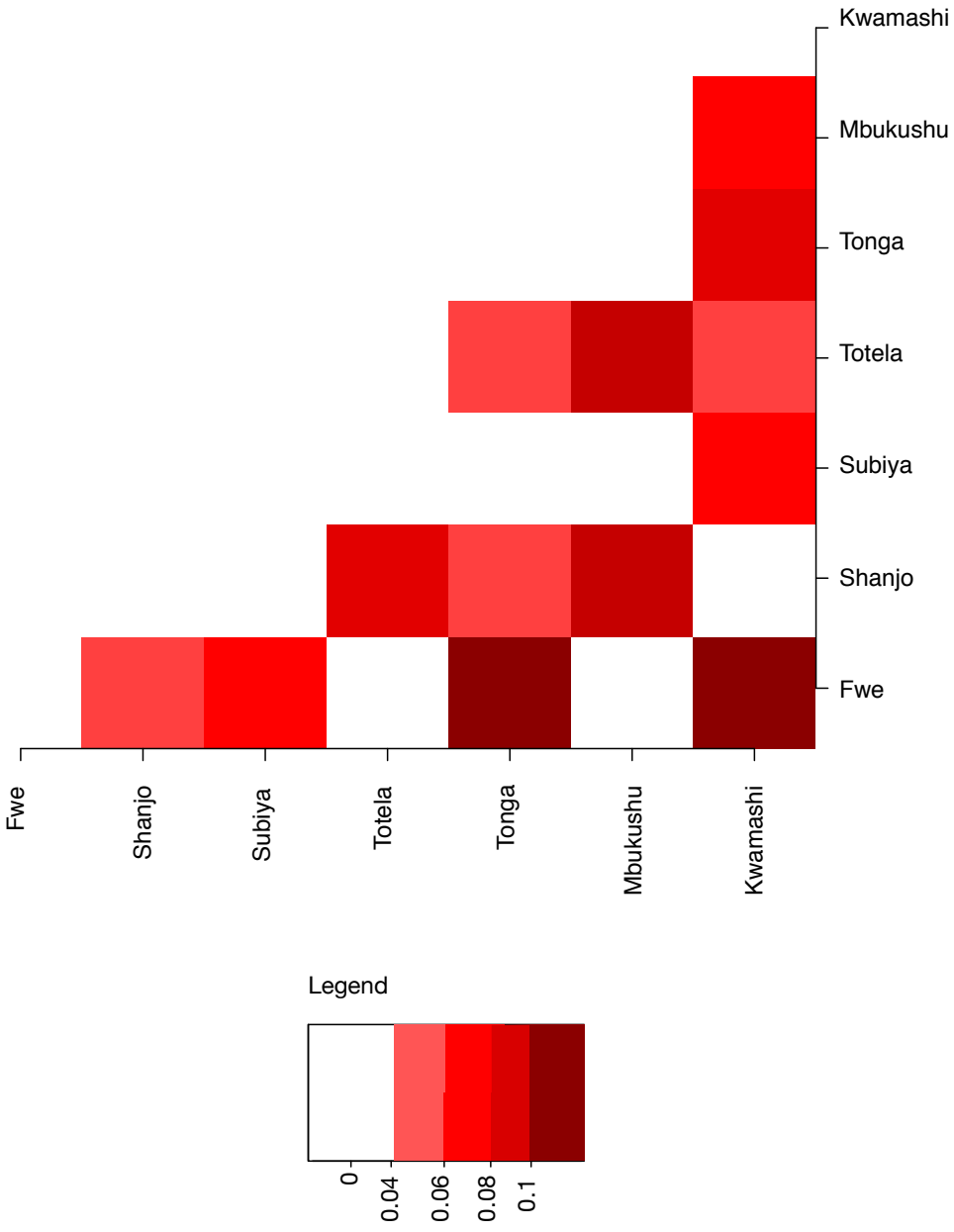
AUTHOR CONTRIBUTIONS

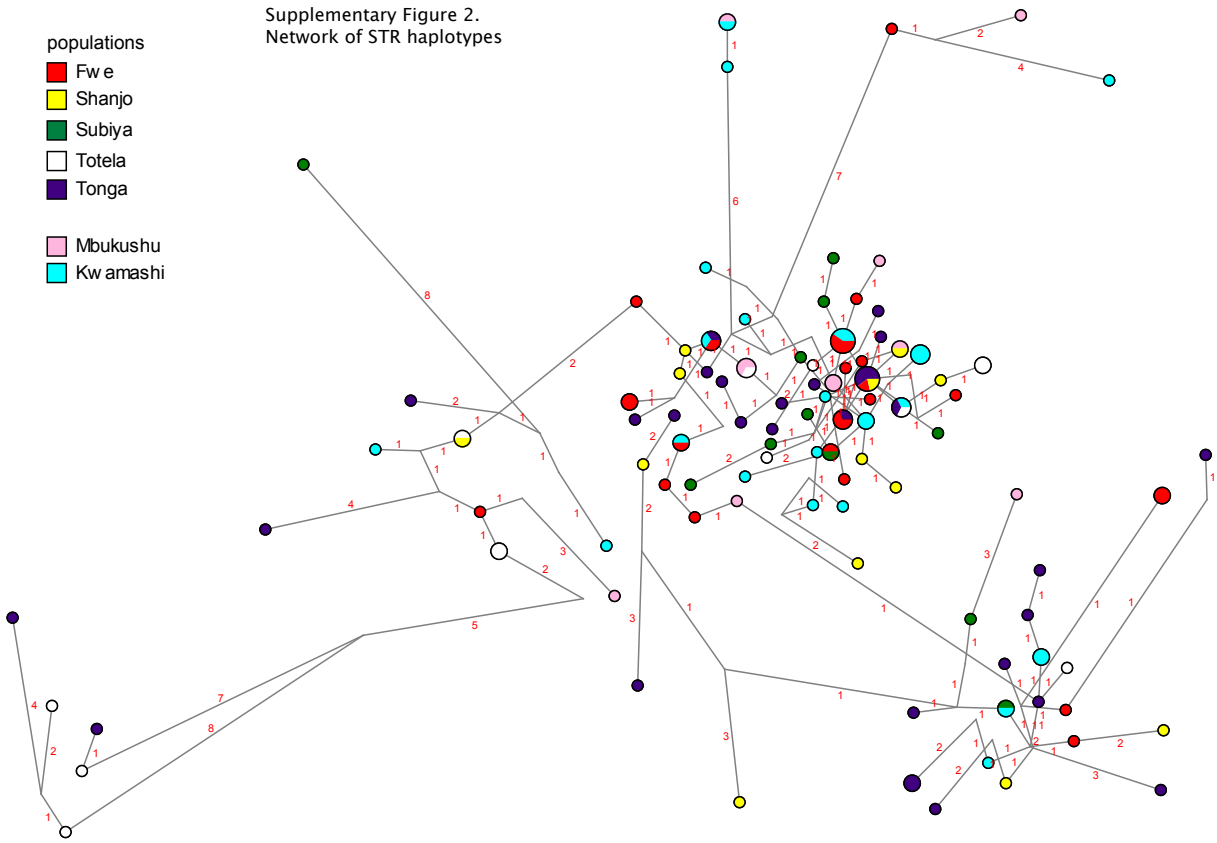
KB and BP conceived the study, AB prepared the sequencing libraries, and CB performed phylogenetic analyses. The paper was written by BP and CB with detailed input from KB.

1 Güldemann T, Stoneking M: A historical appraisal of clicks: a linguistic and genetic population perspective. *Annu Rev Anthropol* 2008; **37**: 93–109.
2 Bostoen K, Sands B: Clicks in south-western Bantu languages: contact-induced vs. language-internal lexical change; in Brenzinger M (ed). *Proceedings of the 6th World Congress of African Linguistics*. Cologne: Köln: Rüdiger Köppe Verlag, 2012 (in press), pp 129–140.
3 Fortune G: The languages of the Western Province of Zambia. *J Lang Assoc East Africa* 1970; **1**: 31–38.
4 Bostoen K: Shanjo and Fwe as part of Bantu Botatwe: a diachronic phonological approach; in Ojo A, Moshi L (eds). *Selected proceedings of the 39th Annual Conference on African Linguistics: Linguistic Research and Languages in Africa*. Sommerville, MA, USA: Cascadilla Press 2009; pp 110–130.
5 Behar DM, Villems R, Soodyall H *et al*: The dawn of human matrilineal diversity. *Am J Hum Genet* 2008; **82**: 1130–1140.
6 Batini C, Ferri G, Destro-Bisol G *et al*: Signatures of the preagricultural peopling processes in Sub-Saharan Africa as revealed by the phylogeography of early Y chromosome Lineages. *Mol Biol Evol* 2011; **28**: 2603–2613.
7 Schlebusch CM, Naidoo T, Soodyall H: SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis* 2009; **30**: 3657–3664.
8 Naidoo T, Schlebusch CM, Makkan H *et al*: Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investig Genet* 2010; **1**: 6.
9 Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J: On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol Biol* 2009; **9**: 80.
10 Pretorius JL: The Fwe of the East Caprivi. *Unpublished MA thesis University of Stellenbosch*, 1975.
11 Phillipson DW: *African archaeology*. Cambridge: Cambridge University Press, 2005.
12 de Luna K: Classifying Botatwe: M60 languages and the settlement chronology of south central Africa. *Afr Linguist* 2010; **16**: 65–96.
13 de Filippo C, Barbieri C, Whitten M *et al*: Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. *Mol Biol Evol* 2011; **28**: 1255–1269.
14 Maricic T, Whitten M, Pääbo S: Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 2010; **5**: e14004–e14004.
15 Excoffier L, Laval G, Schneider S: Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinformatics Online* 2005; **1**: 47–47.
16 Paradis E: pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 2010; **26**: 419–419.
17 Bandelt HJ, Forster P, Rohl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999; **16**: 37–48.
18 Bosch E, Calafell F, Gonzalez-Neira A *et al*: Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet* 2006; **70**: 459–487.
19 StatSoft, Inc. STATISTICA (data analyisis software system), version 10, 2011.
20 Soodyall H, Heeran M, Philip H, Thijessen N: The genetic prehistory of the Khoe and San. *S Afr Humanit* 2008; **20**: 37–48.
21 Wood ET, Stover DA, Ehret C *et al*: Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* 2005; **13**: 867–876.
22 Batini C, Lopes J, Behar DM *et al*: Insights into the Demographic History of African Pygmies from Complete Mitochondrial Genomes. *Mol Biol Evol* 2011; **28**: 1099–1110.
23 Ingman M, Kaessmann H, Pääbo S, Gyllensten U: Mitochondrial genome variation and the origin of modern humans. *Nature* 2000; **408**: 708–713.
24 Arnason U, Gullberg A, Janke A, Kullberg M: Mitogenomic analyses of caniform relationships. *Mol Phylogenet Evol* 2007; **45**: 863–874.
25 Hartmann A, Thieme M, Nanduri LK *et al*: Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum Mutat* 2009; **30**: 115–122.
26 Salas A, Richards M, De la Fe T *et al*: The making of the African mtDNA landscape. *Am J Hum Genet* 2002; **71**: 1082–1111.
27 Soares P, Ermini L, Thomson N *et al*: Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 2009; **84**: 740–759.
28 Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA: Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* 2007; **24**: 757–768.
29 Henn BM, Gignoux CR, Jobin M *et al*: Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA* 2011; **108**: 5154–5162.
30 Quintana-Murci L, Harmant C, Quach H *et al*: Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. *Am J Hum Genet* 2010; **86**: 611–620.

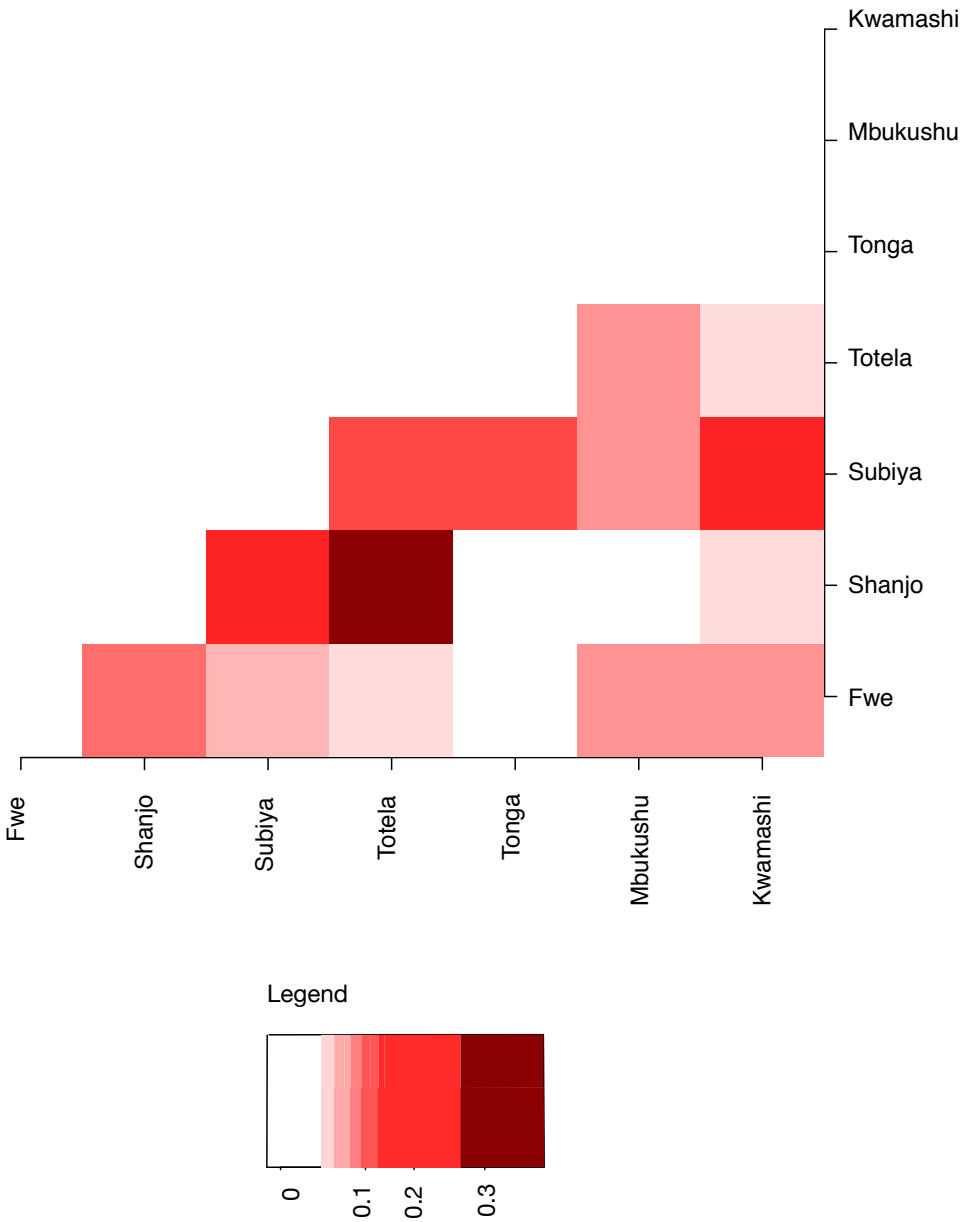
Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Supplementary Figure 1
frequency of shared haplotypes, Y chromosome

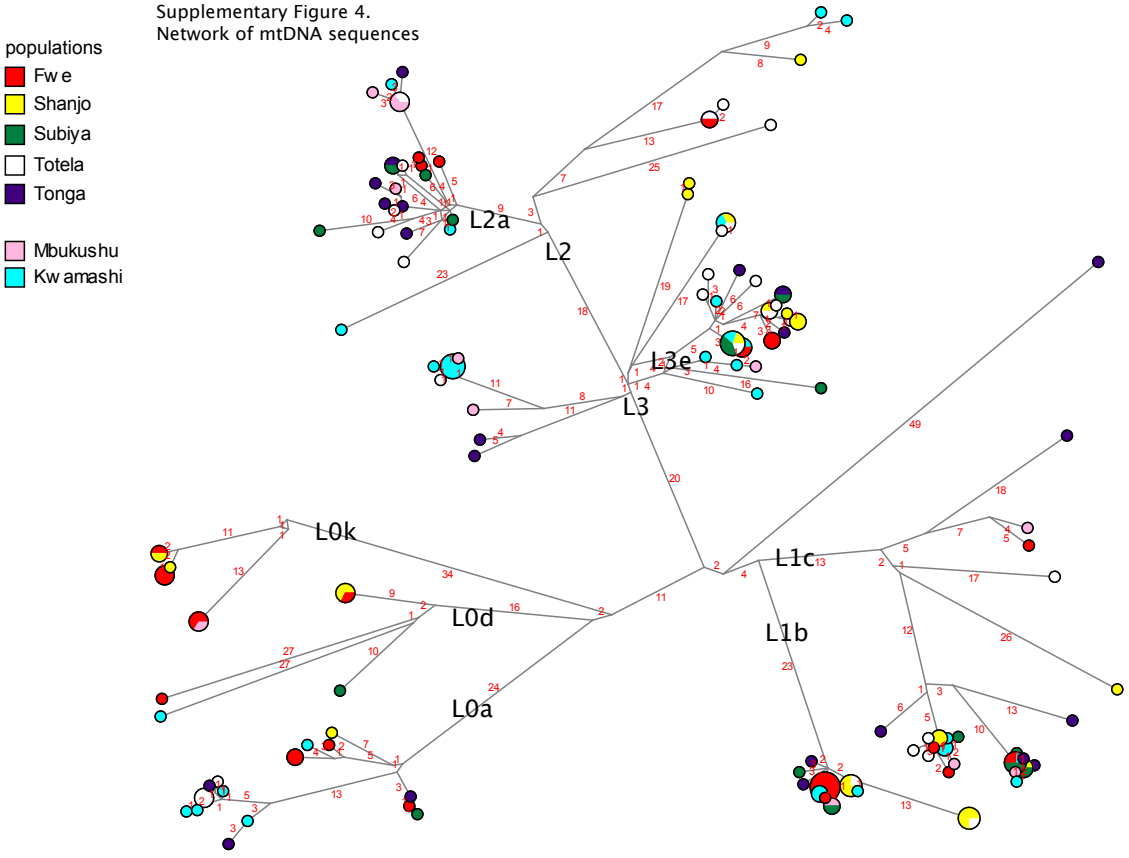



Supplementary Figure 2.
Network of STR haplotypes
populations
Fwe
Shanjo
Subiya
Totela
Tonga
Mbukushu
Kwamashi

Supplementary Figure 3
frequency of shared haplotypes, mtDNA



Supplementary Figure 4.
Network of mtDNA sequences



Supplementary Table 1: Additional information on location sampled

village/town name	lon	lat	number of individuals sampled
Ilonga	24.01525112	-17.04006054	1
Imusho	23.2241122	-17.52991313	15
Itufa	23.31256424	-15.85859039	1
Kalabo	22.68611723	-14.99844653	8
Kaoma	24.79837392	-14.80465642	3
Katunda	24.68049469	-14.83310745	1
Litambuy	23.30912431	-16.06673853	1
Liyuyelo	23.14424942	-15.24126976	1
Mambolomoka	22.14452	-16.12411	16
Mambumbu	23.15387185	-15.17652836	2
Mangango	24.5116296	-14.65710022	1
Masese	24.64445251	-17.28570621	10
Mbaala	23.370763	-17.579957	13
Mbao	24.5116296	-14.65710022	2
Mbume	23.24972629	-16.47654721	15
Mongu	23.14539942	-15.23299187	22
Mulele	23.11455251	-16.6263951	9
Mulonga	22.57937838	-16.36417958	2
Mutomena	23.10983	-16.74916	7
Nalisa	24.25700951	-17.46529369	2
Nalwashi	23.11199728	-16.39024805	3
Namatindi	22.75066366	-15.21167887	12
Nombe	23.18301539	-16.66045245	8
Sambulo	24.97710795	-16.58184486	8
Senanga	23.295865	-16.118018	1
Sesheke	24.27968407	-17.46741934	16
Shango	22.09224924	-16.32090147	6
Sichili	24.95532565	-16.71121686	2
Singembela	23.02576728	-17.31589464	9
Sioma	23.50335874	-16.6012654	3
Winela	23.15288798	-15.24512594	5

Supplementary Table 2: Y chromosome genotypes

ID	population	haplogroup	DYS19	DYS385a	DYS385b	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385sum
ZAM001	Mbukushu	E2	14	15	19	12	28	25	11	11	13	14	11	11	34
ZAM006	Tonga	E1b1a8	15	15	19	13	31	21	10	11	13	14	11	13	34
ZAM020	Tonga	E1b1a8	15	15	19	13	31	21	10	11	13	14	11	14	34
ZAM022	Subiya	B2b	16	15	16	13	29	20	10	11	13	15	10	8	31
ZAM043	Subiya	E1b1a8	15	15	16	13	31	21	11	11	13	14	11	11	31
ZAM058	Kwamashi	E1b1a8	15	16	16	13	31	21	11	11	13	14	11	11	32
ZAM059	Tonga	B2a1a	17	10	11	14	32	23	10	11	13	14	10	13	21
ZAM067	Tonga	E1b1a7a	16	15	18	14	31	21	10	11	15	15	12	13	33
ZAM081	Subiya	E1b1a8	14	16	18	13	31	21	10	11	14	14	11	12	34
ZAM126	Tonga	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	11	33
ZAM138	Tonga	E1b1a7a	16	18	19	12	29	21	10	11	15	14	11	12	37
ZAM140	Tonga	E1b1a7a	16	17	19	13	30	22	11	11	15	14	11	13	36
ZAM143	Kwamashi	E1b1a8	15	15	21	13	30	21	10	11	13	14	11	11	36
ZAM146	Shanjo	E1b1a7a	16	16	18	13	30	21	10	11	15	13	11	11	34
ZAM148	Shanjo	E1b1a7a	16	18	18	13	29	21	10	11	15	14	11	11	36
ZAM151	Mbukushu	E2	14	14	19	12	29	24	10	11	13	14	11	11	33
ZAM155	Totela	E1b1a8	15	14	17	13	32	21	11	11	13	14	11	11	31
ZAM159	Totela	E1b1a	15	13	16	13	30	21	10	11	14	14	11	12	29
ZAM167	Tonga	E1b1a8	15	16	19	13	31	21	11	11	13	14	11	11	35
ZAM171	Tonga	E1b1a8	15	16	16	13	31	21	11	11	13	14	11	11	32
ZAM174	Tonga	E1b1a	17	14	15	12	30	21	10	11	15	14	11	12	29
ZAM176	Subiya	E1b1a7a	16	17	17	14	31	21	10	11	14	14	11	12	34
ZAM179	Subiya	E1b1a8	15	15	19	13	31	21	10	11	13	14	11	11	34
ZAM180	Tonga	E1b1a	15	15	16	12	29	21	11	11	13	14	11	13	31
ZAM181	Tonga	E1b1a8	15	16	17	13	31	21	11	11	13	14	12	11	33
ZAM183	Tonga	E1b1a8	15	15	19	13	30	21	10	11	14	14	11	12	34
ZAM187	Subiya	E1b1a7a	16	17	18	13	31	21	10	11	15	14	11	12	35
ZAM190	Tonga	B2b	14	11	12	12	27	22	10	11	14	14	11	13	23
ZAM201	Tonga	E1b1a8	15	17	18	13	31	21	10	11	13	14	11	11	35
ZAM218	Tonga	E1b1a7a	16	17	18	13	30	21	10	11	14	14	11	12	35
ZAM219	Tonga	E1b1a8	15	16	17	13	31	21	10	11	12	14	11	11	33
ZAM231	Tonga	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	11	33
ZAM232	Tonga	E1b1a8	15	15	19	13	30	21	11	11	14	14	11	11	34
ZAM234	Kwamashi	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	13	33
ZAM235	Shanjo	E1b1a8	15	14	17	13	31	21	11	11	13	14	11	11	31
ZAM241	Kwamashi	E1b1a8	15	16	18	13	28	21	10	11	13	14	11	11	34
ZAM253	Tonga	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	11	33
ZAM318	Kwamashi	E1b1a7a	17	17	18	13	30	21	10	11	14	14	11	12	35
ZAM319	Kwamashi	E1b1a8	15	17	18	13	31	21	11	11	12	14	11	11	35
ZAM320	Kwamashi	E2	15	12	20	12	28	24	11	11	13	14	11	11	32
ZAM321	Kwamashi	E1b1a8	15	16	18	13	31	21	11	11	13	14	11	11	34
ZAM328	Kwamashi	E1b1a	14	14	16	12	29	21	10	11	14	14	11	12	30
ZAM329	Kwamashi	E1b1a8	15	16	17	13	31	21	10	11	13	14	11	11	33

ID	population	haplogroup	DYS19	DYS385a	DYS385b	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385sum
ZAM331	Kwamashi	E1b1a8	15	15	19	13	30	21	10	11	14	14	11	12	34
ZAM332	Kwamashi	E1b1a8	16	16	16	13	31	21	10	11	13	14	11	12	32
ZAM335	Fwe	E1b1a8	15	15	19	13	30	21	12	11	14	14	11	12	34
ZAM337	Kwamashi	E1b1a8	15	16	17	13	30	21	11	11	13	14	11	11	33
ZAM338	Fwe	E1b1a8	15	17	18	13	31	21	10	11	13	14	11	11	35
ZAM339	Fwe	E1b1a8	15	17	18	13	30	21	10	11	13	14	11	11	35
ZAM340	Fwe	E1b1a7a	16	17	19	13	29	22	10	11	15	14	11	11	36
ZAM341	Fwe	E1b1a7a	16	16	19	13	29	21	10	11	15	14	11	11	35
ZAM344	Fwe	E1b1a8	16	16	19	13	30	21	10	11	14	14	11	13	35
ZAM345	Fwe	E1b1a8	15	16	17	13	30	21	11	11	13	14	11	11	33
ZAM347	Mbukushu	E1b1a8	15	15	19	13	30	21	10	11	13	14	11	12	34
ZAM348	Mbukushu	E1b1a8	15	16	17	12	30	21	10	11	13	14	11	11	33
ZAM349	Fwe	E1b1a8	16	16	17	14	32	21	11	11	13	14	11	11	33
ZAM350	Fwe	E1b1a8	15	16	17	13	30	21	11	11	13	14	11	11	33
ZAM351	Fwe	E1b1a8	15	15	19	13	30	21	10	11	14	14	11	12	34
ZAM353	Totela	B2a1a	15	11	11	14	32	24	10	11	13	14	10	12	22
ZAM354	Shanjo	E1b1a7a	17	16	18	15	32	21	10	11	15	14	11	12	34
ZAM355	Fwe	E1b1a	15	14	16	13	30	21	10	11	14	14	11	12	30
ZAM356	Mbukushu	E1b1a7a	16	16	18	13	30	21	10	11	14	14	11	12	34
ZAM357	Fwe	E1b1a8	15	16	17	13	30	21	11	11	13	14	11	11	33
ZAM359	Mbukushu	E1b1a8	15	16	18	13	30	21	11	11	13	14	11	11	34
ZAM401	Totela	E1b1a8	15	16	16	13	31	21	11	11	13	14	11	11	32
ZAM429	Tonga	E1b1a8	15	17	19	13	32	21	11	11	13	14	11	11	36
ZAM435	Kwamashi	E1b1a8	15	17	18	13	31	21	11	11	12	14	11	11	35
ZAM467	Kwamashi	E1b1a7a	16	17	18	13	30	21	10	11	15	14	11	13	35
ZAM477	Totela	B2a1a	16	11	11	14	32	24	9	11	13	15	10	12	22
ZAM479	Totela	E1b1a8	15	17	17	13	31	21	10	11	13	14	12	13	34
ZAM486	Kwamashi	E1b1a8	15	17	18	13	31	21	11	11	12	14	11	11	35
ZAM487	Kwamashi	E1b1a8	16	15	19	13	30	21	10	11	14	14	11	13	34
ZAM488	Kwamashi	E1b1a7a	17	17	18	13	30	21	10	11	14	14	11	12	35
ZAM489	Kwamashi	E2	14	14	19	12	29	24	10	11	13	14	11	11	33
ZAM490	Kwamashi	E1b1a8	15	16	17	13	30	21	11	11	13	14	11	11	33
ZAM491	Kwamashi	E1b1a8	16	16	17	13	31	21	11	11	13	14	11	13	33
ZAM492	Kwamashi	E1b1a8	15	16	18	13	31	21	11	11	13	14	11	11	34
ZAM494	Kwamashi	E1b1a	15	15	15	13	30	21	10	11	13	14	11	12	30
ZAM495	Kwamashi	E1b1a8	15	16	18	14	31	21	11	11	14	14	11	12	34
ZAM500	Fwe	E1b1a7a	16	16	19	13	29	21	10	11	15	14	11	11	35
ZAM502	Shanjo	E1b1a8	15	15	20	13	30	21	10	11	14	14	11	12	35
ZAM508	Tonga	E1b1a8	15	16	18	12	29	21	10	11	13	14	11	12	34
ZAM516	Fwe	E1b1a8	15	15	19	13	30	21	12	11	14	14	11	12	34
ZAM523	Fwe	E1b1a8	15	17	17	13	31	21	10	11	13	14	11	11	34
ZAM525	Fwe	E1b1a8	15	17	18	13	31	21	10	11	13	14	11	11	35
ZAM526	Fwe	E1b1a8	15	16	17	13	32	21	10	11	13	14	11	11	33
ZAM527	Fwe	E1b1a7a	16	17	18	13	30	21	10	11	15	13	11	12	35
ZAM529	Mbukushu	E1b1a7a	16	16	17	14	33	21	10	11	14	14	11	12	33

ID	population	haplogroup	DYS19	DYS385a	DYS385b	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385sum
ZAM530	Fwe	E1b1a7a	16	16	18	13	30	21	10	11	14	14	11	13	34
ZAM531	Fwe	E2	14	14	19	12	28	23	11	11	13	14	11	11	33
ZAM532	Fwe	E1b1a8	15	16	17	12	30	21	11	11	13	14	11	11	33
ZAM533	Mbukushu	E1b1a	14	15	16	13	30	21	11	11	14	14	11	13	31
ZAM535	Fwe	E1b1a8	16	15	19	13	30	21	10	11	14	14	11	13	34
ZAM538	Fwe	E1b1a8	17	15	18	13	31	21	10	11	13	14	11	11	33
ZAM539	Mbukushu	E1b1a8	15	17	17	13	31	21	11	11	13	14	11	11	34
ZAM544	Fwe	E1b1a8	15	16	17	12	29	21	10	11	13	14	11	12	33
ZAM545	Fwe	E1b1a8	15	16	18	13	31	21	11	11	13	14	11	12	34
ZAM546	Mbukushu	E1b1a8	15	16	18	13	30	21	11	11	13	14	11	11	34
ZAM547	Fwe	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	11	33
ZAM602	Subiya	E1b1a8	15	16	18	13	31	21	11	11	13	14	11	12	34
ZAM612	Kwamashi	E2	14	14	18	12	29	24	10	11	13	14	11	11	32
ZAM616	Subiya	E1b1a8	15	16	18	13	30	21	11	11	13	14	11	12	34
ZAM635	Tonga	E1b1a7a	17	17	19	13	30	21	10	11	15	14	11	12	36
ZAM636	Tonga	E1b1a7a	15	16	19	13	29	22	10	11	15	14	11	11	35
ZAM640	Tonga	E1b1a7a	16	17	19	13	29	21	10	11	14	14	11	11	36
ZAM641	Tonga	E1b1a7a	17	17	19	13	30	21	10	11	14	14	11	12	36
ZAM645	Subiya	E1b1a8	14	16	18	13	31	21	10	11	13	14	11	11	34
ZAM675	Totela	B2b	14	11	13	12	27	22	10	11	14	14	11	13	24
ZAM694	Tonga	E1b1a7a	16	16	17	13	30	21	10	11	14	14	11	11	33
ZAM704	Shanjo	E1b1a8	15	17	17	13	31	21	11	11	13	14	11	11	34
ZAM705	Shanjo	E1b1a7a	15	15	18	14	31	21	10	11	14	14	11	12	33
ZAM710	Totela	E1b1a7a	16	17	18	13	30	22	10	11	14	14	11	12	35
ZAM711	Tonga	R	15	13.2	15	14	31	23	11	13	13	14	12	13	28.2
ZAM714	Tonga	E1b1a8	15	15	17	14	31	21	10	11	13	14	11	12	32
ZAM715	Tonga	E1b1a7a	16	17	19	13	30	22	11	11	15	14	11	13	36
ZAM717	Totela	E1b1a	15	15	16	12	29	21	10	11	14	14	11	12	31
ZAM719	Mbukushu	E1b1a8	15	15	19	13	30	21	10	11	13	14	11	12	34
ZAM720	Totela	E1b1a8	15	15	19	13	30	21	10	11	13	14	11	12	34
ZAM721	Shanjo	E1b1a	15	15	16	12	29	21	10	11	14	14	11	12	31
ZAM724	Shanjo	E1b1a8	16	16	17	13	32	21	10	11	13	14	11	14	33
ZAM726	Shanjo	E1b1a8	16	16	18	13	31	21	11	11	13	15	11	11	34
ZAM731	Kwamashi	E1b1a7a	16	17	18	13	31	21	10	11	15	14	11	12	35
ZAM735	Totela	E1b1a8	16	16	17	13	30	21	11	11	13	14	11	11	33
ZAM736	Shanjo	E1b1a8	16	15	20	13	30	21	10	11	14	14	11	12	35
ZAM738	Shanjo	E1b1a8	16	16	18	13	31	21	11	11	13	14	11	11	34
ZAM739	Totela	E1b1a8	15	14	17	13	32	21	11	11	13	14	11	11	31
ZAM741	Shanjo	E1b1a8	15	16	17	13	31	21	11	11	13	14	11	11	33
ZAM742	Totela	E1b1a	15	13	16	13	30	21	10	11	14	14	11	12	29
ZAM755	Tonga	E1b1a7a	16	17	18	14	31	21	10	11	15	14	11	11	35
ZAM761	Subiya	E1b1a8	15	16	17	13	29	21	11	11	13	14	11	11	33
ZAM764	Subiya	E1b1a8	15	16	17	13	29	21	11	11	13	14	12	11	33
ZAM769	Tonga	E1b1a7a	16	17	18	14	27	21	10	11	15	14	11	11	35

Supplementary Table 3: matrix of RST distances (y chromosome STR)

	Fwe	Kwamashi	Mbukushu	Shanjo	Subiya	Tonga	Totela
Fwe	0						
Kwamashi	0.0037	0					
Mbukushu	-0.01622	-0.03667	0				
Shanjo	-0.00786	-0.00557	-0.01696	0			
Subiya	0.0799	0.0237	0.03002	0.00834	0		
Tonga	-0.00573	-0.00149	-0.0135	-0.02134	-0.02981	0	
Totela	0.00482	0.04743	-0.01265	0.00122	0.06597	-0.00882	0

p values

	Fwe	Kwamashi	Mbukushu	Shanjo	Subiya	Tonga	Totela
Fwe	*						
Kwamashi	0.33	*					
Mbukushu	0.65	0.88	*				
Shanjo	0.58	0.52	0.65	*			
Subiya	0.04	0.23	0.23	0.31	*		
Tonga	0.49	0.42	0.55	0.83	0.65	*	
Totela	0.35	0.08	0.60	0.38	0.10	0.46	*

Supplementary Table 4: matrix of PhiST distances (mtDNA sequences)

	Fwe	Kwamashi	Mbukushu	Shanjo	Subiya	Tonga	Totela
Fwe	0						
Kwamashi	0.05262	0					
Mbukushu	0.04744	0.00327	0				
Shanjo	0.00731	0.01788	0.03777	0			
Subiya	0.04386	0.01039	-0.01799	0.01767	0		
Tonga	0.04116	-0.00234	-0.0296	0.02373	-0.0245	0	
Totela	0.06256	-0.00919	-0.00128	0.02058	0.00227	-0.01422	0

p values

	Fwe	Kwamashi	Mbukushu	Shanjo	Subiya	Tonga	Totela
Fwe	*						
Kwamashi	0.004	*					
Mbukushu	0.06	0.38	*				
Shanjo	0.28	0.13	0.09	*			
Subiya	0.05	0.25	0.62	0.19	*		
Tonga	0.03	0.46	0.88	0.10	0.85	*	
Totela	0.003	0.67	0.43	0.12	0.37	0.75	*

Chapter 8

Paper III: Ancient substructure in early mtDNA lineages of southern Africa

This chapter includes the paper "**Ancient substructure in early mtDNA lineages of southern Africa**" written by Barbieri, Chiara, Mário Vicente, Jorge Rocha, Sununguko W. Mpoloka, Mark Stoneking, and Brigitte Pakendorf, as it appears in the published version on *American Journal of Human Genetics*, 2013, 92(2): 285-292

REPORT

Ancient Substructure in Early mtDNA Lineages of Southern Africa

Chiara Barbieri,[1,7,*] Mário Vicente,[3,4] Jorge Rocha,[4,5] Sununguko W. Mpoloka,[6] Mark Stoneking,[2] and Brigitte Pakendorf[1,8]

Among the deepest-rooting clades in the human mitochondrial DNA (mtDNA) phylogeny are the haplogroups defined as L0d and L0k, which are found primarily in southern Africa. These lineages are typically present at high frequency in the so-called Khoisan populations of hunter-gatherers and herders who speak non-Bantu languages, and the early divergence of these lineages led to the hypothesis of ancient genetic substructure in Africa. Here we update the phylogeny of the basal haplogroups L0d and L0k with 500 full mtDNA genome sequences from 45 southern African Khoisan and Bantu-speaking populations. We find previously unreported subhaplogroups and greatly extend the amount of variation and time-depth of most of the known subhaplogroups. Our major finding is the definition of two ancient sublineages of L0k (L0k1b and L0k2) that are present almost exclusively in Bantu-speaking populations from Zambia; the presence of such relic haplogroups in Bantu speakers is most probably due to contact with ancestral pre-Bantu populations that harbored different lineages than those found in extant Khoisan. We suggest that although these populations went extinct after the immigration of the Bantu-speaking populations, some traces of their haplogroup composition survived through incorporation into the gene pool of the immigrants. Our findings thus provide evidence for deep genetic substructure in southern Africa prior to the Bantu expansion that is not represented in extant Khoisan populations.

Sub-Saharan Africa harbors the deepest-rooting lineages of human mitochondrial DNA (mtDNA), in agreement with an African origin of modern humans supported by both fossil and genetic evidence.[1–4] Several studies concurred in placing the root of the mtDNA phylogeny in the southern half of the continent,[5–7] and two deep-rooting clades of this phylogeny—haplogroups L0d and L0k—have been unanimously associated with so-called Khoisan populations.[6–9] The generic term "Khoisan" covers hunter-gatherer and pastoralist populations of southern Africa who speak non-Bantu indigenous languages and share some linguistic features (one of the most characteristic being the heavy use of click consonants in their languages); however, these similarities might be the effect of contact.[10] Haplogroups L0d and L0k are present nearly exclusively in Khoisan populations and neighboring Bantu-speaking populations that have been in documented close contact with them;[11–14] the only known exceptions are sporadic occurrences of haplogroup L0d in East Africa (e.g., in the Sandawe from Tanzania)[7] and in an individual from Yemen[6] as well as an individual from Kuwait[6] who belongs to haplogroup L0k. Specialists recognize three independent language families among Khoisan, namely Tuu, Kx'a, and Khoe-Kwadi,[15–17] which are spoken by a large number of different ethnolinguistic groups comprising both foragers and pastoralists. The forager populations of the central Kalahari, who speak languages belonging to the Tuu and Kx'a families, are assumed to be the descendants of autochthonous Late Stone Age populations, whereas the Khoe-Kwadi languages may have been brought to the area by pastoralist populations around 2,000 years ago.[18–20] The populations speaking Bantu languages, in contrast, are known for their expansion over almost half the African continent and are associated with the concomitant spread of the Bantu language family, an agricultural lifestyle, and iron technology.[3,21,22] Archeological data suggest that they may have reached southern Africa not earlier than 2,000–1,200 years ago,[3,23,24] where they met populations who were probably ancestral to current Khoisan populations.

The most recent comprehensive study that focused on the deepest-rooting lineages of the mtDNA phylogeny was undertaken by Behar et al.,[6] who analyzed a total of 624 full mtDNA sequences belonging to haplogroup L*(xM,N). Although this was the first substantial collection of complete mtDNA genome sequences from Africa, some limitations arose from the inclusion of a large number of sequences from diverse published sources that were not always of high quality; furthermore, for some sequences the source population or the country of origin was not clearly specified. Nevertheless, the sequences considered in that study still represent the vast majority of the haplogroup L*(xM,N) data set included in the most recent version of Phylotree (Build 15, September 2012[25]), a comprehensive database of mtDNA genome sequences that is periodically updated when more data become available.

[1]Max Planck Research Group on Comparative Population Linguistics, [2]Department of Evolutionary Genetics, MPI for Evolutionary Anthropology, Leipzig 04103, Germany; [3]STAB VIDA, Investigação e Serviços em Ciências Biológicas, Lda, Oeiras 2780-182, Portugal; [4]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto, Vairão 4485-661, Portugal; [5]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto 4169-007, Portugal; [6]Department of Biological Sciences, University of Botswana, Gaborone UB 0022, Botswana
[7]Present address: Department of Evolutionary Genetics, MPI for Evolutionary Anthropology, Leipzig 04103, Germany
[8]Present address: Laboratoire Dynamique du Langage, UMR5596, CNRS and Université Lyon Lumière 2, Lyon 69007, France
*Correspondence: chiara_barbieri@eva.mpg.de
http://dx.doi.org/10.1016/j.ajhg.2012.12.010.

It thus represents the most accessible resource for studying mtDNA variation and is a widely used reference for mtDNA nomenclature.[26]

Behar et al.[6] focused particularly on the root of the phylogeny, i.e., the age and variability of the Khoisan-specific haplogroups L0d and L0k, with the aim of investigating the most likely model of origin and isolation of Khoisan populations. With their data they were able to suggest a time frame for the dispersal of the main lineages and the split of Khoisan and other modern humans, which they dated not later than 90 thousand years ago (kya); furthermore, they suggested that the early human settlement of Africa was matrilineally structured. These hypotheses are relevant for the interpretation of early human demography and evolution; however, their results were substantially limited by the fact that only one ethnolinguistically undefined "Khoisan" sample of 38 individuals was included, thereby missing the potentially immense variability of the different ethnolinguistic populations subsumed under the generalized label Khoisan. In addition, only 30 sequences from haplogroup L0d and 7 from L0k were included, representing only a small and probably incomplete fraction of the overall variation in these haplogroups.

We here report analyses of 500 mtDNA genome sequences belonging to haplogroups L0d and L0k, of which 15 have already been published in Barbieri et al.,[14] leading to a more than 10-fold increase in the available complete mtDNA genome sequences from southern Africa (Phylotree ver. 15[25]). With this rich data set, we aim to elucidate the phylogenetic relationships, the patterns of diversity, and the distribution of these relatively understudied haplogroups that represent some of the deepest-rooting lineages in the maternal phylogeny of modern humans. The broader data set from which the subset of L0d and L0k sequences was chosen consists of mtDNA genome sequences generated from saliva samples collected in Botswana, Namibia, Zambia, and Angola after prior approval by the relevant institutional review boards and with the consent of the donors after the aims of the study had been explained to them with the help of local translators, where necessary. Details of the samples have been described elsewhere.[11,27,28] The sequence data set analyzed here comprises 45 ethnolinguistic groups, who speak Khoisan languages belonging to all three accepted language families as well as different Bantu languages; individuals were assigned to populations on the basis of the ethnic affiliation of their maternal grandmother (Table S1 available online).

Libraries enriched for mtDNA[29,30] were sequenced on the Illumina GAIIx platform, resulting in an average 400-fold coverage. Sequences were manually checked with BioEdit and read alignments were screened with ma[31] to exclude alignment errors and confirm indels. The two poly-C regions (np 303–315, 16,183–16,194) were excluded from the analysis. To minimize the impact of missing data, we applied imputation and resolved unknown positions by comparison to at least two otherwise identical haplotypes in the data set. Before imputation, 74 sequences included positions with missing data; after imputation, only 26 sequences still had missing positions. In the final alignment, 32 positions were left with an unknown nucleotide call (26 of which corresponded to polymorphic sites) and were excluded from the analyses (see Table S2 for a list of the excluded positions). Basic haplogroups were defined with the web tool Haplogrep.[26] Mutations that did not fit the overall phylogeny were checked manually in the read alignments to exclude the possibility of erroneous base calls. Although we took into account published data on the frequency of haplogroups L0d and L0k, only the 500 sequences that were generated with the same technology and from individuals for whom we know the place of sampling and ethnicity were included in the phylogenetic analyses. We did not include previously published sequences, because they do not add substantial information to our analysis and often pose problems because of missing positions[6] or missing ethnolinguistic information. The only exceptions are the L0k2 sequence from Yemen[6] and the six L0d3 sequences from South Africa, Kuwait,[6] and Tanzania,[5] which we included to clarify the structure in L0k and L0d3 discussed below.

First, we compared the frequency and distribution of haplogroups L0d and L0k in our data set and in the available literature (where in most cases haplogroups were assigned based on partial mtDNA sequence variation and/or RFLP typing; cf. Table S1 and Figure S1 for details) and plotted the frequencies of each haplogroup (Figures 1A and 1B) with the software Surfer ver. 10.4.799 (Golden Software). The maps show a concentration of both L0d and L0k in the southern part of the continent, with L0d present in high frequency in populations from South Africa, Namibia, and Botswana, and sporadically (<5%) in some populations of Zambia, Mozambique, and Angola, as well as in the Sandawe from Tanzania. The highest frequencies (90%–100%) are found in Khoisan foragers of central Botswana, as well as in South African populations with Khoisan ancestry.[32,33] In general, other studies did not distinguish between L0d and L0k as typical "Khoisan" lineages; and yet, interestingly, the distribution of haplogroup L0k is far more restricted than that of L0d, with a maximum frequency of 33% in the !Xuun foragers of Namibia; it is also found in frequencies >10% in several populations of foragers in Botswana and Namibia who speak languages belonging to all three Khoisan linguistic families (see Table S1), as well as in the Bantu-speaking Fwe from southwestern Zambia.

We next reconstructed a phylogeny of the L0d and L0k mtDNA genome sequences from the most probable tree out of 10 million MCMC chains with BEAST (v1.7.2[34]) and identified the mutations defining different branches by viewing the aligned sequences in BioEdit in comparison to the Reconstructed Sapiens Reference Sequence (RSRS[35]). The node branches were dated with the mutation rate of

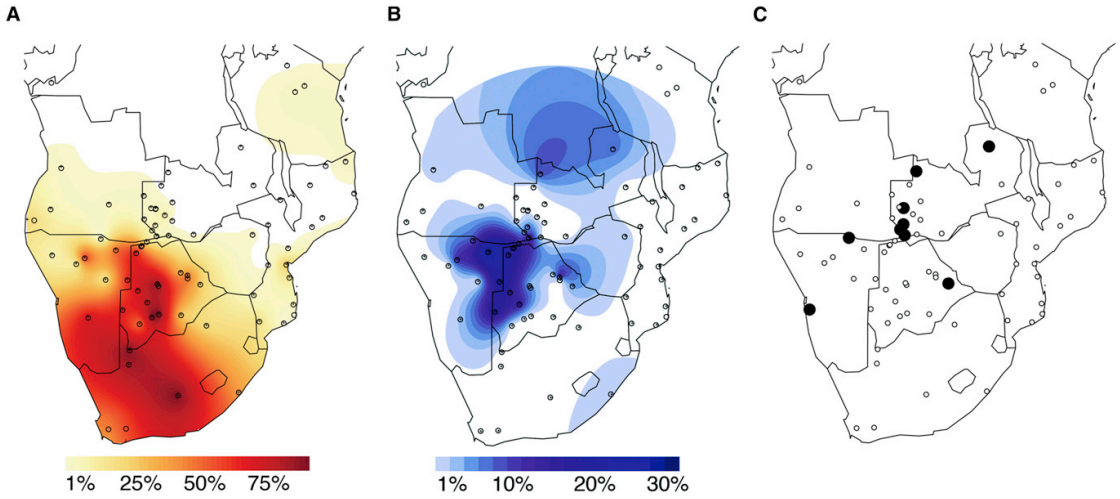


Figure 1. Surfer Maps Displaying the Spatial Distribution of Haplogroup Frequencies
Dots indicate sample locations.
(A) Haplogroup L0d.
(B) Haplogroup L0k. Note that the scale in (B) is different from that in (A).
(C) Presence of haplogroups L0k1b and L0k2 in southern Africa (large black dots). The actual sampling location of one Topnaar Nama individual with haplogroup L0k1b is shown here; in (A) and (B) this individual was included with the general Nama population sample.

1.26×10^{-8} for the coding region only,[36] which makes our estimates comparable to those from Behar et al.[6] The complete tree of sequences showing mutations that characterize the major branches is available in the Supplemental Data (Figure S2); further discussion of some of these mutations is found in Table S3. Figure 2 summarizes the tree topology and the TMRCA of lineages, with confidence intervals indicated for the major nodes.

The tree coalesces 145 kya (95% C.I.: 118–179 kya), corresponding to the time of split between L0d and L0k. From the topology of the tree, different sublineages can be distinguished for both the L0d and L0k haplogroups. For L0d, three main branches (L0d3, L0d1, and L0d2) separate around 95 kya (95% C.I.: 79–121 kya), whereas L0k splits into L0k1 and L0k2 approximately 40 kya (95% C.I.: 28–53 kya). The first branch of L0d is the uncommon L0d3, which is found in a population with South African Khoisan ancestry (Karretjie People) at 13% and in a Coloured population at 10%,[32] as well as being attested in one undefined Khoi and one individual from Kuwait[6] and three Sandawe and one Burunge from Tanzania (our identification, based on sequences from Gonder et al.[5]). In our data set, it is found in only five individuals (two Nama and one Hai||om, who speak Khoe languages, and two Kgalagadi, who speak a Bantu language). As can be seen from the tree (Figure 2), L0d3 splits into two branches (L0d3a and L0d3b) 45 kya (95% C.I.: 30–61 kya), with eight mutations defining L0d3b (Figure S2 and Table S3). Interestingly, this split reflects geographic substructure: L0d3a is restricted to East Africa and the Middle East, being found in the individuals from Kuwait and Tanzania, and L0d3b is restricted to southern Africa, being found in the five individuals of our data set plus the Khoi sequence published by Behar et al.[6]

L0d1 is the most common subhaplogroup: it is present in all Khoisan populations, all Bantu-speaking populations of our data set from Botswana and Namibia, and a few individuals from Bantu-speaking populations of Zambia and Angola. It coalesces approximately 55 kya (95% C.I.: 44–68 kya) and comprises two branches, of which the first includes haplogroups L0d1a and L0d1c. L0d1a is a monophyletic clade; however, two sites, namely T199C and C16266A, previously assumed to define this clade, pose problems for reconstructing the history of mutations (see Table S3 for details).

In L0d1c, substantial variation emerges from our expanded data set that pushes the coalescence date back to 32 kya (95% C.I.: 24–41 kya), 10 ky older than previously estimated.[6] A low posterior probability is associated with the first nodes; these are represented by paraphyletic clades that are characterized by a large number of private mutations. In addition to the paraphyletic clades, L0d1c contains two monophyletic clades. The first is the previously attested L0d1c1, which is defined by only two of the mutations previously associated with it (Figure S2 and Table S3). The second monophyletic clade in L0d1c, which we here define as L0d1c2a, is represented by six haplotypes and supported by four mutations (Figure S2).

The second basal branch of L0d1 is subhaplogroup L0d1b, which coalesces approximately 45 kya (95% C.I.: 35–56 kya) and is thus 10 ky older than previously estimated.[6] As shown by our data, this is characterized by

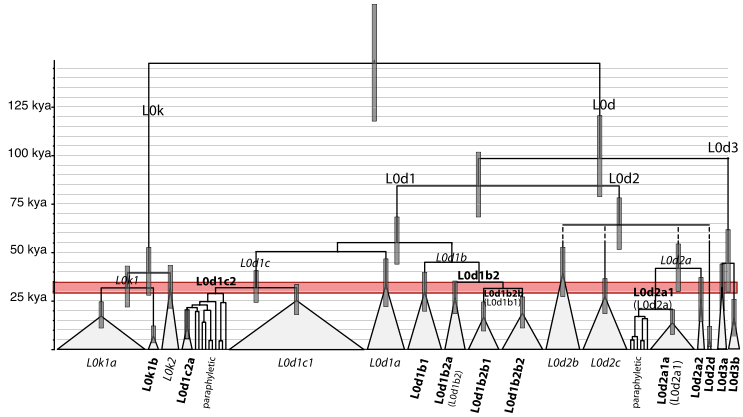


Figure 2. Simplified Tree Topology for the Major Lineages of L0d and L0k, Based on Coding Region Sequences and with Time Scale Indicated
Previously undetected branches are labeled in bold font; when a previously reported branch is renamed, the old label is given in brackets. Confidence intervals for the TMRCA of the major nodes are indicated by vertical bars. The red shading highlights the time span that was associated with the deterioration of climate in the central Kalahari area.

only one mutation, T3618C, splitting immediately into several subhaplogroups. Because the haplogroup previously labeled L0d1b1 is only the second of three hierarchical splits, the nomenclature is revised as follows: we propose to assign the label L0d1b1 to the first branch, which is characterized by four mutations (Figure S2). This is followed by a branch that we label L0d1b2, which is defined by several of the mutations previously assigned to L0d1b (Figure S2). This splits into L0d1b2a—represented by a monophyletic clade labeled in the latest version of Phylotree (ver. 15[25]) as L0d1b2—and L0d1b2b, which was previously defined as L0d1b1 and again contains two subclades: L0d1b2b1 and L0d1b2b2 (Figure 2).

Haplogroup L0d2, which coalesces around 65 kya (95% C.I.: 52–78 kya), is less common than L0d1 and is found at frequencies >10% only in populations from Botswana (mainly Khoisan foragers, but also the Bantu-speaking Tswana and Kgalagadi) and in the pastoralist Nama and forager Hai||om of Namibia (Table S1). With our data the diversity of this part of the tree is substantially increased: the earliest splits appear almost simultaneously, and we are unable to cleanly resolve the phylogeny (with a very low posterior probability for each of the nodes). From these splits arise four monophyletic clades: the previously defined L0d2a, L0d2b, and L0d2c, as well as a previously unreported branch that we here define as L0d2d. Although the clade previously defined as L0d2c is not changed by our data, subhaplogroup L0d2a is much more diverse than previously known, as is also reflected by our TMRCA estimate of approximately 40 ky (95% C.I.: 30–54 kya) versus the previous estimate of 9 ky.[6] Some of the mutations previously thought to be characteristic of L0d2a actually define a subclade of L0d2a (Figure S2), which we here call L0d2a1, whereas the branch previously called L0d2a1 is shown to be a subclade of L0d2a1 and is therefore correspondingly labeled L0d2a1a (Figure 2). Two further previously undetected branches emerge from our data: L0d2a2, a sister clade of L0d2a1, and the very divergent subclade of L0d2 mentioned above, which we here define as L0d2d.

L0k separated from L0d approximately 145 kya (95% C.I.: 118–179 kya) and has a TMRCA of approximately 40 ky (95% C.I.: 28–53 kya). The majority of L0k lineages can be unambiguously assigned to the branch previously defined as L0k1;[6] however, with our expanded data set we are now able to identify variation within L0k1, which consists of two sister clades: L0k1a (proposed in the latest version of Phylotree based on a sequence from Barbieri et al.[14]) and L0k1b (defined here), which we find in four individuals of our data set (Figure 2). Haplogroup L0k2 had previously been found in only one ethnolinguistically undefined individual from Yemen;[6] in our data set, nine individuals from Bantu-speaking populations of Zambia and northeast Botswana belong to this haplogroup (Table S1).

The branching structure of the mtDNA phylogeny may have been shaped by events of climate change occurring at different periods in southern Africa. Thus, the deep splits in haplogroups L0k, L0d1b2, and L0d1c and the diversification of haplogroups L0d1a, L0d1b1, and L0d2c, which all happened approximately 30–40 kya, might be associated with the deterioration of climate in the central Kalahari area ~35–27 kya.[37] The aridification of this area, which was partly concurrent with a milder and more moist climate in the Eastern Cape,[38] would have led to the dispersal of foragers to more suitable environments, with the subsequent separation and isolation of populations leading to the diversification of the mtDNA tree. Conversely, the shallow branches of L0k1a, L0d1c1, and L0d2a1a (Figure S2), which started to diversify 15–10 kya, suggest population expansions that may be associated with the postglacial amelioration of the climate and concomitant environmental diversification.[39] Such expansions are also visible in the Bayesian Skyline Plots generated with BEAST[34] (Figure S3): thus, L0k shows a signal of expansion at ~5 kya and L0d1 and L0d2 expand ~3–4 kya. Archeological evidence suggests an increase in population size beginning approximately 14 kya that peaked ~4 ky,[18] in good accordance with the genetic evidence.

The separation between L0k1a, L0k1b, and L0k2 is particularly evident from a network (Figure 3), where different patterns of diversity characterize the three haplogroups: whereas L0k1a has short branches and shows

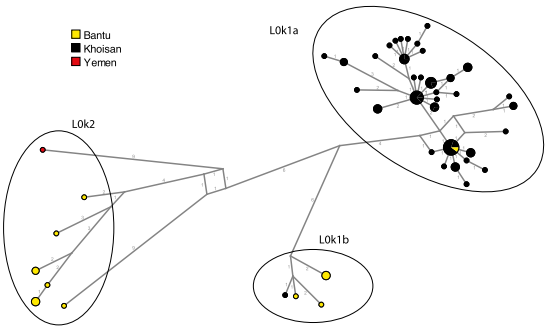


Figure 3. MJ-Network of L0k Based on Full Sequences

signals of expansion in its star-like pattern, L0k1b and especially L0k2 are composed of long separate branches and unique haplotypes that might represent the remains of an ancient and richer diversity. Although L0k was previously tentatively associated with a relatively late immigration of pastoralist Khoe populations rather than with central Kalahari foragers,[40] our more comprehensive data demonstrate that this haplogroup, together with L0d, is in fact characteristic of the central Khoisan genetic profile, being absent only from South Africa. Seventy sequences are identified as belonging to L0k1a, coming predominantly from the Khoisan populations of Botswana (plus the Khoe-speaking Hai||om of Namibia) that also carry high frequencies of L0d1. In contrast, the distribution of L0k1b and L0k2 is highly restricted, being found only in the northern range of the L0d/L0k distribution, predominantly in Zambia (Figure 1C). Interestingly, and in contrast to L0k1a, L0k1b and L0k2 are found almost exclusively in Bantu-speaking populations (Figure 3; Table S1), who probably acquired it after contact with Khoisan groups; the only exceptions are an individual from Yemen with L0k2 and a Topnaar Nama (speaking a Khoe language) with L0k1b.

The near-exclusive presence of L0k1b and L0k2 haplotypes in Bantu-speaking populations rather than in Khoisan groups requires an explanation. The early separation from L0k1a of L0k2 (almost 40 kya) and L0k1b (around 30 kya) and the absence of recent diversification and branching might in principle suggest a very ancient incorporation into Bantu-speaking populations and subsequent isolation of these relic haplotypes. However, because there is no evidence for people speaking Bantu languages in southern Africa before 2,200 years ago,[3] and because L0k is not found in the place of origin of the ancestors of the Bantu-speaking populations in western and central Africa,[41,42] the contact between Khoisan and Bantu is unlikely to predate this period.

There are two possible alternative explanations. (1) These L0k1b and L0k2 lineages were incorporated into the Bantu-speaking populations through contact with now-extinct populations whose mtDNA haplogroup composition differed from that found in extant Khoisan groups in that they possessed the divergent L0k types. (2) The ancestors of extant Khoisan populations did possess the divergent L0k types and thus contributed them to Bantu-speaking populations (along with L0d and L0k1a lineages), but the haplogroup composition of the ancestral Khoisan groups was subsequently affected by drift, leading to the loss of L0k1b and L0k2.

We investigated these two alternative scenarios by assessing the probability that L0k1b and L0k2 would be lost from a Khoisan population by drift while being retained in Bantu-speaking populations after incorporation through contact. To do this, we assumed a relatively small effective population size for the Khoisan foragers, who throughout their history have lived in small nomadic bands,[18,43] and a 10- to 100-fold higher effective population size for the Bantu-speaking food-producing groups. We simulated the variation in frequency of L0k1b/L0k2 for both the Khoisan and the Bantu virtual populations under three scenarios: (1) assuming $N_e = 50$ for Khoisan and $N_e = 5{,}000$ for Bantu speakers; (2) assuming $N_e = 100$ for Khoisan and $N_e = 1{,}000$ for Bantu speakers; and (3) assuming $N_e = 1{,}000$ for Khoisan and $N_e = 10{,}000$ for Bantu speakers. All the tests were iterated 10,000 times over 71 generations (about 2,000 years assuming 28 years per generation[44]), and the final haplogroup composition was checked in a random sample of 30 individuals from each population.

First we evaluated the likelihood of losing L0k1b/L0k2 for a range of initial frequencies of L0k1b/L0k2 in Khoisan (Table 1). The probability of losing L0k1b/L0k2 in the Khoisan is at least 95% for initial frequencies of not more than 3% for $N_e = 50$, 1.5% for $N_e = 100$, and 0.3% for $N_e = 1{,}000$. We next investigated the minimum amount of unidirectional migration from the Khoisan population necessary to ensure the presence of L0k1b/L0k2 in Bantu-speaking populations in more than 5% of the 10,000 simulated cases (Table 2). To do so, we chose three initial frequencies of L0k1b/L0k2 from Table 1 for $N_e = 50$, $N_e = 100$, and $N_e = 1{,}000$ that resulted in loss in more than 90% of the simulations, and we created hypothetical ancestral Khoisan populations carrying those frequencies; the rest of the population was assumed to carry other Khoisan haplogroups (i.e., L0d or L0k1a). Finally, we determined the frequency of Khoisan haplogroups other than L0k1b/L0k2 in the Bantu population after 71 generations (Table 2).

Overall, the results of these analyses indicate that there is a high probability of loss of L0k1b/L0k2 lineages in ancestral Khoisan populations if their initial frequency was not more than 1.5% (for $N_e = 100$, Table 1). With this initial frequency, a Bantu-speaking population with $N_e = 1{,}000$ could have retained the L0k1b/L0k2 lineages with a migration rate of 0.012 (Table 2). However, with this migration rate we would expect to find other Khoisan haplogroups (L0d or L0k1a) at a frequency of at least 57% in extant Bantu-speaking populations—and yet the frequency of L0d/L0k1a haplogroups in the Bantu-speaking

Table 1. Values of the Initial Frequency of L0k1b/L0k2 in the Simulated Khoisan Population with Associated Probabilities of Losing Them after 71 Generations, Based on 10,000 Iterations

	Initial Frequency of L0k1b/L0k2 in Khoisan									
	0.05	**0.03**	**0.02**	**0.015**	**0.01**	**0.006**	**0.005**	**0.004**	**0.003**	**0.002**
$N_e = 50$	93.3	96.3	96.9	100	100	100	100	100	100	100
$N_e = 100$	86	91.4	94.1	97.2	97.2	100	100	100	100	100
$N_e = 1{,}000$	45	63.4	74.3	80	85.8	91.9	92.9	94.3	95.4	97

Three hypothetical cases are considered, with an N_e for Khoisan of 50, 100, and 1,000. Probabilities are expressed in percent.

populations with L0k1b/L0k2 haplotypes is significantly lower in all cases (chi-square test p values < 0.05 for all Bantu-speaking populations). The scenario based on an N_e of 1,000 for Khoisan (Tables 1 and 2) appears even more unlikely. Here the maximum frequency of L0k1b/L0k2 in the ancestral Khoisan population that could be lost by drift with a probability >95% is 0.3% (Table 1); a migration rate of 0.023 is needed in order to retain L0k1b/L0k2 haplogroups in the Bantu-speaking group, which would in turn lead to the incorporation of at least 87% other Khoisan haplogroups (Table 2). The only scenario that would lead to an incorporation of L0d/L0k1a in the Bantu-speaking immigrants compatible with the observed values are that of a Khoisan population of size 50 in contact with a Bantu-speaking population of size 5,000. In this case, if the initial frequency of L0k1b/L0k2 in the Khoisan group was 3%, it could have been incorporated into the Bantu-speaking population with a migration rate of 0.002 and subsequently been lost by drift in the Khoisan group. With such a migration rate, one would expect to find 13% other Khoisan haplogroups in the Bantu speakers, a value compatible with what is found in the Bantu-speaking populations carrying the divergent L0k lineages (Table S1). However, this scenario is based on an implausibly small N_e for the ancestral Khoisan population—because even though these foraging groups live in small bands, the bands are in contact with each other and exchange marriage partners.[43] This ethnographic evidence in favor of a larger effective population size in Khoisan is supported by Bayesian Skyline plots for individual Khoisan populations, which show consistent population sizes of at least 1,000 (data not shown).

Overall, the results of this analysis indicate that it is very unlikely that the highly divergent L0k1b/L0k2 lineages were incorporated into the Bantu-speaking populations via gene flow from a population that was ancestral to a Khoisan population in our sample but subsequently lost from the Khoisan population via drift. Instead, these results support the hypothesis that the ancestors of the Bantu-speaking populations carrying the divergent L0k lineages (who now live mainly in Zambia) experienced gene flow from a pre-Bantu population that is nowadays extinct. Alternatively, it is possible that descendants from this pre-Bantu population do exist but have not yet been included in population genetic studies; however, our extensive sampling of populations from Botswana, Namibia, and West Zambia (which includes representatives of nearly all known Khoisan groups) makes it highly unlikely that this pre-Bantu Khoisan population has not yet been sampled. Our data thus indicate the existence of considerable genetic substructure in southern Africa prior to the Bantu expansion (cf. Barbieri et al.[14]) that is not represented in Khoisan groups today. Unfortunately, individuals from the relevant geographic areas have not yet been included in studies of autosomal DNA variation, making it impossible to assess the overall impact of this substructure on modern genetic diversity in southern Africa. However, from existing Y chromosomal data it appears that the admixture between the pre-Bantu autochthonous groups and the Bantu-speaking immigrants was restricted to the maternal line: the Y chromosome haplogroups found in the Zambian populations included here are not distinct from other sub-Saharan African groups.[27] These findings highlight the importance of

Table 2. Migration Rates from Khoisan into Bantu Required to Retain L0k1b/L0k2 in Bantu with a Probability of at Least 5% over 10,000 Iterations, and Corresponding Estimates of the Frequency of Other "Khoisan" Haplogroups Retained in the Bantu

	Khoisan $N_e = 50$, Bantu $N_e = 5{,}000$			**Khoisan $N_e = 100$, Bantu $N_e = 1{,}000$**			**Khoisan $N_e = 1{,}000$, Bantu $N_e = 10{,}000$**		
Initial frequency of L0k1b/L0k2 in Khoisan	0.05	0.03	0.02	0.03	0.02	0.015	0.004	0.003	0.002
Minimum migration rate	0.001	0.002	0.0025	0.0015	0.003	0.012	0.01	0.023	–[a]
Frequency of other "Khoisan haplogroups"	5%	13%	18%	13%	18%	57%	55%	87%	ND

[a]The migration rate necessary to incorporate L0k1b/L0k2 into the Bantu-speaking group would be higher than 0.1, so the number of migrants would be larger than the N_e of the Khoisan population.

investigating in more detail other relic haplogroups in more regions of sub-Saharan Africa that might testify to a wider genetic variation in the cradle of modern humans.

In conclusion, with this extensive data set of L0d and L0k sequences, we considerably increase our knowledge of the variation in these basal haplogroups. Our results concerning the geographic and genetic structure within haplogroups L0d and L0k reveal interesting patterns. Whereas L0d1 is common to all the Khoisan populations of our data set and in published sources,[6,7,32,33] L0d2 and L0k show a restricted distribution. The presence of divergent L0k haplotypes in populations speaking Bantu languages and their absence from Khoisan populations indicates that it will be possible to learn more about the prehistoric distribution of southern African pre-Bantu peoples by studying Bantu-speaking populations. Several promising areas of southern Africa have yet to be sampled in detail, most notably Zimbabwe, Malawi, and parts of South Africa, Zambia, and Angola; with the retrieval of genetic data from populations located in these areas, we should be able to gain a more complete picture of the genetic variation in southern Africa and better understand the ancient genetic structure.

Supplemental Data

Supplemental Data include three figures and three tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We are grateful to all the individuals who voluntarily participated in this study and to the governments of Angola, Botswana, Namibia, and Zambia for supporting our research. We thank Hongyang Xu for help with the imputation process, Roland Schröder for lab assistance, and Martin Kircher and Mingkun Li for support in processing raw data and read alignments. This work has been carried out within the EUROCORES Programme EuroBABEL of the European Science Foundation and was supported by funds from the Deutsche Forschungsgemeinschaft and the Max Planck Society. J.R. was partially supported by a grant from Fundação para a Ciência e a Tecnologia (FCT; PTDC/BIA-BDE/68999/2006) and M.V. is supported by STAB VIDA, Investigação e Serviços em Ciências Biológicas, Lda., and by the Portuguese Ministry for Science, Technology and Higher Education through PhD grant SFRH/BDE/51828/2012.

Received: October 22, 2012
Revised: November 29, 2012
Accepted: December 19, 2012
Published: January 17, 2013

Web Resources

The URLs for data presented herein are as follows:

BioEdit Software, http://www.mbio.ncsu.edu/BioEdit/bioedit.html
GenBank, http://www.ncbi.nlm.nih.gov/genbank/
HaploGrep, http://haplogrep.uibk.ac.at/
Phylotree, http://www.phylotree.org/

Accession Numbers

The GenBank accession numbers for the 485 sequences reported in this paper are KC345764–KC346248.

References

1. Campbell, M.C., and Tishkoff, S.A. (2010). The evolution of human genetic and phenotypic variation in Africa. Curr. Biol. *20*, R166–R173.
2. Blum, M.G.B., and Jakobsson, M. (2011). Deep divergences of human gene trees and models of human origins. Mol. Biol. Evol. *28*, 889–898.
3. Phillipson, D.W. (2005). African Archaeology (Cambridge: Cambridge University Press).
4. Tattersall, I. (2009). Out of Africa: modern human origins special feature: human origins: out of Africa. Proc. Natl. Acad. Sci. USA *106*, 16018–16021.
5. Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A., and Tishkoff, S.A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. *24*, 757–768.
6. Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al.; Genographic Consortium. (2008). The dawn of human matrilineal diversity. Am. J. Hum. Genet. *82*, 1130–1140.
7. Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B., Ramakrishnan, U., et al. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol. Biol. Evol. *24*, 2180–2195.
8. Chen, Y.S., Olckers, A., Schurr, T.G., Kogelnik, A.M., Huoponen, K., and Wallace, D.C. (2000). mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am. J. Hum. Genet. *66*, 1362–1383.
9. Knight, A., Underhill, P.A., Mortensen, H.M., Zhivotovsky, L.A., Lin, A.A., Henn, B.M., Louis, D., Ruhlen, M., and Mountain, J.L. (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr. Biol. *13*, 464–473.
10. Güldemann, T. (1997). The Kalahari basin as an object of areal typology: A first approach. In Language, Identity and Conceptualization among the Khoisan, M. Schladt, ed. (Köln: Rüdiger Köppe), pp. 137–169.
11. Coelho, M., Sequeira, F., Luiselli, D., Beleza, S., and Rocha, J. (2009). On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evol. Biol. *9*, 80.
12. Schlebusch, C.M., Naidoo, T., and Soodyall, H. (2009). SNaPshot minisequencing to resolve mitochondrial macrohaplogroups found in Africa. Electrophoresis *30*, 3657–3664.
13. Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G., and Behar, D.M. (2010). Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. Am. J. Hum. Genet. *86*, 611–620.
14. Barbieri, C., Butthof, A., Bostoen, K., and Pakendorf, B. (2012). Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur. J. Hum. Genet. Published online August 29, 2012. http://dx.doi.org/10.1038/ejhg.2012.192.

15. Heine, B., and Honken, H. (2010). The Kx'a family: a new Khoisan genealogy. J. Asian Afr. Stud. *79*, 5–36.
16. Güldemann, T. (2005). Studies in Tuu (Southern Khoisan). University of Leipzig Papers on Africa, Languages and Literatures 23 (Leipzig: Institut für Afrikanistik, Universität Leipzig).
17. Güldemann, T., and Elderkin, E.D. (2010). On external genealogical relationships of the Khoe family. In Khoisan Languages and Linguistics: Proceedings of the 1st International Symposium January 4–8, 2003, M. Brenzinger and C. König, eds. (Riezlern/Kleinwalsertal. Quellen zur Khoisan-Forschung. Köln: Rüdiger Köppe,), pp. 15–52.
18. Deacon, H.J., and Deacon, J. (1999). Human Beginnings in South Africa: Uncovering the Secrets of the Stone Age (Walnut Creek, CA: Altamira Press).
19. Güldemann, T., and Stoneking, M. (2008). A historical appraisal of clicks: a linguistic and genetic population perspective. Annu. Rev. Anthropol. *37*, 93–109.
20. Güldemann, T. (2008). A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. South Afr. Humanit. *20*, 93–132.
21. Ehret, C. (2001). Bantu expansions: Re-envisioning a central problem of early African history. Int. J. Afr. Hist. Stud. *34*, 5–41.
22. Pakendorf, B., Bostoen, K., and de Filippo, C. (2011). Molecular perspectives on the Bantu expansion: a synthesis. Language Dynamics and Change *1*, 50–88.
23. Kinahan, J. (2011). From the beginning: the archaeological evidence. In A History of Namibia: From the Beginning to 1990, M. Wallace. (London: Hurst and Company), pp. 15–43.
24. Reid, A., Sadr, K., and Hanson-James, N. (1998). Herding traditions. In Ditswa MMung: The Archaeology of Botswana, P. Lane, A. Reid, and A. Segobye, eds. (Gaborone: Pula Press and The Botswana Society), pp. 81–100.
25. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. *30*, E386–E394.
26. Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., and Kronenberg, F. (2011). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum. Mutat. *32*, 25–32.
27. de Filippo, C., Barbieri, C., Whitten, M., Mpoloka, S.W., Gunnarsdóttir, E.D., Bostoen, K., Nyambe, T., Beyer, K., Schreiber, H., de Knijff, P., et al. (2011). Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol. Biol. Evol. *28*, 1255–1269.
28. Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. Nat. Commun. *3*, 1143.
29. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, pdb.prot5448.
30. Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE *5*, e14004.
31. Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science *325*, 318–321.
32. Schlebusch, C.M., de Jongh, M., and Soodyall, H. (2011). Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. J. Hum. Genet. *56*, 623–630.
33. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. USA *108*, 5154–5162.
34. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.
35. Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.L., Silva, N.M., Kivisild, T., Torroni, A., and Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am. J. Hum. Genet. *90*, 675–684.
36. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA *100*, 171–176.
37. Thomas, D.S.G., and Shaw, P.A. (2002). Late Quaternary environmental change in central southern Africa: new data, synthesis, issues and prospects. Quat. Sci. Rev. *21*, 783–797.
38. Lewis, C.A. (2008). Late Quaternary climatic changes, and associated human responses, during the last 45000 yr in the Eastern and adjoining Western Cape, South Africa. Earth Sci. Rev. *88*, 167–187.
39. Mitchell, P. (2002). The Archaeology of Southern Africa (Cambridge: Cambridge University Press).
40. Schlebusch, C.M. (2010). Genetic variation in Khoisan-speaking populations from southern Africa. PhD thesis, University of the Witwatersrand, Johannesburg.
41. Veeramah, K.R., Connell, B.A., Ansari Pour, N., Powell, A., Plaster, C.A., Zeitlyn, D., Mendell, N.R., Weale, M.E., Bradman, N., and Thomas, M.G. (2010). Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. BMC Evol. Biol. *10*, 92.
42. Batini, C., Lopes, J., Behar, D.M., Calafell, F., Jorde, L.B., van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D. (2011). Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol. Biol. Evol. *28*, 1099–1110.
43. Barnard, A. (1992). Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples (Cambridge, UK: Cambridge University Press).
44. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. *128*, 415–423.

American Journal of Human Genetics, Volume *92*

Supplemental Data

Ancient Substructure in Early mtDNA Lineages of Southern Africa

Chiara Barbieri, Mário Vicente, Jorge Rocha, Sununguko W. Mpoloka, Mark Stoneking, and Brigitte Pakendorf

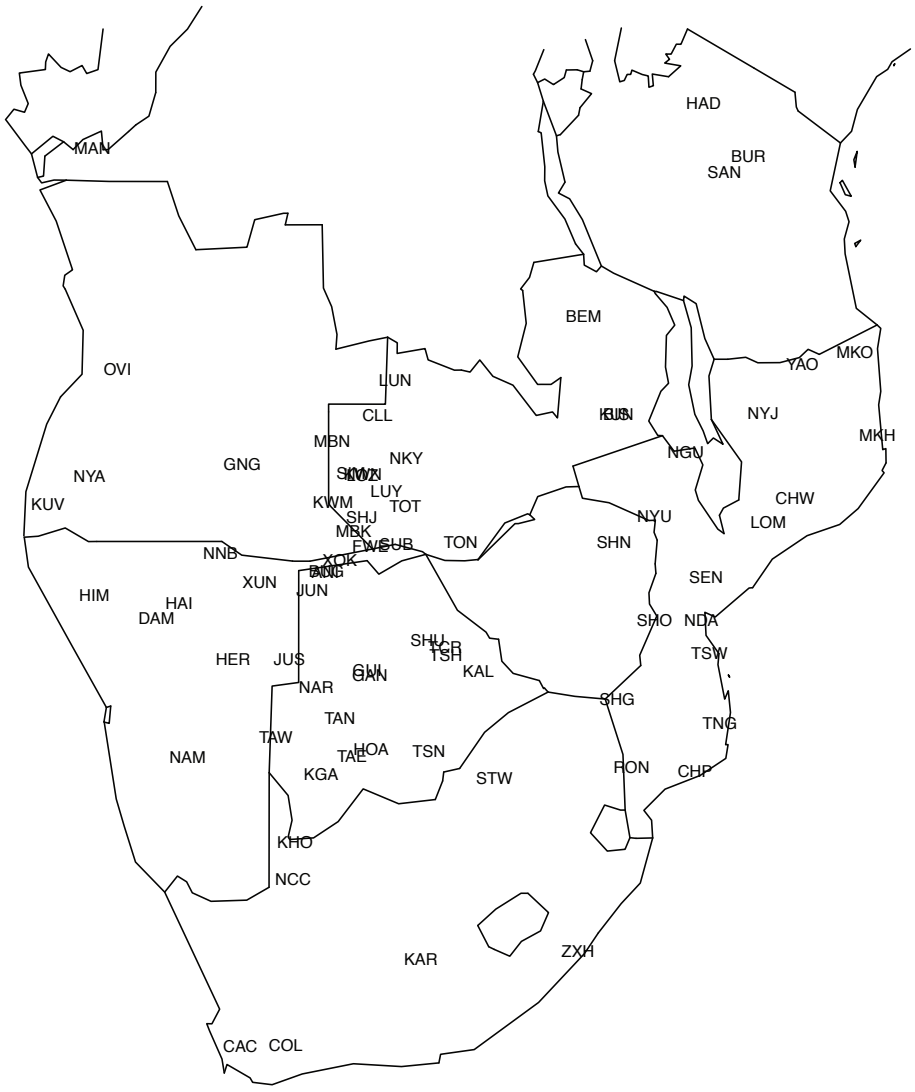


Figure S1: Map of approximate locations of the populations included in the surfer map (Figure 1 in the main text).

Population codes as indicated in Supplementary Table 1.

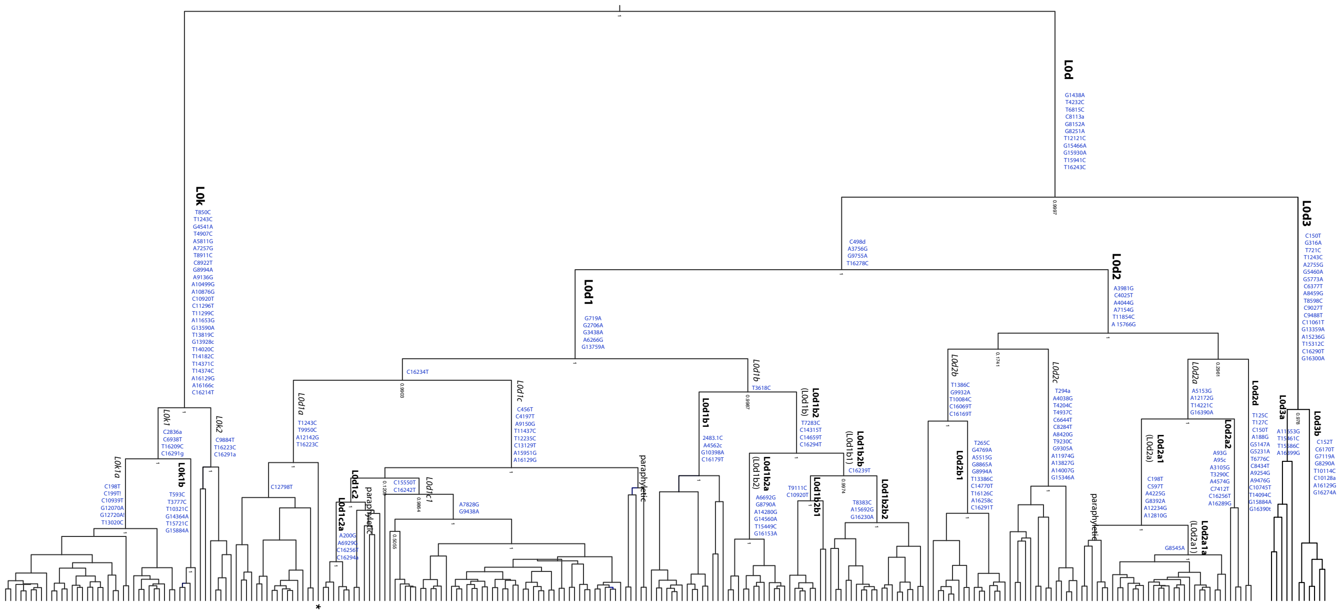


Figure S2: Tree of the 254 unique haplotypes of the dataset.

The tree is based on full sequences, without a time scale, and includes the major branches L0k, L0d1, L0d2 and L0d3. Sub-branches that have not changed are labeled in italic font; new branches defined here are labeled in bold font. When previously defined branches have to be renamed, the older label is indicated in brackets. The posterior probabilities associated with major nodes are shown. Mutations defining branches are shown in blue font: transversions are indicated with lowercase, and back mutations to an ancestral state are indicated with an exclamation mark (!). The individual marked with an asterisk is mentioned in the note about positions 199 and 16266 in Supplementary Table 3.

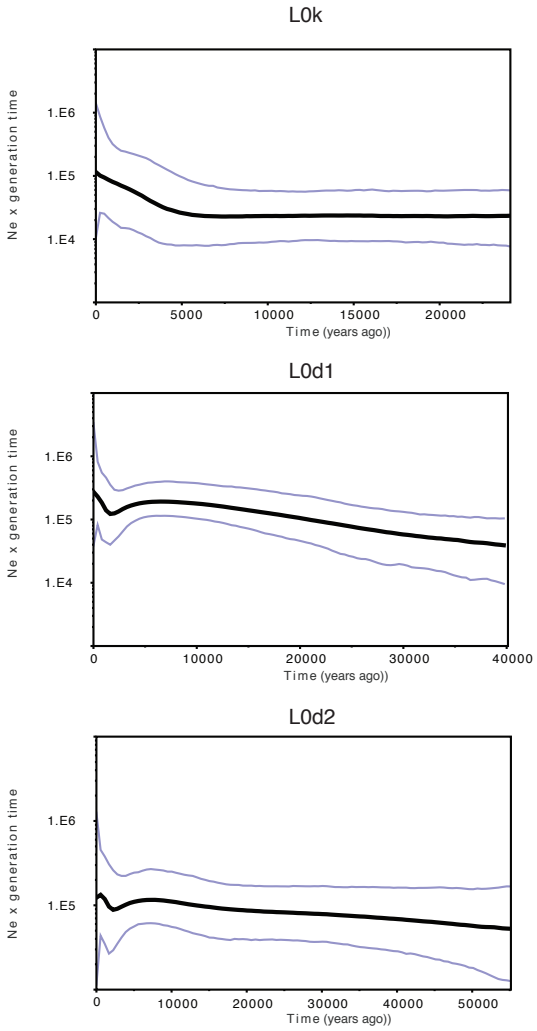


Figure S3: Bayesian Skyline Plots (BSP) of the L0k, L0d1 and L0d2 haplogroups.

The BSPs are based on the mtDNA coding region, estimated with 10 million iterations. The y axis for each plot is the product of the effective population size and the generation time and the x axis shows time using a linear relaxed clock with the substitution rate of 1.26×10-8 per site per year.

Table S1: List of African populations considered in the study, with frequencies of haplogroups L0d and L0k; frequencies of subhaplogroups are given only for populations from the present study.

Population	Code	Country (main)	Language family	lon	lat	N	L0d (%)	L0k (%)	L0d1	L0d2	L0d3	L0k1a	L0k1b	L0k2	Source
Kalanga	KAL	Botswana	Bantu	27.035522	-21.524627	17	**29.4**	**5.9**	29.4	0.0	0.0	0.0	0.0	5.9	PRESENT STUDY
Kgalagadi	KGA	Botswana	Bantu	21.751098	-24.796708	19	**52.6**	**0**	26.3	15.8	10.5	0.0	0.0	0.0	PRESENT STUDY
Tswana	TSN	Botswana	Bantu	25.3656	-24.066528	17	**29.4**	**0**	11.8	17.6	0.0	0.0	0.0	0.0	PRESENT STUDY
Herero	HER	Namibia	Bantu	18.7878333	-21.1344167	30	**16.7**	**0**	13.3	3.3	0.0	0.0	0.0	0.0	PRESENT STUDY
Himba	HIM	Namibia	Bantu	14.1235013	-19.100676	21	**9.5**	**0**	9.5	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
North Namibia Bantu	NNB	Namibia	Bantu	18.365478	-17.748687	10	**0**	**10**	0.0	0.0	0.0	0.0	10.0	0.0	PRESENT STUDY
Bemba	BEM	Zambia	Bantu	30.57312	-10.185187	12	**0**	**8.3**	0.0	0.0	0.0	0.0	0.0	8.3	PRESENT STUDY
Chokwe, Luchazi, Luvale	CLL	Zambia	Bantu	23.638916	-13.346865	33	**0**	**0**	0.0	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Kwangwa	KWN	Zambia	Bantu	23.152888	-15.2451259	36	**2.8**	**0**	2.8	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Lozi	LOZ	Zambia	Bantu	23.133544	-15.284185	118	**5.1**	**0.8**	4.2	0.8	0.0	0.0	0.8	0.0	PRESENT STUDY
Lunda	LUN	Zambia	Bantu	24.240417	-12.243392	9	**0**	**11.1**	0.0	0.0	0.0	0.0	0.0	11.1	PRESENT STUDY
Luyana	LUY	Zambia	Bantu	23.948364	-15.792254	8	**12.5**	**0**	12.5	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Mbunda	MBN	Zambia	Bantu	22.113647	-14.179186	67	**4.5**	**0**	4.5	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Nkoya	NKY	Zambia	Bantu	24.6118007	-14.723394	32	**0**	**0**	0.0	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Simaa	SIM	Zambia	Bantu	22.7506637	-15.2116789	44	**4.5**	**0**	4.5	0.0	0.0	0.0	0.0	0.0	Barbieri et a. 2012
Mbukushu	MBK	Zambia/ Namibia	Bantu	22.8387	-17.0567	20	**10**	**5**	5.0	5.0	0.0	0.0	5.0	0.0	Barbieri et a. 2012
Fwe	FWE	Zambia	Bantu	23.2241122	-17.5299131	33	**6.1**	**18.2**	6.1	0.0	0.0	0.0	6.1	12.1	Barbieri et a. 2012
Kwamashi	KWM	Zambia	Bantu	22.14452	-16.12411	32	**3.1**	**0**	3.1	0.0	0.0	0.0	0.0	0.0	Barbieri et a. 2012
Shanjo	SHJ	Zambia	Bantu	23.1145525	-16.6263951	24	**8.3**	**8.3**	4.2	4.2	0.0	0.0	0.0	8.3	Barbieri et a. 2012
Subiya	SUB	Zambia	Bantu	24.2796841	-17.4674193	17	**5.9**	**0**	0.0	5.9	0.0	0.0	0.0	0.0	Barbieri et a. 2012
Totela	TOT	Zambia	Bantu	24.596557	-16.256867	29	**0**	**0**	0.0	0.0	0.0	0.0	0.0	0.0	Barbieri et a. 2012
Tonga	TON	Zambia	Bantu	26.451416	-17.392579	35	**0**	**0**	0.0	0.0	0.0	0.0	0.0	0.0	Barbieri et a. 2012
Ganguela	GNG	Angola	Bantu	19.068603	-14.902322	20	**5**	**0**	5.0	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY
Kuvale	KUV	Angola	Bantu	12.564697	-16.214675	53	**24.5**	**0**	22.6	1.9	0.0	0.0	0.0	0.0	PRESENT STUDY
Nyaneka-Nkhumbi	NYA	Angola	Bantu	13.992919	-15.30538	59	**8.4**	**0**	8.4	0.0	0.0	0.0	0.0	0.0	PRESENT STUDY

Ovimbundu	OVI	Angola	Bantu	14.915771	-11.888853	60	**3.3**	**1.6**	1.7	1.7	0.0	1.7	0.0	0.0	PRESENT STUDY
‖Anikhoe	ANI	Botswana	Khoe	21.8850954	-18.3734521	18	**44.4**	**22.2**	44.4	0.0	0.0	22.2	0.0	0.0	PRESENT STUDY
‖Xokhwe	XOK	Botswana	Khoe	22.3761	-17.9957	17	**17.6**	**11.8**	17.6	0.0	0.0	11.8	0.0	0.0	PRESENT STUDY
Bugakhoe	BUG	Botswana	Khoe	21.9367	-18.3219	14	**42.9**	**28.6**	42.9	0.0	0.0	28.6	0.0	0.0	PRESENT STUDY
Naro	NAR	Botswana	Khoe	21.5840541	-22.0320817	35	**77.1**	**17.1**	51.4	25.7	0.0	17.1	0.0	0.0	PRESENT STUDY
G‖ana	GAN	Botswana	Khoe	23.3889	-21.6523	15	**93.3**	**6.7**	80.0	13.3	0.0	6.7	0.0	0.0	PRESENT STUDY
G|ui	GUI	Botswana	Khoe	23.2946698	-21.486584	31	**93.5**	**3.2**	51.6	41.9	0.0	3.2	0.0	0.0	PRESENT STUDY
Hai‖om	HAI	Namibia	Khoe	16.9694944	-19.3450768	51	**68.6**	**13.7**	39.2	27.5	2.0	13.7	0.0	0.0	PRESENT STUDY
Nama	NAM	Namibia	Khoe	17.2608889	-24.2660935	29	**79.3**	**3.4**	37.9	34.5	6.9	0.0	3.4	0.0	PRESENT STUDY
Damara	DAM	Namibia	Khoe	16.2257392	-19.8301838	38	**13.2**	**0**	10.5	2.6	0.0	0.0	0.0	0.0	PRESENT STUDY
Shua	SHU	Botswana	Khoe	25.3321307	-20.5502369	42	**35.7**	**2.4**	35.7	0.0	0.0	2.4	0.0	0.0	PRESENT STUDY
TcireTcire	TCR	Botswana	Khoe	25.9166477	-20.7658488	12	**50**	**16.7**	41.7	8.3	0.0	16.7	0.0	0.0	PRESENT STUDY
Tshwa	TSH	Botswana	Khoe	25.9365757	-21.0249347	22	**54.5**	**0**	50.0	4.5	0.0	0.0	0.0	0.0	PRESENT STUDY
ǂHoan	HOA	Botswana	K'xa	23.4351167	-23.9989176	13	**100**	**0**	92.3	7.7	0.0	0.0	0.0	0.0	PRESENT STUDY
!Xuun	XUN	Botswana	K'xa	19.6826306	-18.6907202	27	**55.5**	**33.3**	44.4	11.1	0.0	33.3	0.0	0.0	PRESENT STUDY
Ju|'hoan North	JUN	Botswana	K'xa	21.4524476	-18.9372569	40	**72.5**	**22.5**	50.0	22.5	0.0	22.5	0.0	0.0	PRESENT STUDY
Ju|'hoan South	JUS	Botswana	K'xa	20.6815392	-21.151918	44	**70.5**	**25**	50.0	20.5	0.0	25.0	0.0	0.0	PRESENT STUDY
Taa East	TAE	Botswana	Tuu	22.8206545	-24.2365162	30	**100**	**0**	46.7	53.3	0.0	0.0	0.0	0.0	PRESENT STUDY
Taa North	TAN	Botswana	Tuu	22.4158579	-23.0145647	25	**84**	**16**	68.0	16.0	0.0	16.0	0.0	0.0	PRESENT STUDY
Taa West	TAW	Botswana	Tuu	20.2727412	-23.639938	31	**74.2**	**22.6**	51.6	22.6	0.0	22.6	0.0	0.0	PRESENT STUDY
Shona	SHN	Zimbawe	Bantu	31.593017	-17.413546	59	**1.7**	**1.7**							Castrì et al. 2009
Kunda	KUN	Zambia	Bantu	31.671753	-13.325485	36	**2.8**	**0**							De Filippo et al. 2010
Bisa	BIS	Zambia	Bantu	31.67175	-13.325483	46	**0**	**0**							De Filippo et al. 2010, present study
SA Coloured	COL	South Africa	Indoeuropean	20.562744	-33.449777	563	**60**	**0**							Quintana-Murci et al. 2010
Chopi	CHP	Mozambique	Bantu	34.317627	-24.726875	27	**0**	**0**							Salas et al. 2002
Chwabo	CHW	Mozambique	Bantu	37.679443	-16.003576	20	**0**	**0**							Salas et al. 2002
Lomwe	LOM	Mozambique	Bantu	36.778564	-16.762468	20	**0**	**0**							Salas et al. 2002
Makhwa	MKH	Mozambique	Bantu	40.447998	-13.987376	20	**0**	**0**							Salas et al. 2002
Makonde	MKO	Mozambique	Bantu	39.700927	-11.350797	19	**5.3**	**0**							Salas et al. 2002
Ndau	NDA	Mozambique	Bantu	34.537353	-19.890723	19	**30**	**0**							Salas et al. 2002

Nguni	NGU	Mozambique	Bantu	34.010009	-14.51978	11	**0**	**0**							Salas et al. 2002
Nyanja	NYJ	Mozambique	Bantu	36.602783	-13.304103	20	**0**	**0**							Salas et al. 2002
Nyungwe	NYU	Mozambique	Bantu	32.955322	-16.594081	20	**0**	**0**							Salas et al. 2002
Ronga	RON	Mozambique	Bantu	32.186279	-24.58709	21	**19**	**0**							Salas et al. 2002
Sena	SEN	Mozambique	Bantu	34.691162	-18.521283	21	**0**	**0**							Salas et al. 2002
Shangaan	SHG	Mozambique	Bantu	31.70288	-22.411029	22	**4.5**	**0**							Salas et al. 2002
Shona	SHO	Mozambique	Bantu	32.955322	-19.911384	18	**0**	**0**							Salas et al. 2002
Tonga	TNG	Mozambique	Bantu	35.152587	-23.180764	20	**5**	**0**							Salas et al. 2002
Tswa	TSW	Mozambique	Bantu	34.801025	-20.96144	19	**15.8**	**0**							Salas et al. 2002
Yao	YAO	Mozambique	Bantu	37.965087	-11.716788	10	**0**	**0**							Salas et al. 2002
Karretjie Mense	KAR	South Africa	Indoeuropean [a]	25.101013	-30.712638	30	**100**	**0**							Schlebusch et al. 2011
Cape Colured	CAC	South Africa	Indoeuropean	19.037475	-33.495598	20	**45**	**0**							Schlebusch 2010
Khomani	KHO	South Africa	Tuu	20.872192	-26.971038	57	**98.2**	**0**							Schlebusch 2010
Manyanga	MAN	DRC	Bantu	14.058837	-4.82826	14	**0**	**0**							Schlebusch2010
Northern Cape Coloured	NCC	South Africa	Indoeuropean	20.804443	-28.149503	40	**92.5**	**0**							Schlebusch 2010
Sotho Tswana	STW	South Africa	Bantu	27.572021	-24.926295	22	**22.7**	**0**							Schlebusch 2010
Zulu Xhosa	ZXH	South Africa	Bantu	30.384521	-30.448674	36	**44.4**	**2.8**							Schlebusch 2010
Burunge	BUR	Tanzania	Cushitic	36.119384	-5.090944	38	**3**	**0**							Tishkoff et al. 2007
Hadza	HAD	Tanzania	Khoisan (isolated)	34.603271	-3.403758	79	**0**	**0**							Tishkoff et al. 2007
Sandawe	SAN	Tanzania	Khoisan (isolated)	35.306396	-5.594118	82	**5**	**0**							Tishkoff et al. 2007

[a] this population used to speak a Tuu language but has shifted to Afrikaans.

Table S2: List of positions (numbered in accordance with the RSRS/rCRS) with missing data that were excluded from the analysis. Polymorphic sites are underlined.

316
1243
3106
3492
3516
3981
4232
5515
5936
6716
6938
7412
8563
10550
10589
11854
13020
13198
13386
14770
15530
15930
15941
16069
16093
16169
16212
16215
16230
16242
16243
16474

Table S3: Notes on some of the haplogroup-defining mutations.

Mutation	Remarks
C152T	This mutation defines L0d3b, but is also present in an individual belonging to L0d3a as well as being found sporadically in other branches of L0d and L0k.
A188G	In Supplementary Figure 2, this is shown only for L0d2d; however, this mutation also occurs in nearly all the individuals belonging to L0d1b1, with only 2 exceptions.
C198T	In Supplementary Figure 2, this is shown for L0d2a1 and L0k1a; however, this mutation also defines a minor subbranch of L0d1c1 (rather than defining L0d1c1 as a whole, as previously thought).
199	The evolutionary pathway involving L0k cannot be resolved, since L0k2 and L0k1b carry a C at this position, while L0k1a carries a T, which is the state reconstructed for the RSRS. In Supplementary Figure 2 we show the C199T back mutation as defining L0k1a; however, with our dataset it is equally likely that two independent T-C transitions occurred on the branches leading to L0k2 and L0k1b, with L0k1a retaining the ancestral T. In addition, L0d1a carries a C at this position with the exception of three lineages not forming a clade. One of these is a deeply divergent lineage represented by only one individual from Botswana (indicated by an asterisk in Supplementary Figure 2); thus, one could postulate either three back mutations from the mutation defining L0d1a as a whole, or consider T199C a defining mutation only for the subclade L0d1a1, with two back mutations having occurred subsequently. Since C16266a is also missing in the divergent lineage (see below), one should perhaps consider both T199C and C16266a as mutations defining the subclade L0d1a1, with subsequent back mutations (C199T) or novel mutations (A16266G) in some individuals.
294	In Supplementary Figure 2, we show the T-A transversion defining L0d2c; in addition, a T-C transition defines a subbranch of L0d1c1.
A7828G	Rather than defining branch L0d1c1 as a whole, as previously suggested, this is missing from one individual and thus defines only a subbranch, as shown in Supplementary Figure 2.
C8922T	This is found in L0k2, L0k1b, and several branches of L0k1a, but is missing from one subbranch of L0k1a. The most plausible reconstruction is that the transition occurred on the branch leading to L0k, as previously assumed,
G8994A	In Supplementary Figure 2, this is shown only for L0d2b1 and L0k; however, this mutation also defines a small subbranch of the paraphyletic branch of L0d1c.
A9136G	This mutation defining L0k mutates back to A in a subbranch of L0k2.
A9347G	This mutation is at the root of haplogroup L0, but almost all of the L0k2 individuals present a back mutation at this site, with the exception of the sample from Yemen.

G9438A	Rather than defining branch L0d1c1 as a whole, as previously suggested, this is missing from one individual and thus defines only a subbranch, with a further back mutation to A9438G found in one sequence.
A11653G	This mutation defines L0k as well as L0d3a.
C15550T	Together with C16242T, this is the only mutation defining branch L0d1c1; A7828G and G9438A are missing from one divergent lineage and thus define only a subset of L0d1c1 (with a further back mutation to A9438G found in one sequence), while C198T defines an even smaller branch within L0d1c1.
T15586C	This mutation defining L0d3a mutates back to T in one individual of the same subbranch.
A16129G	This mutation defines L0d1c, L0k and L0d3b, as well as a subbranch of L0d1b2a. Given the hypervariability of this position, it is not surprising that several back mutations to A occur in the tree – the most notable being a back mutation in the individual from Yemen whose sequence up to now was the only lineage known for L0k2. Therefore, A16129G was previously considered a mutation defining only L0k1; with our extended dataset we show that it defines all of L0k.
T16209C	This mutation, which defines L0k1, also appears in a subbranch of L0d1a.
C16242T	Together with C15550T, this is the only mutation defining branch L0d1c1; A7828G and G9438A are missing from one divergent lineage and thus define only a subset of L0d1c1 (with a further back mutation to A9438G found in one sequence), while C198T defines an even smaller branch within L0d1c1.
16266	Like the T-C transition at 199, C16266a is not found in all the sequences belonging to L0d1a; rather, four sequences carry a G at this position. Since one of these is the divergent lineage represented by an asterisk in Supplementary Figure 2 (as mentioned for position 199 above), one should perhaps consider both T199C and C16266a as mutations defining the subclade L0d1a1, with subsequent back mutations (C199T) or novel mutations (A16266G) in some individuals.
16291	While a C-T transition defines branch L0d2b1, it also defines a subbranch of the paraphyletic sister clade of L0d1c2. Furthermore, L0k1 is defined by a G at this position, with a subsequent G to A transition on a subbranch of L0k1a; L0k2 carries an A at this position. While Phylotree (http://www.phylotree.org/tree/subtree_L.htm, Build 15) reconstructs a C-G transversion for L0k as a whole and a G-A transition for L0k2, from the data available to us it appears impossible to decide whether a C-G or C-A transversion took place on the branch leading to L0k. Therefore, in Supplementary Figure 2 the mutations defining L0k1 and L0k2 are both listed as transversions, even though the actual evolutionary path would have involved just one transversion (on the branch leading to L0k) and one transition (on either L0k1 or L0k2).
16294	While a C-T transition defines branch L0d1b2, and a C-A transversion defines branch L0d1c2a, the paraphyletic sister branch of L0d1c2a is defined by a G at this position, with the exception of one individual who carries an A.
A16300G	This mutation, which defines L0d3, mutates back to A in two individuals of branch L0d3b

The mutations are numbered in accordance with the RSRS/rCRS sequence.

Supplemental References:

Barbieri, C., Butthof, A., Bostoen, K., and Pakendorf, B. (2012). Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet. doi: 10.1038/ejhg.2012.192. Aug 29. [Epub ahead of print]

Castrì, L., Tofanelli, S., Garagnani, P., Bini, C., Fosella, X., Pelotti, S., Paoli, G., Pettener, D., and 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. Am J Phys Anthropol 140, 302-311.

de Filippo, C., Heyn, P., Barham, L., Stoneking, M., and Pakendorf, B. (2010). Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. Am J Phys Anthropol 141, 382-394.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G., and Behar, D.M. (2010). Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. Am J Hum Genet 86, 611-620.

Schlebusch, C.M., de Jongh, M., and Soodyall, H. (2011). Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. J Hum Genet 56, 623-630.

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

Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B., Ramakrishnan, U., et al. (2007). History of click-speaking Populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24, 2180-2195

Chapter 9

PAPER IV: Unravelling the Complex Maternal History of Southern African Khoisan Populations

This chapter includes the paper "**Unravelling the Complex Maternal History of Southern African Khoisan Populations**" written by Barbieri, Chiara, Tom Güldemann, Christfried Naumann, Linda Gerlach, Falko Berthold, Hirosi Nakagawa, Sununguko W. Mpoloka, Mark Stoneking, and Brigitte Pakendorf, which is currently under review at *Molecular Biology and Evolution*.

Molecular Biology and Evolution – Article: Discoveries

Unraveling the complex maternal history of southern African Khoisan populations

Chiara Barbieri[1,7], Tom Güldemann[2,3], Christfried Naumann[2,3], Linda Gerlach[1,8], Falko Berthold[1,8], Hirosi Nakagawa[4], Sununguko W. Mpoloka[5], Mark Stoneking[6], Brigitte Pakendorf[1,9]

[1]Max Planck Research Group on Comparative Population Linguistics, MPI for Evolutionary Anthropology, Leipzig

[2]Seminar für Afrikawissenschaften, Humboldt University, Berlin

[3]Department of Linguistics, MPI for Evolutionary Anthropology, Leipzig

[4]Institute of Global Studies, Tokyo University of Foreign Studies, Tokyo

[5]Department of Biological Sciences, University of Botswana, Gaborone

[6]Department of Evolutionary Genetics, MPI for Evolutionary Anthropology, Leipzig

[7]Current affiliation: Department of Evolutionary Genetics, MPI for Evolutionary Anthropology, Leipzig

[8]Current affiliation: Department of Linguistics, MPI for Evolutionary Anthropology, Leipzig and Seminar für Afrikawissenschaften, Humboldt University, Berlin

[9]Current affiliation: Laboratoire Dynamique du Langage, UMR5596, CNRS and Université Lyon Lumière 2, Lyon

Corresponding authors:

Chiara Barbieri: barbieri.chiara@gmail.com

Brigitte Pakendorf: Brigitte.Pakendorf@ish-lyon.cnrs.fr

ABSTRACT

The Khoisan populations of southern Africa are known to harbor some of the deepest-rooting lineages of human mitochondrial DNA; however, their relationships are as yet poorly understood. Here, we report the results of analyses of complete mtDNA genome sequences from nearly 700 individuals representing 26 populations of southern Africa who speak diverse Khoisan and Bantu languages. Our data reveal a multilayered history of the indigenous populations of southern Africa, who are likely to be the result of admixture of different genetic substrates, such as resident forager populations and pre-Bantu pastoralists from East Africa. We find high levels of genetic differentiation of the Khoisan populations, which can be explained by the effect of drift together with a partial uxorilocal/multilocal residence pattern. Furthermore, there is evidence of extensive contact, not only between geographically proximate groups, but also across wider areas. The results of this contact are especially evident in the Khoisan populations of the central Kalahari, where they may have played a role in the diffusion of common cultural and linguistic features.

INTRODUCTION

African populations are increasingly the focus of genetic studies, in particular those characterized by the simultaneous presence of an ancestral way of subsistence (predominantly foraging) together with deep-rooting genetic lineages, like the Pygmies of central Africa and the Khoisan of southern Africa (Veeramah et al. 2011, Tishkoff et al. 2009, Lachance et al. 2012, Henn et al. 2011, Patin et al. 2009, Pickrell et al. 2012, Schlebusch et al. 2012, Schuster et al. 2010). With the term "Khoisan" we refer to the hunter-gatherer and pastoralist populations of southern Africa that speak indigenous non-Bantu languages characterized by heavy use of click consonants, without any assumption about their genetic or linguistic unity (cf. Barnard 1992). The term was first proposed by the anthropologist Leonard Schultze (Schultze 1928) to subsume the pastoralist Khoekhoe and the neighboring foraging San in

South Africa under a single label; despite its early biological connotation it became widespread by association with linguistics when Greenberg adopted it to name one of the linguistic phyla of Africa (Greenberg 1963). However, specialists of Khoisan languages agree that the languages joined by Greenberg in his Khoisan phylum are unlikely to be all genealogically related (Westphal 1971, Sands 1998, Güldemann 2008a).

There is archeological evidence of continuous presence of foragers in the Kalahari region since the Late Stone Age ~30,000 years ago (Denbow 1984, Deacon and Deacon 1999). Much later in time, signals of pastoralist and Iron Age agriculturalist cultures begin to appear in the archeological record. Pottery and remains of domesticated animals appear almost simultaneously ~2000 years ago in the coastal regions of what is now South Africa and Namibia, and in northern Botswana. One hypothesis suggests that this pastoralist culture originated in East Africa, where the animal species were domesticated (Phillipson 2005, Deacon and Deacon 1999), and was brought to southern Africa by an immigration of East African herders, spreading rapidly over the entire territory (Deacon and Deacon 1999, Smith 1992, Mitchell 2002, Pleurdeau et al. 2012). In contrast, some archeologists prefer an explanation of cultural diffusion, according to which "hunters with sheep" would have autonomously embarked on the transition to a new way of subsistence after coming into contact with populations of herders from the north (Sadr 1998, Kinahan 1991). However, such a rapid shift in lifestyle and cultural paradigm is hard to reconcile with the ethnographic evidence (Smith 1990, Barnard 2008). Furthermore, the Nama and now extinct Khoekhoe in South Africa, who are described in historical records, represent an uncommon case of very specialized herding cultures that are assumed to have emerged only after a long period of interaction with livestock (Fauvelle-Aymar 2008). The pastoralist tradition predates the arrival of the agriculturalist Bantu speakers, whose culture appears in the archeological record of southern Africa not earlier than 2000-1200 years ago (Phillipson 2005, Kinahan 2011, Reid et al. 1998).

Khoisan populations speak languages that belong to three distinct families (see Figure 1): Kx'a (Heine & Honken 2010), Tuu (Güldemann 2005), and Khoe-Kwadi (Güldemann 2004; Güldemann & Elderkin 2010); in addition, two isolates of Tanzania, Sandawe and Hadza, were included in the "Khoisan" phylum by Greenberg (1963). Kx'a and Tuu share some linguistic features, and their distribution is mostly centered over the Kalahari and its immediate surroundings. Speakers of dialects belonging to the Ju branch of Kx'a are settled mainly somewhat to the northwest of the Kalahari, in northeast Botswana, northern Namibia,

and southern Angola. The Tuu language family was formerly more widely distributed than today, covering most of South Africa as well as parts of Botswana and Namibia; however, in South Africa most Khoisan populations have assimilated culturally and linguistically to neighboring populations. Khoe-Kwadi languages are distributed over a large geographic area, including the Kalahari, western Namibia, the Okavango river delta, and the salt pans to the east of the Kalahari; the now extinct language Kwadi was spoken in southern Angola. While all speakers of Kx'a and Tuu languages are (or were) foragers, Khoe-Kwadi speakers are diverse in terms of way of subsistence: the majority are (or were) foragers (with a focus on fishing in the Okavango river), but the now extinct Kwadi of Angola and the Nama of Namibia were traditionally pastoralists, and the Damara had a mixed pattern of subsistence involving hunting and gathering as well as herding of small livestock (Barnard 1992). Lastly, there are phenotypic differences: while the majority of Khoisan populations have on average light skin pigmentation and relatively short stature (a phenotype we here refer to as the "Khoisan phenotype"), the Damara from Namibia as well as populations of the eastern Kalahari and of the Okavango region are characterized by on average taller stature and darker skin pigmentation; the latter two groups were therefore known as "Black Bushmen" (e.g. Weiner 1964, Jenkins 1986, Gusinde 1966).

From a genetic perspective, Khoisan populations are known to harbor the deepest-rooting clades of uniparental lineages (Behar et al. 2008, Naidoo et al. 2010, Schlebusch et al. 2009), but until recently not much was known about the relationships between individual populations and the distribution of genetic variation in these populations. Two novel studies of autosomal DNA diversity in extended datasets of Khoisan populations from southern Africa demonstrate an ancient split that dates within the past ~30,000 years, dividing Khoisan populations of the northwest Kalahari Basin from those settled to the southeast or south (Pickrell et al. 2012, Schlebusch et al. 2012). Furthermore, both studies detect genetic links with East Africa in the Nama and other Khoe-Kwadi speakers. This is in good accordance with the hypothesis that the Khoe-Kwadi languages were brought to southern Africa by the pre-Bantu immigration of pastoralists detectable in the archeological record (Güldemann 2008b). A similar link of the Khwe from southern Angola/the Caprivi Strip, who speak a Khoe-Kwadi language, with East African pastoralists was detected in the shared presence at high frequency of the Y-chromosomal haplogroup E-M293 (Henn et al. 2008) that is rare elsewhere in Africa (de Filippo et al. 2011). Finally, there is evidence of varying degrees of non-Khoisan ancestry in all Khoisan populations, which could reflect contact with pre-Bantu pastoralists and/or Bantu-

speaking populations that took place at different periods of time in different areas (Pickrell et al. 2012).

The mtDNA variation of most Khoisan populations is characterized by high frequencies of the deepest clades of the mtDNA phylogeny, namely haplogroups L0d and L0k (Behar et al. 2008, Schlebusch et al. 2009, Barbieri et al. 2013). A minor presence of these haplogroups in neighboring Bantu-speaking populations can be explained by gene flow after the ancestors of these populations reached these southernmost areas of their migration route; the proportion of L0d and L0k in Bantu-speaking populations is higher than that of the characteristic Y-chromosomal haplogroups A-M91 and B-M112, in line with sex-biased gene flow after contact (Coelho et al. 2009; Schlebusch et al. 2011; Quintana-Murci et al. 2010; Barbieri et al. 2012a). The source of haplogroups other than mitochondrial L0d/L0k and Y-chromosomal A-M91/B-M112 in Khoisan foragers has often been identified with Bantu agriculturalists (Schlebusch 2010); however, the possibility of gene flow from pastoralists or other pre-Bantu populations should not be dismissed out of hand.

This study is one of the first to investigate the history of Khoisan populations using complete mtDNA genome data from a large set of populations. The extensive diversity of haplogroups L0d and L0k and the resulting changes to the mtDNA phylogeny are discussed in Barbieri et al. (2013), while here we focus on the maternal prehistory of the Khoisan peoples. We analyze a total of nearly 700 complete mtDNA genome sequences from 19 Khoisan populations covering the three linguistic families Kx'a, Tuu and Khoe-Kwadi and including both hunter-gatherers and pastoralists, as well as from seven neighboring Bantu-speaking populations. Our dataset covers most of the extant variability in Khoisan populations, but lacks samples from South African populations whose heritage languages belonged to the Tuu family and the Khoekhoe group of Khoe-Kwadi, as well as the extinct Kwadi of Angola; for this reason we refer to the "Khoe family" and "Khoe speakers" instead of the "Khoe-Kwadi family" and "Khoe-Kwadi speakers" in the remainder of this article. With these data we aim at investigating the relationships among Khoisan populations as well as evidence for gene flow among them. In particular, we focus on the following research questions: 1) How is the maternal genetic component structured in Khoisan, and does it mirror the genetic structure emerging from the genome-wide data? 2) How much contact was there between different Khoisan populations, and to what extent does contact correlate with geographic proximity? 3) Can we detect traces of the hypothesized East African ancestry of populations speaking Khoe languages?

MATERIALS AND METHODS

The dataset

Samples were collected in Botswana and Namibia between 2009 and 2011. The collection was approved by the ethical review board of the University of Leipzig and authorized by the governments of Botswana and of Namibia (Research permit CYSC 1/17/2 IV (8) from the Ministry of Youth Sport and Culture of Botswana, and 17/3/3 from the Ministry of Health and Social Services of Namibia). Each individual gave written consent after the purpose of the study was explained with the help of local translators. Details on the sample collection and DNA extraction from saliva have been reported in the Supplementary Material of Pickrell et al. 2012. While in that study a reduced set of 187 individuals was chosen for genome-wide SNP typing from a total of 22 African populations, in this study we consider almost all the unrelated individuals from the same sample collection from Botswana and Namibia. Relatives were excluded from the analysis as far as they could be ascertained from the information provided, as were individuals with unclear ethnolinguistic family background, resulting in a dataset of 665 individuals belonging to 19 Khoisan and five Bantu-speaking populations from Botswana and Namibia. This dataset was augmented with 22 Tonga and 12 Mbukushu sequences from Zambia (Barbieri et al. 2012a); these Mbukushu sequences were merged with data from Mbukushu samples obtained in Namibia, after checking for genetic homogeneity. The Tonga were chosen for comparison because they represent a relatively unadmixed Bantu-speaking population (Barbieri et al. 2012a) with clear ethnolinguistic self-affiliation from an area adjacent to the region that is the focus of this study.

Nineteen sequences were not included in analyses based on population comparisons because they belong to populations with sample sizes below 12 individuals; these are speakers of Khoe languages from Botswana (8 individuals) and of Bantu languages from Namibia (11 individuals). These sequences were included only in comparisons of haplotypes (i.e. network analyses). We assigned the remaining 680 individuals to 26 populations on the basis of ethnolinguistic self-affiliation during sample collection. The Khoisan populations and their linguistic affiliations are: Kx'a family: Ju|'hoan North, Ju|'hoan South, and !Xuun (Ju dialect cluster) as well as ǂHoan; Khoe branch of the Khoe-Kwadi family: G|ui, G‖ana, Naro, Tshwa, Tcire Tcire, Shua, ‖Ani, Buga, ‖Xo, Nama, Hai‖om, and Damara; and Tuu family: Taa East, Taa North, and Taa West, which are populations defined for this study on geographic and linguistic grounds, but who actually speak several dialects of a single language (cf. Figure 1

for an overview of the linguistic affiliations of the populations included). The Bantu-speaking populations are the Herero and Himba from Namibia, the Tonga and Mbukushu from Zambia (with some Mbukushu from Namibia, as mentioned above), and the Kalanga, Kgalagadi, and Tswana from Botswana.

Populations were grouped together according to their geographic distribution, and in some cases taking into consideration their linguistic affiliation and way of subsistence, into eight clusters (see Table 1). This was done to simplify the interpretation of sequence sharing and networks, and for analyses performed in BEAST, where larger sample sizes improve the performance of the methods. These clusters are: NORTHWEST (comprising Kx'a-speaking Ju|'hoan North, Ju|'hoan South, and !Xuun, together with Khoe-speaking Hai‖om), SOUTH-CENTRAL (comprising Tuu-speaking Taa East, Taa North, and Taa West, together with Kx'a-speaking ǂHoan), CENTRAL (comprising G|ui, G‖ana, and Naro), OKAVANGO (comprising ‖Ani, Buga, and ‖Xo, all from the Okavango River), EAST (comprising Shua, Tshwa and Tcire Tcire), NAMA (comprising only the pastoralist Nama), NORTHWEST-NAMIBIA (NW-NAMIBIA; comprising the pastoralist Bantu-speaking Himba and Herero together with the Khoe-speaking Damara), and BANTU (comprising the other Bantu-speaking populations: Kalanga, Kgalagadi, Tswana, Tonga, Mbukushu, and other Bantu-speakers from Namibia).

Sequence and data analysis

Genomic libraries were made from sheared DNA, following protocols described in Meyer and Kircher (2010) and Maricic et al. (2010); see also Supplementary Text in Barbieri et al. (2012b). Fragments were tagged with both single-indexing and double-indexing methods. Libraries were enriched for mtDNA with in-solution capture on streptavidin treated baits. The libraries were sequenced on the Illumina GAIIx platform, using either single or paired end runs of 76 bp length, resulting in an average coverage of ~400x. Sequences were manually checked with Bioedit (www.mbio.ncsu.edu/BioEdit/bioedit.html) and read alignments were screened with ma (Briggs et al. 2009) to exclude alignment errors and confirm INDELS. The sequences belonging to haplogroups L0d and L0k were already submitted to Genbank (http://www.ncbi.nlm.nih.gov/genbank/; Barbieri et al.2013) and given accession numbers KC345764-KC346248; the remaining 218 sequences were given accession numbers XXXX. The two poly-C regions (np 303-315, 16183-16194), which are prone to sequencing errors, were trimmed from the final alignment used in the analysis.

In the final alignment of 699 sequences there are 1290 polymorphic positions, of which 7% were missing in one or more sequences. To minimize the impact of missing data, we applied imputation using stringent criteria, replacing missing sites with the nucleotide that was present in more than two other identical haplotypes of the dataset. After imputation, the maximum number of missing sites per sample was three, and the amount of polymorphic positions with missing sites was 1.5%. Positions with missing sites were excluded from the analysis. Haplogroup assignment was performed with the online tool Haplogrep (Kloss-Brandstätter et al. 2012).

Values of nucleotide diversity and variance were calculated in R with the package Pegas (Paradis 2010). CA analysis was performed with the package ca (Nenadic and Greenacre 2007). Non-metric Multi-Dimensional Scaling (MDS) analyses were performed with the function "isoMDS" from the package MASS (Venables and Ripley 2002). AMOVA, values of sequence diversity and Φst matrices of distances were computed in Arlequin ver. 3.11. A Mantel test was performed between genetic (Φst) and geographic distances with the R package vegan (Oksanen et al. 2012); geographic distances between populations were averaged over GPS data from the individual sampling locations with the function rdist.earth of the package fields (Furrer et al. 2012). A population tree was generated from a Φst matrix of distances with the function "nj" of the package ape (Paradis et al. 2004).

Median-joining networks (Bandelt et al. 1999) were computed with Network 4.11 (www.fluxus-engineering.com), and branches showing starlike signals of expansions were dated using the rho statistic (Forster et al. 1996) implemented in Network, with the mutation rate for the complete mtDNA genome of one mutation every 3624 years (Soares et al. 2009). In the L0d1 network, branches are labeled with subhaplogroup names, according to the nomenclature proposed in Barbieri et al. (2013).

BEAST (v1.7.2 Drummond et al. 2012) was used to construct Bayesian Skyline Plots, based on the mtDNA coding region only and using the mutation rate of Soares et al. (2009). A Generalized Time Reversible model was applied, and multiple runs were performed for each dataset, using 30 million chains.

Simulations were performed in Serial Simcoal (Anderson et al. 2005) to estimate the probability of retaining identical sequence types after a given number of generations following a population split, starting from effective population sizes of 100, 1000, 5,000 and 10,000 individuals. We based our simulations on the two groups emerging from the

autosomal data – NW Kalahari and SE Kalahari – which are estimated to have split within the last 30,000 years (Pickrell et al. 2012). The populations included in the two groups were chosen according to Figure S18 of Pickrell et al. (2012): the Northwest Kalahari group (NW Kalahari) included the Ju|'hoan South, Ju|'hoan North, !Xuun, and Hai‖om (and thus corresponds to our NORTHWEST cluster), and the Southeast Kalahari group (SE Kalahari) included the Taa North, Taa East, Taa West, ǂHoan, G‖ana, Shua and Tshwa. The resulting groups had sample sizes of 162 for the NW Kalahari and 209 for the SE Kalahari, with 7 haplotypes shared between the groups. We proceeded as follows: the initial population was split in two populations, N_e was kept constant, and no migration was considered. The time after the split was calculated applying a generation time of 25 years (Fenner 2005). The possibility of generating new haplotypes was taken into account, calculated from the mutation rate for full mtDNA genomes from Soares et al. (2009). For each effective population size and split time we ran 1000 iterations, and calculated both the probability of retaining identical haplotypes and the average number of haplotypes retained, sampling 162 and 209 individuals.

RESULTS AND DISCUSSION

Khoisan mtDNA variation, population size and demography

The haplogroups L0d and L0k are the most common haplogroups in our dataset: L0d1 is present at 38%, L0d2 at 16% and L0k at 11%. As discussed in detail in Barbieri et al. (2013), these haplogroups are present in higher proportions in most Khoisan populations than in populations speaking Bantu languages (Supplementary Table); the highest percentages are found in Khoisan populations residing in the central Kalahari, like the ǂHoan, G‖ana, Taa North and Taa East, where L0d and L0k comprise 100% of the haplogroup composition. Apart from L0d and L0k, the other haplogroups found in the dataset have a non-uniform distribution, and they mainly characterize and distinguish Bantu-speaking populations from each other; a few of these haplogroups are also present in certain Khoisan populations. L2a, which is a widely distributed African haplogroup (Salas et al. 2002), is common in the EAST cluster (Shua, Tcire Tcire, and Tshwa; 9-17%) and present in lower frequency in the ‖Ani and ‖Xo (~6%); it is otherwise present in the Bantu-speaking Tonga, Mbukushu, Kalanga and

Tswana (18-35%). An uncommon haplogroup in our dataset is L2b, which is present exclusively in ǁXo (12%) and Buga (7%), although it is widespread in the rest of the continent (Behar et al. 2008). Another rare haplogroup is L1b, present at low frequency only in the OKAVANGO cluster (6-7%) and Haiǁom (4%), as well as in the Bantu-speaking Tonga and Mbukushu (5-9%); in Africa, it is present at highest frequency in West Africa (Rosa et al. 2004). L3d is frequent in Bantu-speaking populations and probably accompanied the Bantu expansion (Salas et al. 2002, Beleza et al. 2005); it is present at high frequencies (38 to 63%) in the NW NAMIBIA cluster, mainly represented by its subhaplogroup L3d3a, and in lower frequencies (3-17%) in various populations speaking languages of the Khoe family as well as in !Xuun (Supplementary Table). L3e, which probably originated in central Africa and spread to the rest of the continent with the Bantu expansion (Bandelt 2001), is found in the Bantu-speaking Mbukushu, Tonga, Himba, Kalanga, and Kgalagadi at frequencies of 14-19%, and at varying frequency in most Khoe-speaking populations (as well as in Ju|'hoan North), with high frequencies in the Shua (17%) and the ǁXo (35%). In contrast, L3f is characteristic of the Himba and Herero (29-33%) and also present at 12% in Kalanga and at 9% in Shua; other studies suggest that this haplogroup is likely to have originated in eastern Africa, where it is more frequent and more diverse (Salas et al. 2002, Soares et al. 2012). L4 is present at very low frequencies in NORTHWEST populations and in Naro (2-5%): this is mostly a typical eastern African haplogroup, with highest frequencies in Hadza and Sandawe (Tishkoff et al. 2007). Lastly, L5 is exclusive of the Tshwa and Shua (EAST), with 18 and 5% frequency, respectively; elsewhere, it is found in East Africa (Salas et al. 2002), including Sandawe (at 5%, Tishkoff et al. 2007), and at higher frequency in Pygmies, where some sequences fall into a clade with some East African individuals (Batini et al. 2011).

In the MDS analysis based on sequence data (Figure 2), the first dimension separates populations of foragers resident in the Kalahari from Bantu-speaking populations from Zambia and Namibia plus the Damara. The OKAVANGO and EAST clusters as well as the Bantu-speaking populations from Botswana are located in an intermediate position. Notwithstanding their geographic location in southern-central Namibia and their pastoralist subsistence, the Nama are genetically similar to foraging populations of the central Kalahari speaking languages belonging to all three language families. The Damara, who speak dialects of the same language as the Nama and Haiǁom and have a mixed subsistence of small-stock pastoralism and foraging, cluster closely with the pastoralist Bantu-speaking Himba and Herero. Overall, the MDS plot does not display any clear structure, with no distinct

linguistically or geographically defined clusters emerging. Only the Himba, Herero and Damara (who speak Bantu languages and a Khoe language, respectively) are grouped at a distance from all the other populations, as are the G|ui and ǂHoan, who speak a Khoe and a Kx'a language.

The presence of haplogroups L0d and L0k has a strong influence on the genetic structure of the Khoisan populations, as can be deduced from the CA analysis (Supplementary Figure 1a and b), where populations with a high frequency of L0d1, L0d2, and L0k are grouped closely together. In the CA plots, the distinction between most Khoisan and the Bantu-speaking populations is emphasized more than in the MDS analysis, as is the distinction between the Kalahari foragers (NORTHWEST, SOUTH-CENTRAL, AND CENTRAL) and the populations of the OKAVANGO and EAST clusters. The absence of genetic outliers among the Kalahari foragers suggests that the G|ui and ǂHoan, who are separated in the MDS, do not differ from their Khoisan neighbors with respect to their haplogroup composition. While strong genetic drift as well as the small sample size might account for the distinction of the ǂHoan, the G|ui are characterized by high frequencies of divergent sequence types belonging to haplogroup L0d2 (Supplementary Figure 2).

The overall lack of ethnolinguistic or geographic distinctions between the populations evident in the MDS and CA plots is confirmed by AMOVA analyses (Table 2). These underline the considerable heterogeneity of the maternal genepool in southern Africa, with a very high and significant variance observed between populations, both for the whole dataset of 26 populations (21%), as well as for the set of 19 Khoisan populations (16.6%). Focusing on the 19 Khoisan populations, different groupings were tested (Table 2). The variance between groups is very low (3.4%) and non-significant when grouping by the three language families, suggesting that simple linguistic classification is not a good predictor of genetic variation between populations. Dividing the populations in four groups by rough geographic criteria results in a significant between-group variance of 6.7%, but the between-population variance is still higher (11.3%). The between-group variance is even higher when grouping by the two phenotypes, i.e. "Khoisan phenotype" vs. "non-Khoisan phenotype"; phenotypic variation therefore correlates with genetic structure, with the highly significant between-group variance (16.7%) higher than that between populations (7.5%). This result is not unexpected, given that phenotypic traits have a biological basis and are thus more likely to be linked to populations than their linguistic affiliation or geographic location. Nevertheless, the highest between-group variance (19.4%, as opposed to only 3.9% variance between populations) is found

when grouping the Khoisan populations by the clusters selected on geographic, linguistic and subsistence criteria, suggesting that all these factors contribute to structuring the genetic variation in Khoisan (cf. Schlebusch et al.2012).

The high level of between-population variance at the maternal level emerging from the AMOVA is an important feature of our dataset. In fact, this value of between-population diversity is strikingly different from that found in other African datasets of full mtDNA sequences (Barbieri et al. 2012a, b), where the variance between distinct ethnolinguistic populations is <2% of the total. These studies focused on agriculturalist patrilocal societies with a social structure that has been shown to homogenize the maternal gene pool across different ethnolinguistic groups (Gunnarsdottir et al. 2011, Barbieri et al. 2012a, b) in the presence of strict exogamy (Kumar et al. 2006). The majority of Khoisan societies, however, are traditionally foragers, and patrilocality is not the predominant system. While the ethnographic record for the populations included in this study is often incomplete (Barnard 1992), uxorilocal postmarital residence is documented for several foraging populations: it implies residence with the bride's band for the first years after marriage and up to the birth of the third child, in association with bride service that the husband has to provide for the bride's father (Lee 1984, Heinz 1994, Silberbauer 1981, Widlok 1999). In addition, this extended period of stay with the bride's parents frequently results in permanent settlement of the young couple with the woman's band. While not strictly uxorilocal, this social behavior results in reduced female mobility in comparison to the more common patrilocal practice, and could have influenced the distribution of the maternal lineages through generations. A comparison with the paternal gene pool might shed further light on this hypothesis and complete the genetic picture of a potentially sex-biased social structure (cf. Oota et al. 2001, Gunnarsdottir et al. 2011, Heyer et al. 2012).

Genetic drift might have further increased the structure of the maternal genepool caused by reduced female mobility, since Khoisan foragers traditionally led a nomadic lifestyle within a restricted territory, where the core unit was represented by small bands of related individuals (Barnard 1992). This is confirmed by the low nucleotide diversity values found in some populations of the CENTRAL and SOUTH-CENTRAL clusters (Table 1), like ǂHoan, Taa East and Gǁana (values below 0.002), while the Bantu-speaking sedentary agriculturalists Tonga, Mbukushu, and Kalanga have the highest values (0.0042). Bayesian Skyline plots (Supplementary Figure 3), too, show reduced effective population sizes in the foraging populations of the Kalahari area (especially for the SOUTH-CENTRAL and CENTRAL clusters).

This low and constant effective population size found in SOUTH-CENTRAL and CENTRAL contrasts with the higher effective population size of the agriculturalist Bantu speakers, who furthermore show a population expansion with a steep increase at ~7,000 ya. A life-style based on hunting and gathering might be associated with conditions of long-term demographic stability, contrary to the instability characteristic of agricultural societies, where the resource supply allows for storage of surplus (resulting in population expansions) but also for the effects of famine or epidemic diseases (resulting in population bottlenecks) (Coale 1974, Bates 1955, Caldwell and Caldwell 2003).

To summarize, the majority of Khoisan populations are confirmed to be distinct in their mtDNA from their Bantu-speaking neighbors and more generally from sub-Saharan Africans. They are also quite heterogeneous in their mtDNA composition, irrespective of the high frequency of haplogroups L0d and L0k in several groups of Kalahari foragers and in contrast to perceived wisdom of their constituting a linguistically, culturally, and biologically unified group: this population heterogeneity matches the autosomal data to a certain degree (Pickrell et al. 2012, Schlebusch et al. 2012). The major social factor that could have played a role in shaping this high mtDNA diversity is the tendency for multilocal postmarital residence patterns, with a strong uxorilocal tradition in the first years after marriage, which characterizes some of the populations. In addition, in the Kalahari foragers in particular, low diversity values reflect low effective population size, making it likely that genetic drift further increased population differences. While there is genetic structure overall, Khoisan populations cannot be split into distinct groups; however, their genetic variability is best explained by the small clusters defined here on the grounds of geographic, linguistic and subsistence variation, indicating that all these factors helped shape the maternal diversity of Khoisan populations.

The impact of geography on mtDNA variation and the northwestern-southeastern split

There is a significant association between Φst distances and geographic distances for the 19 Khoisan populations (Mantel test, Z=0.33, p=0.001), indicating that geography plays a role in shaping genetic variation. The distribution of sequence types as seen in networks and analyses of haplotype sharing can provide further insights into the geographic component of the mtDNA variation. A network based on sequences belonging to haplogroup L0d1 (Figure 3) highlights the presence of common haplotypes shared between different geographic/linguistic clusters. This contrasts with the presence of long isolated branches consistent with a

considerable time depth and development in isolation (cf. Barbieri et al. 2013). Subhaplogroup L0d1b2 (with frequencies of 15% in NORTHWEST, 7% in SOUTH-CENTRAL, and 11% in CENTRAL) is composed of 31 haplotypes, nearly half of which (14, i.e. 45%) belong to NORTHWEST; the CENTRAL and SOUTH-CENTRAL clusters make up only 19% and 10% of the L0d1b2 haplotypes, respectively, with six and three haplotypes. On the opposite side of the network, L0d1c1 is the most widely represented subhaplogroup, with frequencies of 14% in the NORTHWEST, 34% in the SOUTH-CENTRAL, and 31% in the CENTRAL clusters. There are 45 haplotypes distributed at roughly equal frequency between CENTRAL (17 haplotypes, 37% of the L0d1c1 haplotypes), SOUTH-CENTRAL (15 haplotypes) and NORTHWEST (13 haplotypes) populations. A branch of haplogroup L0d1c1 is characterized by a haplotype shared by several clusters (SOUTH-CENTRAL, NORTHWEST and CENTRAL, as well as one Nama and one Tswana individual) surrounded by many NORTHWEST, SOUTH-CENTRAL and CENTRAL haplotypes in a star-shaped pattern, with other Khoe and BANTU haplotypes represented to a lesser extent. Out of a total of 40 haplotypes, only nine are shared (22.5%); of these, four are shared between SOUTH-CENTRAL and CENTRAL, and a fifth is shared between SOUTH-CENTRAL, CENTRAL and EAST. There is thus clear evidence of close ties between the SOUTH-CENTRAL and the CENTRAL clusters, who share six haplotypes on this branch, as opposed to the NORTHWEST cluster, who share only one haplotype with SOUTH-CENTRAL and two with CENTRAL.

The striking star-like pattern in L0d1c1 is consistent with a population expansion which is dated with the rho statistic (Forster et al. 1996) to be 7,290 (±1,920) years old. An explanation for this genetically detectable population expansion is not obvious: the signal of expansion is restricted to this branch of L0d1c1, which is hard to reconcile with a demographic expansion that would have affected all of the populations represented in this star-like cluster, and that should thus have left a trace in several haplogroups. An alternative explanation for the expansion detectable solely in L0d1c1 is positive selection. There is one non-synonymous mutation on the branch leading to L0d1c1, namely G9438A, which results in an amino acid change from glycine to serine in the COX3 gene, the terminal component of the respiratory chain involved in the aerobic production of energy (Fontanesi et al. 2006). This mutation is not exclusive to L0d1c1; it is present eight additional times in the entire human mtDNA phylogeny (according to Phylotree v. 15, van Oven et al. 2009), with two events occurring within the African haplogroup L2. It is thus not obvious why selection might have occurred on L0d1c1.

From the heatplot of haplotypes shared between clusters (Figure 4) we can see how the majority of haplotypes is shared between the NORTHWEST, SOUTH-CENTRAL, and CENTRAL clusters. CENTRAL displays the most sharing, with 29 haplotypes (53% of 55 haplotypes) shared with other clusters; one third of these is shared with NORTHWEST, one third is shared with SOUTH-CENTRAL. The SOUTH-CENTRAL populations share 18 of their 44 haplotypes (41%); of these, 66% are shared with CENTRAL as opposed to only 22% shared with NORTHWEST. In contrast, NORTHWEST populations share only 23% of their 94 haplotypes with other populations; of these 22 shared haplotypes, they share 50% with CENTRAL and 18% with SOUTH-CENTRAL. These numbers indicate a closer connection between SOUTH-CENTRAL and CENTRAL than between SOUTH-CENTRAL and NORTHWEST, as was also seen in the network of L0d1 (Figure 3). Furthermore, the NORTHWEST cluster emerges as being somewhat isolated from the other clusters, as evidenced by the relatively low number of haplotypes they share with others (23%), in spite of their representing the largest sample size of the dataset (162 individuals and 94 haplotypes); this predominance of exclusive haplotypes in the NORTHWEST cluster can also be seen in Figure 3.

Sharing is frequent between populations that belong to the same geographic cluster (Supplementary Figure 4), as expected from the positive correlation between genetic and geographic distances emerging in the Mantel test, which could easily derive from situations of contact between neighbors. However, many haplotypes are also shared between distinct clusters, especially the most frequent haplotypes (in the bottom of the heatplot) which are shared over wider areas. For example, excluding the first most common haplotype which is shared only between populations from Namibia (Himba, Herero, Damara, Nama and Hai‖om), the second most common haplotype is shared among the Taa, ǂHoan, G|ui, Naro, Shua and Tshwa (thus connecting SOUTH-CENTRAL and CENTRAL with EAST), and the third most common is found in Buga, ǁXo, Nama, Damara, Himba and Tonga, and is therefore found mostly in the north (with the exception of the Nama; Supplementary Figure 4).

The matrix of pairwise genetic distances (Figure 5) displays populations ordered by geographic cluster: several populations are visibly distinguished as having large genetic distances from almost all of the other populations, for example the Himba, Herero, Damara, ǁXo, Tonga, Mbukushu. Non-significant genetic distances (after applying a Bonferroni correction for multiple tests) are highlighted in the matrix: they are frequent between populations of the same cluster but also between populations from different clusters. In accordance with the signal of haplotype sharing, populations of the SOUTH-CENTRAL,

CENTRAL and NORTHWEST clusters appear genetically close to each other (with non-significant p values), with the exception of the G|ui, Taa East and ǂHoan, who are significantly different from several of their Kalahari neighbors; these are also separated in the MDS plot. Another signal of genetic proximity comes from the non-significant distances between Buga and ǁAni and the NORTHWEST populations, who are geographically close, and between the Bantu speakers from Botswana and the EAST and OKAVANGO clusters.

Overall, a moderate genetic differentiation of the NORTHWEST cluster emerges from the distribution of L0d1b2 haplotypes (Figure 3) and from the high proportion of haplotypes exclusive to this cluster, in contrast to a signal of genetic proximity of the SOUTH-CENTRAL and CENTRAL clusters. The EAST cluster appears genetically more distinct than the SOUTH-CENTRAL and CENTRAL populations, but shares haplotypes preferentially with these. A similar split was detected in the autosomal data (Pickrell et al.2012), where a much clearer structure emerged between northwestern and southeastern Kalahari Khoisan: the split between these two groups was dated within the last 30,000 years – a date in good accordance with that estimated by Schlebusch et al. (2012) based on a very different set of populations. The NW Kalahari group detected by Pickrell et al. corresponds to the NORTHWEST cluster defined here; the SE Kalahari group detected by Pickrell et al. corresponds largely to our SOUTH-CENTRAL, EAST, and CENTRAL clusters. In our data, the NW and SE Kalahari groups each contain a total of 94 haplotypes, with a large amount of haplotype sharing within each group (29% for NW Kalahari, 50% for SE Kalahari); in contrast, only seven haplotypes (7.5%) are shared between the two groups. However, in other analyses the division of the NW and SE Kalahari groups is not so clear-cut: i) an AMOVA performed with populations grouped into NW and SE Kalahari as defined in Figure S18 of Pickrell et al. (2012) (Table 2) gives a very low and non-significant between-group variance of 0.86; ii) the two groups are not separated as clearly in the MDS plot (Figure 2) as in the PCA plot based on the autosomal data; iii) some populations falling into the NW Kalahari and SE Kalahari group are not significantly differentiated (for example the Taa West, which are not significantly differentiated from any of the NORTHWEST populations, or the Ju|'hoan North, which are not differentiated from the Taa North or Taa West; cf. Figure 5); and iv) there is some sharing of haplotypes between groups (Figure 4). The split between the NW and SE Kalahari populations detected by Pickrell et al. in the autosomal data was based on analyses biased towards genetic variation specific to central Kalahari Khoisan populations (with a PC plot based on SNPs ascertained in a Ju|'hoan and with a tree constructed after excluding the effect of non-Khoisan admixture).

We therefore constructed a neighbor-joining tree based on Φst distances using only L0d and L0k sequences (Figure 6) for those populations with at least 10 individuals carrying L0d and L0k haplogroups. This separates populations of the SE Kalahari group (Taa North, Taa East, ǂHoan, Gǁana, G|ui, and Tshwa) from those of the NW Kalahari group (Ju|'hoan North, Ju|'hoan South, !Xuun, and Haiǁom). However, differences between the mtDNA sequences and the autosomal data emerge, too: the Taa West and the Shua, who in the autosomal analyses fall into the SE Kalahari group, fall on the branch with the NW Kalahari populations in the tree based on L0d/L0k sequences. Overall the mtDNA analyses thus suggest an initial population divergence between the NW and SW Kalahari groups followed by more recent contact, which was not captured in the autosomal analyses of Pickrell et al. (2012).

In order to investigate whether the mtDNA sequences shared between the NW and SE Kalahari groups are compatible with a 30,000 year old separation, we performed simulations to test how long shared haplotypes are retained after a population split (Table 3). Since new mutations (calculated as one every 3624 years, with the rate of Soares et al. 2009) will eventually erase the signal of shared haplotypes, our simulations investigate how long shared haplotypes are retained after two populations diverge, in the absence of any further contact. The results show that the probability of keeping shared haplotypes when the populations split more than 15 ky ago is zero. Shared haplotypes are present with a probability >0.05 only up to 7500 years after the split. If we take into consideration that there are seven unique haplotypes shared between the NW and SE Kalahari groups, the split would have had to occur 1000-1250 years ago in the absence of subsequent migration. Our results thus suggest that some migration and exchange throughout the area must have taken place after the split that was inferred with autosomal data to have happened within the last 30,000 years. Distinguishing recent shared ancestry from contact is difficult with autosomal SNP data; mtDNA analyses can thus complement such autosomal data, as shared mtDNA genome sequences provide a clear signal of recent contact.

Nowadays the Kalahari and surrounding areas represent the core area of settlement of the indigenous populations of southern Africa (Barnard 1992), but the presence of these hunter-gatherers in the central Kalahari itself can only be relatively recent: this area was covered with water until ~10kya, when post-glacial conditions dried the Makgadikgadi Lake (one of the largest ancient basins) and filled it with alluvial debris (Ebert and Hitchcock 1978, Cooke 1979). The lake could have represented a geographic barrier dividing northwestern populations (currently mainly speakers of Ju dialects (Kx'a family)) from southeastern

populations (currently speakers of Taa (Tuu family), ǂHoan (Kx'a family), and Khoe languages), resulting in the signal of genetic structure observed in the autosomal data (Pickrell et al. 2012 and Schlebusch et al. 2012). This deep division may also be reflected in the divergent branches in the L0d1 network, especially in L0d1b2, which makes up 15% of the NORTHWEST haplotypes, who in turn represent almost half of the total haplotypes of this branch (Figure 3). A subsequent colonization of the basin, once it dried up, is compatible with the signal of recent areal contact that emerges from the shared haplotype distribution.

In conclusion, geography plays a role in connecting neighboring populations, but the effect of contact also involves populations that are distant geographically and linguistically. Some differences emerge between northwestern and southeastern Kalahari populations, with the NORTHWEST cluster in particular appearing distinct from the southeastern populations. The possibility of an early divergence of the NW and SE Kalahari groups, which is strongly supported by the autosomal data, is complemented by the added signal of recent contact emerging from the mtDNA. Thus, comparing the structure emerging from the autosomal and the mtDNA data reveals a highly complex pattern of prehistoric population movements. However, for comprehensive insights into the prehistoric processes that may have had an impact on Khoisan genetic structure, data from extant representatives of South African and Angolan Khoisan populations are needed.

Contact and social structure in the Kalahari foragers

In the previous section, a major signal of contact and sharing emerged between three clusters: NORTHWEST, CENTRAL and SOUTH-CENTRAL, confirmed by the sharing of haplotypes in the L0d1 network and in the heatplot (Figures 3 and 4) and in mostly low and non-significant genetic distances between populations (Figure 5). The populations from these three clusters belong to the same geographic region: the core area of the Kalahari Basin. They also share common traits like the "Khoisan phenotype" and a traditional way of subsistence based on foraging. Genetically, they are characterized by very high frequencies of mtDNA haplogroups L0d and L0k and a common trend for low values of nucleotide diversity associated with not so low (or even high) values of sequence diversity (Table 1) with the exception of the ǂHoan, who are characterized by very low sequence diversity. Low nucleotide diversity values indicate reduced admixture with populations with a different genetic composition, such as the herders who migrated to the area 2,000 years ago, or the Bantu-speaking agriculturalists who

arrived later. This reduced admixture on the maternal side is in good agreement with sex-biased gene flow, since the economically "advanced" agriculturalists and herders would have a higher status and could afford to pay a lower tribute to the bride's family (if any) for forager women, while the contrary would be very unlikely (Deacon and Deacon 1999). The genetic isolation of these populations is probably also enhanced by environmental constraints represented by the harsh conditions of the semi-arid land they inhabit, where agriculturalist Bantu would not succeed. Only populations with a deep knowledge of the territory and how to locate food and water can survive for a long time, in equilibrium with the limited resources and with the cycles of the wet and dry seasons. Nevertheless, the presence of a non-Khoisan genetic component in the autosomal data (Pickrell et al. 2012) indicates that some admixture must have occurred, probably in the paternal line.

The common features displayed are probably the result of areal contact. However, this contact is not strong enough to make these populations genetically homogeneous (when pooled together in one group, the between-population variance of 7% is significant, cf. Table 2). Further evidence of potential contact can be revealed by comparisons of linguistic and genetic relationships. In the south, Taa speakers (Tuu family) predominate: while the Taa East are genetically distinct from the Taa West, the Taa North are not genetically distinct from either Taa East or Taa West (Figure 5). The Taa West and Taa North are also not significantly differentiated from any NORTHWEST or CENTRAL population, with the exception of the G|ui. There is thus a signal of genetic proximity between the Taa and their northern neighbors who speak unrelated languages that could be explained by contact or recent population divergence.

Speakers of Ju languages of the Kx'a family (Figure 1), who are settled in the northwestern Kalahari area, are genetically undifferentiated in the maternal line (cf. Table 2, Figure 5). In contrast, their linguistic relatives the ǂHoan, who live in southern Botswana, differ from the Ju|'hoan North and Ju|'hoan South, but share haplotypes with the geographically neighboring G|ui, Taa, Naro, and the Tshwa and Shua from the EAST cluster (Supplementary Figure 4); furthermore, they are not significantly differentiated from the CENTRAL populations (Figure 5). This proximity of the ǂHoan to their geographic neighbors rather than to their linguistic relatives mirrors the results from the autosomal data (Pickrell et al. 2012) and is in good agreement with linguistic evidence for contact among these populations (Traill & Nakagawa 2000, Güldemann & Loughnane 2012).

The CENTRAL cluster includes foragers of the Kalahari who speak a West Kalahari Khoe language: these are the G|ui, G‖ana and Naro, who are not significantly differentiated from

each other (Figure 5). The MDS plot separates the G|ui; since they do not differ from other central Kalahari foragers in terms of haplogroup composition (Supplementary Figure 1a and 1b), this distinction can be associated with a predominance of specific lineages of L0d2 localized in a single branch (Supplementary Figure 2). Analyses of genome-wide SNP data similarly found a very high level of private haplotypes in a mixed sample of G|ui and G‖ana with some possible Kgalagadi ancestry (Schlebusch et al. 2012). Together with the low mtDNA diversity values (Table 1), this high level of private haplotypes for both autosomal DNA and mtDNA might be caused by partial isolation and the effect of drift. The Naro are genetically closely related to both the Ju and the Taa (Figure 5), which is in agreement with autosomal evidence that they are the result of admixture between northwestern and southeastern Kalahari populations (Pickrell et al. 2012). Irrespective of their genetic affinities with the Taa and ǂHoan, the G|ui and G‖ana are distinct with respect to mtDNA from other populations speaking Khoe languages. This is in good accordance with the hypothesis of a language shift of the G|ui and G‖ana to the Khoe languages they speak nowadays (Güldemann 2008b). There is also linguistic and historical evidence for contact between speakers of G|ui and Taa (Traill & Nakagawa 2000).

Summing up, similarities between Khoisan populations are particularly evident in the core area of the Kalahari Basin, where contact has played a large role in shaping the genetic makeup of the resident foragers, and admixture with other immigrants did not leave evident traces in the maternal genetic material. These populations display values of diversity in accordance with low levels of exogamy. In this arid territory, the effect of areal contact had a visible influence on populations who speak languages belonging to all three linguistic families, as reflected in the direct exchange of haplotypes. This can be paralleled on linguistic ground: the intuition of extensive language contact within the Kalahari Basin led to its definition as a "sprachbund" (Güldemann 1998, Güldemann and Loughnane 2012).

Khoe pastoralists and a putative East African origin

The majority of the Khoe-speaking populations live in peripheral areas of the Kalahari, and it has been hypothesized that they represent the descendants of a migration of Khoe-Kwadi speakers with a herding economy (Güldemann 2008b). The putative origin of these Khoe-Kwadi populations is in East Africa, where livestock was first domesticated (Phillipson 2005, Deacon and Deacon 1999). There is some genetic evidence in support of this hypothesis: the

distribution of Y chromosome haplogroup E-M293, in association with microsatellite diversity, suggests an expansion from Tanzania to southern Africa that does not overlap with the Bantu migration (Henn et al. 2008). Autosomal data (Schlebusch et al.2012) provides evidence of shared ancestry between the Nama and East African Maasai, together with the presence of the same genetic variant for lactase persistence in both populations, which supports the suggested pastoralist character of this demographic event. Autosomal data (Pickrell et al. 2012) also suggest a tentative link to East Africa for the Nama as well as other Khoe populations, especially the Shua. Once the migrating pastoralists reached the Kalahari, it is likely that there was intensive exchange and sex-biased gene flow with resident foraging populations (Deacon and Deacon 1999): this would be reflected in a major contribution of mtDNA haplogroups L0d and L0k in the immigrating pastoralists, and a consequent homogenization of the forager and pastoralist populations.

Can a genetic signature of the pastoralist Khoe migration be identified from the mtDNA data? A potential signature would be mtDNA haplogroups and haplotypes shared among modern Khoe speakers if the pastoralist migration included female migrants, since this is assumed to have taken place not more than 2000 years ago. The lineages mostly shared by Khoe populations are haplogroups L0d (present in all populations) and L0k (present in most of them). These might represent retentions from an original shared East African ancestor, which would explain the traces of L0d in the Sandawe of Tanzania (Tishkoff et al. 2007), who speak a language possibly related to the Khoe languages (Güldemann & Elderkin 2010). However, L0d and L0k are rare outside of southern Africa (Barbieri et al. 2013), and are highly characteristic of the NORTHWEST, CENTRAL, and SOUTH-CENTRAL clusters. Thus, the presence of these lineages in the Khoe populations might rather be the result of contact with local foragers. As found for the Y chromosome, some haplogroups might retain traces of the putative East African origin of the Khoe, assuming that not all of these lineages were incorporated via direct contact with Bantu-speaking agriculturalists. A potential East African candidate is haplogroup L5, common in East Africa and present exclusively in the Shua and Tshwa (at 5 and 18%). A further trace of the Khoe migration might be sought in the presence of a minimal common genetic denominator that could be interpreted as a genetic signal of shared ancestry of these populations: however, the Khoe clusters of putative East African origin (OKAVANGO, EAST, NAMA) harbor different proportions of non-L0d/L0k haplogroups (Supplementary Table); genetic drift and/or subsequent contact with other Khoisan populations may have played a role in increasing this differentiation. A possible exception is

represented by haplogroup L3d, which is present in Khoe-speaking individuals belonging to the EAST, NAMA, and OKAVANGO clusters, and in three Hai‖om and one G|ui individual (as well as two !Xuun). However, haplogroup L3d is present at highest frequency in NW-NAMIBIA, which comprises the Khoe-speaking Damara and the Bantu-speaking Himba and Herero. The L3d network (Figure 7) shows a common haplotype shared by 28 individuals (26 NW-NAMIBIA, one Nama, and one Hai‖om, indicated with an asterisk in the network) and surrounded by 15 other haplotypes in a star-shaped form, suggesting a recent expansion. The time of this expansion is dated with the rho statistic (Forster et al. 1996) to ~1,850 years ago (± 500 years), which would coincide with the arrival of the pastoralist migrants. This haplotype stems from a motif carried by seven Khoe-speaking individuals from various regional clusters (indicated by an arrow), suggesting that the ancestors of the Khoe-Kwadi speakers could have initially carried it to the area and subsequently spread it, creating the resulting signal of expansion. Strong female gene flow could then have incorporated L3d lineages into the gene pool of the ancestors of the pastoralist Himba, Herero, and Damara (NW NAMIBIA cluster).

Among Khoisan populations, the Nama show the clearest signal of ancestry with East Africa in the autosomal data (Schlebusch et al. 2012, Pickrell et al. 2012), which strongly contrasts with the mtDNA results: the Nama do not harbor any characteristic East African mtDNA lineages, and they are genetically close to the foragers from the NORTHWEST, SOUTH-CENTRAL and CENTRAL clusters, especially to the linguistically closely related Hai‖om (Figure 2, Figure 5, Supplementary Figure 4). It is possible that high levels of contact with local foragers in the maternal line erased any original signal of East African maternal ancestry in the Nama, while a signal of East African ancestry was retained in the autosomal data, and/or the pastoralist migration was heavily male-mediated.

In summary, the variation present in the non-L0d/L0k lineages (which are less likely to stem from contact with Kalahari foragers) does not provide a strong genetic link of the Khoe-speaking populations with eastern Africa. L3d is the only genetic marker that may have been brought to southern Africa by the Khoe-Kwadi immigration, but this signal is not unequivocal. The putative genetic background carried by the maternal ancestors of the Khoe-Kwadi may have been diluted through gene flow from local foragers and Bantu-speaking migrants and further been erased by drift in some of the populations. A male-dominated migration could also have played a role in leaving a more evident signal of Eastern African origin in the Y chromosome (Henn et al. 2008), while the maternal genetic component would

stem from autochthonous foragers. The mtDNA results could also be interpreted as rejecting the hypothesis of an East African immigration of pastoralists bringing the Khoe languages to southern Africa, but this would contradict the evidence from the Y chromosome and the lactase persistence mutation that are shared with East Africa (Henn et al. 2008, Schlebusch et al. 2012). The hypothesized eastern African origin of the Khoe requires more investigation, and this line of research would greatly benefit from the availability of more representative samples, in particular from more pastoralist populations of East Africa.

CONCLUSIONS

With this dataset of complete mtDNA genome sequences we greatly extend our knowledge about the history and demography of Khoisan foragers and pastoralists of southern Africa. Most importantly, we show that the Khoisan populations are genetically differentiated, and that areal contact involving especially the Kalahari foragers played a role in shaping their mtDNA diversity. This contact may also have played a role in the diffusion of common cultural and linguistic features. Our main findings can be summarized as follows:

- The high between-population variance, not found in previous studies of complete mtDNA sequences in African populations, can be explained by the effect of drift together with a partial uxorilocal/multilocal residence pattern, in contrast to neighboring patrilocal Bantu-speaking societies. Furthermore, some forager populations show a pattern of low and constant population size that contrasts with the higher population size of the agriculturalist Bantu-speakers, who show signals of expansion.

- There is at most a subtle signal of the older divergence between northwest and southeast Kalahari populations, and a much clearer signal of recent contact between Khoisan populations, than is evident in the autosomal DNA data (Pickrell et al. 2012). Thus, analyses of complete mtDNA genome sequences can complement insights from genome-wide SNP data with evidence of subsequent contact and mixing throughout the Kalahari region.

- The Kalahari foragers are characterized by the presence of a Khoisan phenotype, a hunter-gatherer way of life, and the highest percentage of haplogroups L0d and L0k. The distinct arid environment in which they live might have played a role in keeping these indigenous populations in relative isolation, with minimal gene flow from other immigrants.

- The hypothesis of an eastern African origin for Khoe-Kwadi speakers, who would descend from a pastoral migration separate from the Bantu migration, is supported by Y chromosome

and autosomal data, but not by mtDNA. The absence of a clear mtDNA signal of this migration might be explained by subsequent intense contact with the resident foragers, and the joint effect of drift – and/or the East African component is not detected in the maternal line because the migration was male mediated.

Our results reveal a multilayered genetic perspective of the demographic patterns of populations resident in southern Africa, who are likely to be the result of admixture of different genetic substrates, such as resident forager populations and pre-Bantu pastoralists from East Africa. However, the picture presented here is limited by our lack of comparable data from descendants of Khoisan populations from South Africa and Angola. In future work, analyses of the Y-chromosome will contribute to our understanding of the genetic variation of these populations, and will complete the picture of the socio-demographic factors (in particular, those that are sex-biased) that have had an impact during Khoisan prehistory.

ACKNOWLEDGEMENTS

This study focuses on the prehistory of populations as reflected in their genetic variation. It does not intend to evaluate the self-identification or cultural identity of any group, which consist of much more than just genetic ancestry. We sincerely thank all the sample donors for their participation in this study, the governments of Botswana, Namibia, and Zambia for supporting our research, Justin Magabe and Berendt Nakwe for assistance with sample collection, and Serena Tucci, Vera Lede, Roland Schröder and Anne Butthof for assistance with sample preparation. We thank Gertrud Boden for helpful comments on the manuscript. This work, as part of the European Science Foundation EUROCORES Programme EuroBABEL, was supported by grants from the Deutsche Forschungsgemeinschaft (to BP and TG), by a Grant-in-Aid for Scientific Research (B), Ref. 19401019, Japan Society for the Promotion of Science (to HN), as well as by funds from the Max Planck Society (to BP and MS).

TABLES

TABLE 1: Populations included in the study with values of diversity

Population	Linguistic affiliation	Phenotype	Geography	Geo-linguistic cluster	nuc.div (π)	Variance	Seq.div	sd
Taa East	Tuu	Khoisan	Center	SOUTH-CENTRAL	0.0015	0.000001	0.95	0.02
Taa North	Tuu	Khoisan	Center	SOUTH-CENTRAL	0.0022	0.000001	0.94	0.03
Taa West	Tuu	Khoisan	Center	SOUTH-CENTRAL	0.0028	0.000002	0.96	0.02
ǂHoan	Kx'a	Khoisan	Center	SOUTH-CENTRAL	0.0010	0.000000	0.79	0.11
G|ui	Khoe	Khoisan	Center	CENTRAL	0.0022	0.000001	0.92	0.03
G‖ana	Khoe	Khoisan	Center	CENTRAL	0.0018	0.000001	0.98	0.03
Naro	Khoe	Khoisan	Center	CENTRAL	0.0029	0.000002	0.99	0.01
Ju|'hoan North	Kx'a	Khoisan	North	NORTHWEST	0.0028	0.000002	0.92	0.03
Ju|'hoan South	Kx'a	Khoisan	Center	NORTHWEST	0.0029	0.000002	0.98	0.01
!Xuun	Kx'a	Khoisan	North	NORTHWEST	0.0031	0.000002	0.99	0.02
Hai‖om	Khoe	Khoisan	West	NORTHWEST	0.0035	0.000003	0.98	0.01
Nama	Khoe	Khoisan	West	NAMA	0.0033	0.000003	0.99	0.01
ǁAni	Khoe	non-Khoisan	North	OKAVANGO	0.0037	0.000004	0.96	0.03
Buga	Khoe	non-Khoisan	North	OKAVANGO	0.0037	0.000004	0.90	0.06
ǁXo	Khoe	non-Khoisan	North	OKAVANGO	0.0041	0.000004	0.86	0.07
Tshwa	Khoe	non-Khoisan	East	EAST	0.0039	0.000004	0.94	0.03
Tcire Tcire	Khoe	non-Khoisan	East	EAST	0.0039	0.000004	0.97	0.04
Shua	Khoe	non-Khoisan	East	EAST	0.0039	0.000004	0.95	0.02
Damara	Khoe	non-Khoisan	West	NW-NAMIBIA	0.0028	0.000002	0.89	0.04
Herero	Bantu	-	-	NW-NAMIBIA	0.0025	0.000002	0.94	0.03
Himba	Bantu	-	-	NW-NAMIBIA	0.0024	0.000002	0.93	0.04
Kgalagadi	Bantu	-	-	BANTU	0.0037	0.000003	0.97	0.03
Tswana	Bantu	-	-	BANTU	0.0037	0.000004	0.99	0.02
Kalanga	Bantu	-	-	BANTU	0.0042	0.000005	1.00	0.02
Tonga	Bantu	-	-	BANTU	0.0042	0.000004	1.00	0.01
Mbukushu	Bantu	-	-	BANTU	0.0042	0.000005	0.99	0.02

Nuc.div: Nucleotide Diversity, sd: Standard Deviation, Seq.div: Sequence Diversity

TABLE 2: AMOVA analyses based on Φst

		Percentage of variance	
1 group		**between pops**	**within pops**
All 26 pops		20.99**	79.01
19 Khoisan pops		16.59**	83.41
11 Kalahari forager pops[a]		6.98**	93.02
OKAVANGO [b]		4.21	95.79
EAST[b]		5.81*	94.19
Ju dialect cluster[c]		1.87	98.13
Grouping Criteria (only Khoisan)	**between groups**	**between pops/within groups**	**within pops**
3 language families (Tuu, Kx'a, Khoe)[b]	3.38	14.37**	82.25
4 geographic groups (West, North, Center, East)[b]	6.68*	11.32**	82
2 phenotypes ("Khoisan", "non-Khoisan")[b]	16.67**	7.54**	75.79
7 geolinguistic clusters[b] - excluding Bantu	19.39**	3.88**	76.73
2 groups - NW Kalahari vs SE Kalahari[d]	0.86	11.15**	88

*p value<0.05

**p value<0.01

[a] Taa North, Taa East, Taa West, ǂHoan, Ju|'hoan North, Ju|'hoan South, !Xuun, Haiǁom, G|ui, Gǁana, Naro

[b] As indicated in Table 1

[c] !Xuun, Ju|'hoan North, Ju|'hoan South (see Figure 1.b)

[d] NW Kalahari: Ju|'hoan South, Ju|'hoan North, !Xuun, and Haiǁom. SE Kalahari: Taa North, Taa East, Taa West, ǂHoan, Gǁana, Shua and Tshwa (as indicated in main text)

TABLE 3: Results of simulations, with probability of retaining shared haplotypes (p) and average number of haplotypes (n) retained, for populations with different effective sizes (N_e).

n Generations		30	40	50	100	150	200	250	300	400	600	800
Years after split		750	1000	1250	2500	3750	5000	6250	7500	10000	15000	20000
N_e=100	p	0.84	0.70	0.62	0.28	0.13	0.06	0.02	0.01	0	0	0
	n	1.4	1.2	1.2	1.0	1.0	1.0	1.0	1.0	1.0	NA	NA
N_e=1000	p	1.00	0.99	0.97	0.70	0.35	0.19	0.07	0.03	0.01	0	0
	n	4.4	3.5	2.9	1.6	1.1	1.0	1.0	1.0	1.0	NA	NA
N_e=5000	p	1.00	1.00	1.00	0.92	0.66	0.34	0.16	0.08	0.01	0	0
	n	10.2	8.2	6.5	2.6	1.6	1.2	1.1	1.1	1.1	NA	NA
N_e=10,000	p	1.00	1.00	1.00	0.96	0.74	0.45	0.24	0.09	0.03	0	0
	n	14.0	11.1	8.9	3.5	1.8	1.3	1.1	1.1	1.0	NA	NA

FIGURE LEGENDS

Figure 1: a) Map of approximate location of the 26 populations included in this study, colored by linguistic affiliation. The gray area indicates the Kalahari semi-desert. b) Schema of Khoisan linguistic relationships.

Figure 2: Multidimensional Scaling plot based on Φst distances colored by linguistic affiliation as shown in Figure 1. Stress value: 7.97

Figure 3: Network of L0d1 haplotypes colored by geo-linguistic clusters as shown in Table 1.

Figure 4: Heatplot displaying the amount of haplotypes shared between geo-linguistic clusters. The most common haplotypes are in the bottom of the plot.

Figure 5: Matrix of pairwise Φst distances. The non-significant distances (after Bonferroni correction) are highlighted with a black dot.

Figure 6: Neighbor Joining tree based on Φst distances of L0d and L0k sequences

Figure 7: Network of L3d haplotypes. The dashed line indicates a branch that has been shortened for graphic purposes.

REFERENCES

Anderson, CNK, U Ramakrishnan, YL Chan, EA Hadly. 2005. Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics 21:1733-1734.

Bandelt, HJ, J Alves-Silva, PEM Guimaraes, et al. 2001. Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Ann Hum Genet 65:549-563.

Bandelt, HJ, P Forster, A Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48.

Barbieri, C, A Butthof, K Bostoen, B Pakendorf. 2012a. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet. doi: 10.1038/ejhg.2012.192. Aug 29. [Epub ahead of print]

Barbieri, C, M Whitten, K Beyer, H Schreiber, M Li, B Pakendorf. 2012b. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. Mol Biol Evol 29:1213-1223.

Barbieri, C, M Vicente, J Rocha, Sununguko W Mpoloka, M Stoneking, B Pakendorf. 2013. Ancient Substructure in Early mtDNA Lineages of Southern Africa. Am J Hum Genet 92:285-292.

Barnard, A. 1992. Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples. Cambridge ; New York: Cambridge University Press.

Barnard, A. 2008. Ethnographic analogy and the reconstruction of early Khoekhoe society. Southern African Humanities 20:61-75.

Bates, M. 1955. The prevalence of people. New York: Scribner.

Batini, C, J Lopes, DM Behar, F Calafell, LB Jorde, L van der Veen, L Quintana-Murci, G Spedini, G Destro-Bisol, D Comas. 2011. Insights into the Demographic History of African Pygmies from Complete Mitochondrial Genomes. Mol Biol Evol 28:1099-1110.

Behar, DM, R Villems, H Soodyall, et al. 2008. The Dawn of Human Matrilineal Diversity. Am J Hum Genet 82:1130-1140.

Beleza, S, L Gusmao, A Amorim, A Carracedo, A Salas. 2005. The genetic legacy of western Bantu migrations. Hum Genet 117:366-375.

Briggs, AW, JM Good, RE Green, et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318-318.

Caldwell, JC, BK Caldwell. 2003. Was there a Neolithic mortality crisis? Journal of Population Research 20:153-168.

Coale, AJ. 1974. The history of the human population. Scientific American 231.

Coelho, M, F Sequeira, D Luiselli, S Beleza, J Rocha. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evol Biol 9:80.

Cooke, HJ. 1979. The origin of the Makgadikgadi Pans. Botswana notes and records 11:37-42.

de Filippo, C, C Barbieri, M Whitten, et al. 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol 28:1255-1269.

Deacon, HJ, J Deacon. 1999. Human beginnings in South Africa: uncovering the secrets of the Stone Age. Walnut Creek, CA: Altamira Press.

Denbow, J. 1984. Prehistoric herders and foragers of the Kalahari: the evidence for 1500 years of interaction. In: C Schrire, editor. Past and Present in Hunter Gatherer Studies. Orlando: Academic Press. p.175-193.

Drummond, AJ, MA Suchard, D Xie, A Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Ebert, JI, RK Hitchcock. 1978. Ancient Lake Makgadikgadi, Botswana: mapping, measurement and palaeoclimatic significance. Palaeoecology of Africa 10:47-56.

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

Fenner, JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415-423.

Fontanesi, F, IC Soto, D Horn, A Barrientos. 2006. Assembly of mitochondrial cytochrome c-oxidase, a complicated and highly regulated cellular process. Am. J. Physiol., Cell Physiol. 291 (6): C1129–47.

Forster, P, R Harding, A Torroni, HJ Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935-945.

Furrer, R, D Nychka, S Sain. 2012. fields: Tools for spatial data. R package version 6.7 http://CRAN.R-project.org/package=fields

Greenberg, JH. 1963. The Languages of Africa. International Journal of American Linguistics 29:R5-177.

Güldemann, T. 1998. The Kalahari basin as an object of areal typology - a first approach. In: M Schladt, editor. Language, identity and conceptualization among the Khoisan. Köln: Rüdiger Köppe. p. 137-169.

Güldemann, T. 2004. Reconstruction through de-construction: The marking of person, gender, and number in the Khoe family and Kwadi. Diachronica 21:251-306.

Güldemann, T. 2005. Studies in Tuu (Southern Khoisan). Papers on Africa, Languages and Literatures 23. Leipzig: Institut für Afrikanistik, Universität Leipzig

Güldemann, T. 2008a. Greenberg's "case" for Khoisan: the morphological evidence. In: D Ibriszimow, editor. Problems of linguistic-historical reconstruction in Africa. Köln: Rüdiger Köppe. p. 123-153.

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

Güldemann, T, ED Elderkin. 2010. On external genealogical relationships of the Khoe family. In: M Brenzinger, C König, editors. Khoisan languages and linguistics: proceedings of the 1st International Symposium January 4-8, 2003: Riezlern/Kleinwalsertal. Quellen zur Khoisan-Forschung. Köln: Rüdiger Köppe,. p. 15-52.

Güldemann, T, R Loughnane. 2012. Are there “Khoisan” roots in body-part vocabulary? On linguistic inheritance and contact in the Kalahari Basin. Language Dynamics & Change 2: 1-44

Gunnarsdottir, ED, MR Nandineni, M Li, S Myles, D Gil, B Pakendorf, M Stoneking. 2011. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nature Communications 2:228.

Gusinde, M. 1966. Von Gelben und Schwarzen Buschmännern: Eine untergehende Allkultur im Süden Afrikas. Graz: Akademische Druckund Verlagsanstalt

Heine, B, H Honken. 2010. The Kx'a Family: A New Khoisan Genealogy. J Asian Afr Stud 79:5-36.

Heinz, HJ. 1994. Social organization of the! Kõ Bushmen. Köln: R. Köppe.

Henn, BM, C Gignoux, AA Lin, PJ Oefner, P Shen, R Scozzari, F Cruciani, SA Tishkoff, JL Mountain, PA Underhill. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. P Natl Acad Sci USA 105:10693-10693.

Henn, BM, CR Gignoux, M Jobin, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. P Natl Acad Sci USA 108:5154-5162.

Heyer, E, R Chaix, S Pavard, F Austerlitz. 2012. Sex-specific demographic behaviours that shape human genomic variation. Mol Ecol 21:597-612.

Jenkins, T. 1986. The prehistory of the San and Khoikhoi as recorded in their blood. In: R Vossen and K Keuthmann. Contemporary Studies on Khoisan. Hamburg, Helmut Buske Verlag. **2:** 51-77.

Kinahan, J. 1991. Pastoral Nomads of the central Namib Desert: the people history forgot. Windhoek: Namibia Archaeological Trust.

Kinahan, J. 2011. From the beginning: the archaeological evidence. In: M Wallace. A History of Namibia: From the Beginning to 1990. London: Hurst and Company. p. 15-43.

Kloss-Brandstätter, A, D Pacher, S Schönherr, H Weissensteiner, R Binna, G Specht, F Kronenberg. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25-32.

Kumar, V, BT Langstieh, KV Madhavi, VM Naidu, HP Singh, S Biswas, K Thangaraj, L Singh, BM Reddy. 2006. Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. PLoS Genet *2*(4), e53.

Lachance, J, B Vernot, CC Elbers, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150:457-469.

Lee, RB. 1984. The Dobe! Kung. Case studies in cultural anthropology. New York: Holt, Rinehart and Winston,.

Maricic, T, M Whitten, S Pääbo. 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE 5:e14004-e14004.

Meyer, M, M Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols 2010(6), pdb-prot5448

Mitchell, P. 2002. The Archaeology of Southern Africa. Cambridge: Cambridge University Press.

Naidoo, T, CM Schlebusch, H Makkan, P Patel, R Mahabeer, JC Erasmus, H Soodyall. 2010. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investig Genet 1:6.

Nenadic, O, M Greenacre. 2007. Correspondence analysis in R, with two-and three-dimensional graphics: the ca package. Journal of Statistical Software 20(3):1-13.

Oksanen, J, FG Blanchet, R Kindt, P Legendre, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MRH Stevens, H Wagner. 2012. vegan: Community Ecology Package. R package version 2.0-5. http://CRAN. R-project. org/package= vegan.

Oota, H, W Settheetham-Ishida, D Tiwawech, T Ishida, M Stoneking. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:20-21.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26:419-419.

Paradis, E, J Claude, K Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.

Patin, E, G Laval, LB Barreiro, et al. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5:e1000448.

Phillipson, DW. 2005. African archaeology. Cambridge: Cambridge University Press.

Pickrell, JK, N Patterson, C Barbieri, et al. 2012. The genetic prehistory of southern Africa. Nature

Communications 3. doi:10.1038/ncomms2140

Pleurdeau, D, E Imalwa, F Detroit, J Lesur, A Veldman, JJ Bahain, E Marais. 2012. "Of sheep and men": earliest direct evidence of caprine domestication in southern Africa at leopard cave (Erongo, Namibia). PLoS ONE 7:e40340.

Quintana-Murci, L, C Harmant, H Quach, O Balanovsky, V Zaporozhchenko, C Bormans, PD van Helden, EG Hoal, DM Behar. 2010. Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. Am J Hum Genet 86:611-620.

Reid, A, K Sadr, N Hanson-James. 1998. Herding traditions. In: P Lane, A Reid, A Segobye, editors. Ditswa MMung: The Archaeology of Botswana. Gaborone: Pula Press and The Botswana Society. p. 81-100.

Rosa, A, A Brehm, T Kivisild, E Metspalu, R Villems. 2004. MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. Ann Hum Genet 68:340-352.

Sadr, K. 1998. The first herders at the Cape of Good Hope. African Archaeological Review 15:101-132.

Salas, A, M Richards, T De la Fe, MV Lareu, B Sobrino, P Sanchez-Diz, V Macaulay, A Carracedo. 2002. The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111.

Sands, BE. 1998. Eastern and southern African Khoisan: evaluating claims of distant linguistic relationships. Köln: Rüdiger Köppe,.

Schlebusch, CM, T Naidoo, H Soodyall. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis 30:3657-3664.

Schlebusch, CM. 2010. Genetic variation in Khoisan-speaking populations from southern Africa. Johannesburg: University of the Witwatersrand.

Schlebusch, CM, M de Jongh, H Soodyall. 2011. Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. Journal of Human Genetics 56:623-630.

Schlebusch, CM, P Skoglund, P Sjodin, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374-379.

Schultze, L. 1928. Zur Kenntnis des Körpers der Hottentotten und Buschmänner. Jena: Fisher, G.

Schuster, SC, W Miller, A Ratan, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943-947.

Silberbauer, GB. 1981. Hunter and habitat in the central Kalahari Desert. Cambridge: Cambridge University Press.

Smith, AB. 1990. On becoming herders: Khoikhoi and San ethnicity in southern Africa. African Studies 49:51-73.

Smith, AB. 1992. Pastoralism in Africa: origins and development ecology. London: Hurst & Company.

Soares, P, L Ermini, N Thomson, M Mormina, T Rito, A Rohl, A Salas, S Oppenheimer, V Macaulay, MB Richards. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740-759.

Soares, P, F Alshamali, JB Pereira, V Fernandes, NM Silva, C Afonso, MD Costa, E Musilová, V Macaulay, MB Richards. 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 29:915-927.

Tishkoff, SA, MK Gonder, BM Henn, et al. 2007. History of click-speaking Populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180-2195.

Tishkoff, SA, FA Reed, FR Friedlaender, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035-1044.

Traill, A, H Nakagawa. 2000. A historical! Xóõ-| Gui contact zone: linguistic and other relations. In: H Batibo, J Tsonope, editors. The state of Khoesan languages in Botswana. Gaborone: Basarwa Languages Project. p. 1-17.

van Oven, M, M Kayser. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394.

Veeramah, KR, D Wegmann, A Woerner, FL Mendez, JC Watkins, G Destro-Bisol, H Soodyall, L Louie, MF Hammer. 2011. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal re-sequencing data. Mol Biol Evol. 29(2), 617-630.

Venables, WN, BD Ripley. 2002. MASS: modern applied statistics with S: New York: Springer.

Weiner, JS, R Harris, GA Harrison, R Singer, W Jopp. 1964. Skin Colour in Southern Africa. Hum Biol 36:294-&.

Westphal, E. 1971. The click languages of Southern and Eastern Africa. In: J Berry, JH Greenberg, editors. Linguistics in Sub-Saharan Africa. The Hague/ Paris: Mouton. p. 367-420.

Widlok, T. 1999. Living on Mangetti: 'Bushman' autonomy and Namibian independence. Oxford: Oxford University Press.

FIGURES

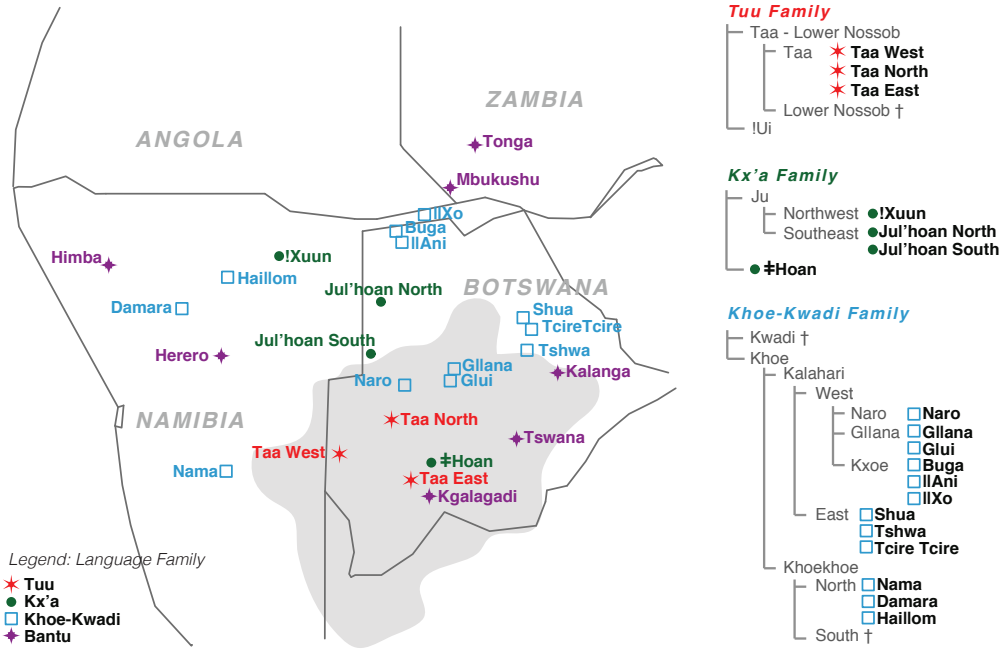


Figure 1: a) Map of approximate location of the 26 populations included in this study, colored by linguistic affiliation. The gray area indicates the Kalahari semi-desert. b) Schema of Khoisan linguistic relationships.

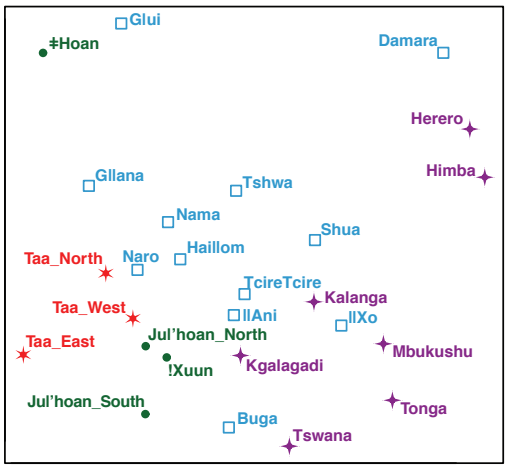


Figure 2: Multidimensional Scaling plot based on Φst distances colored by linguistic affiliation as shown in Figure 1. Stress value: 7.97

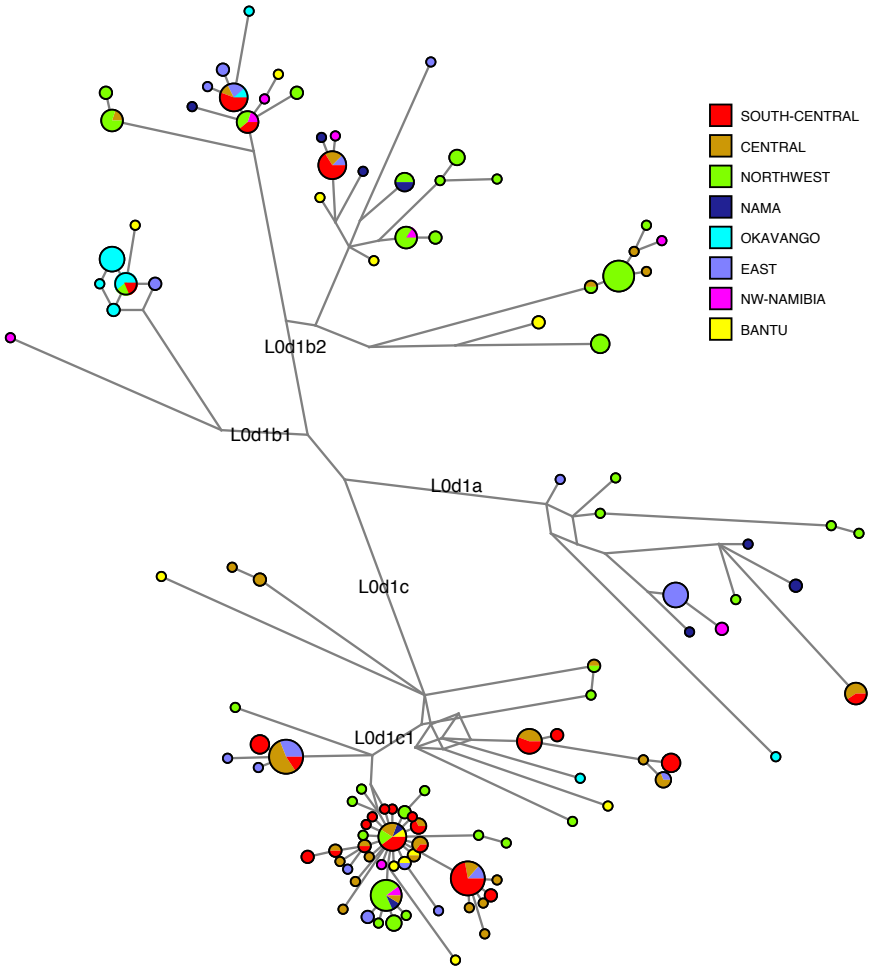


Figure 3: Network of L0d1 haplotypes colored by geo-linguistic clusters as shown in Table 1.

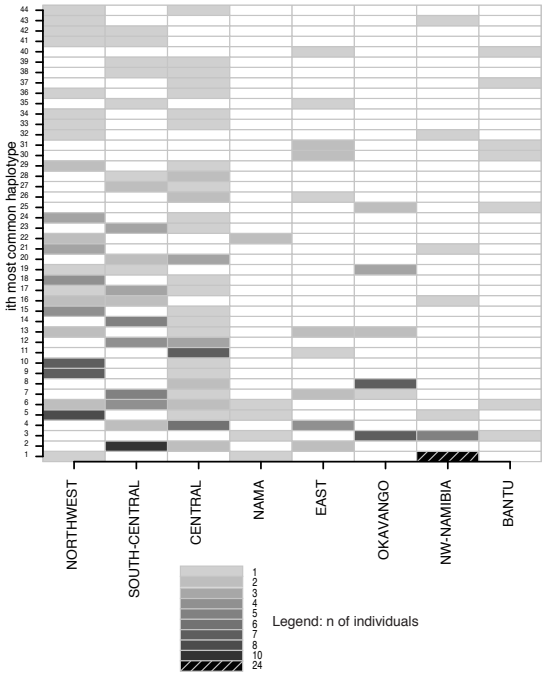


Figure 4: Heatplot displaying the amount of haplotypes shared between geo-linguistic clusters. The most common haplotypes are in the bottom of the plot.

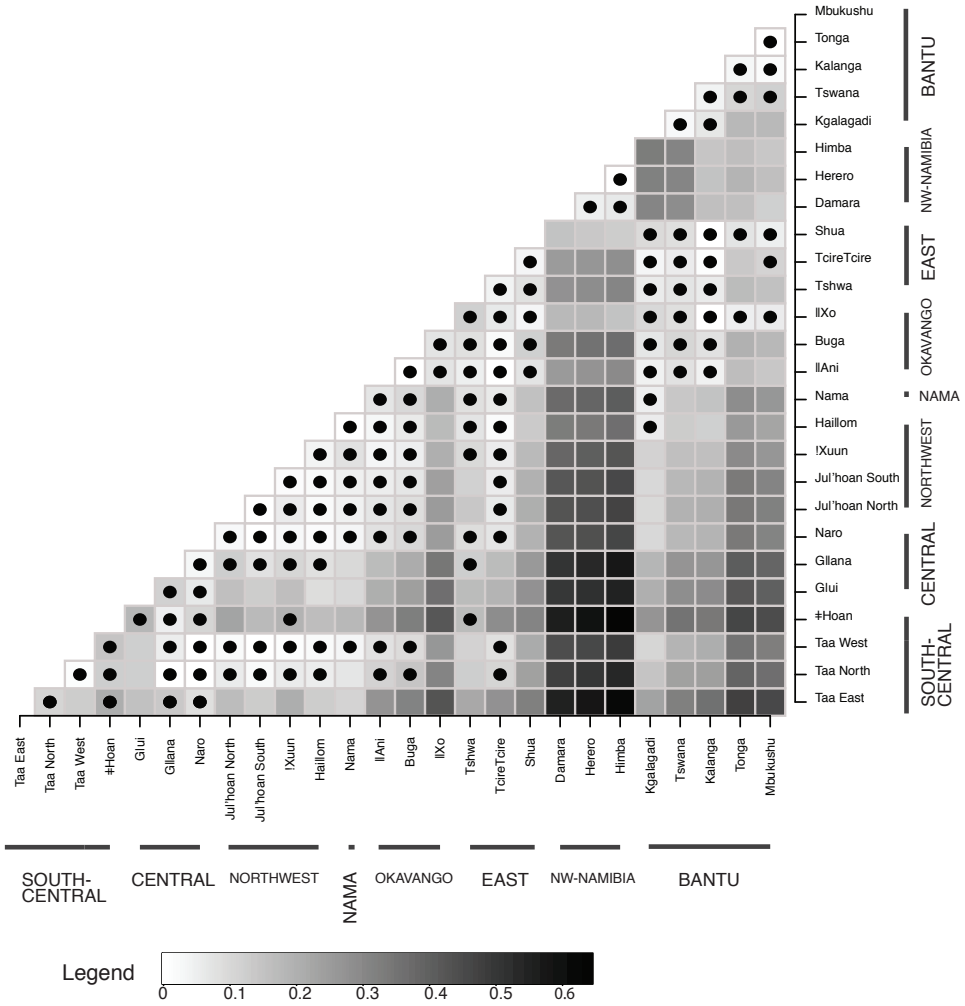


Figure 5: Matrix of pairwise Φst distances. The non-significant distances (after Bonferroni correction) are highlighted with a black dot.

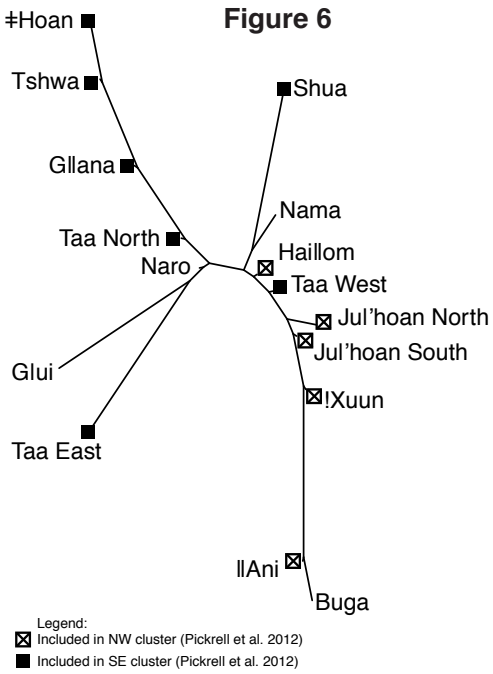


Figure 6: Neighbor Joining tree based on Φst distances of L0d and L0k sequences

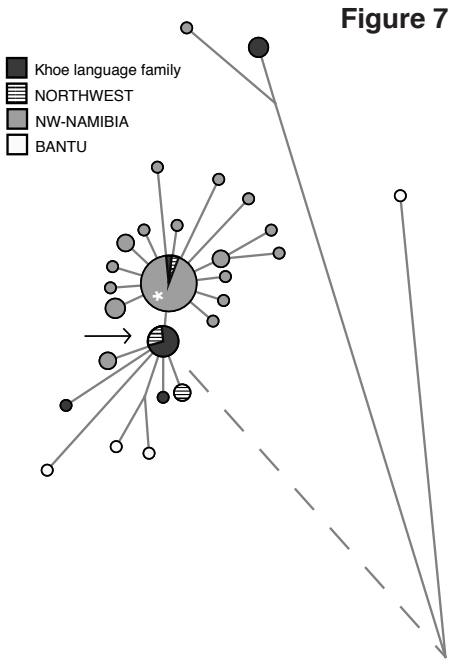
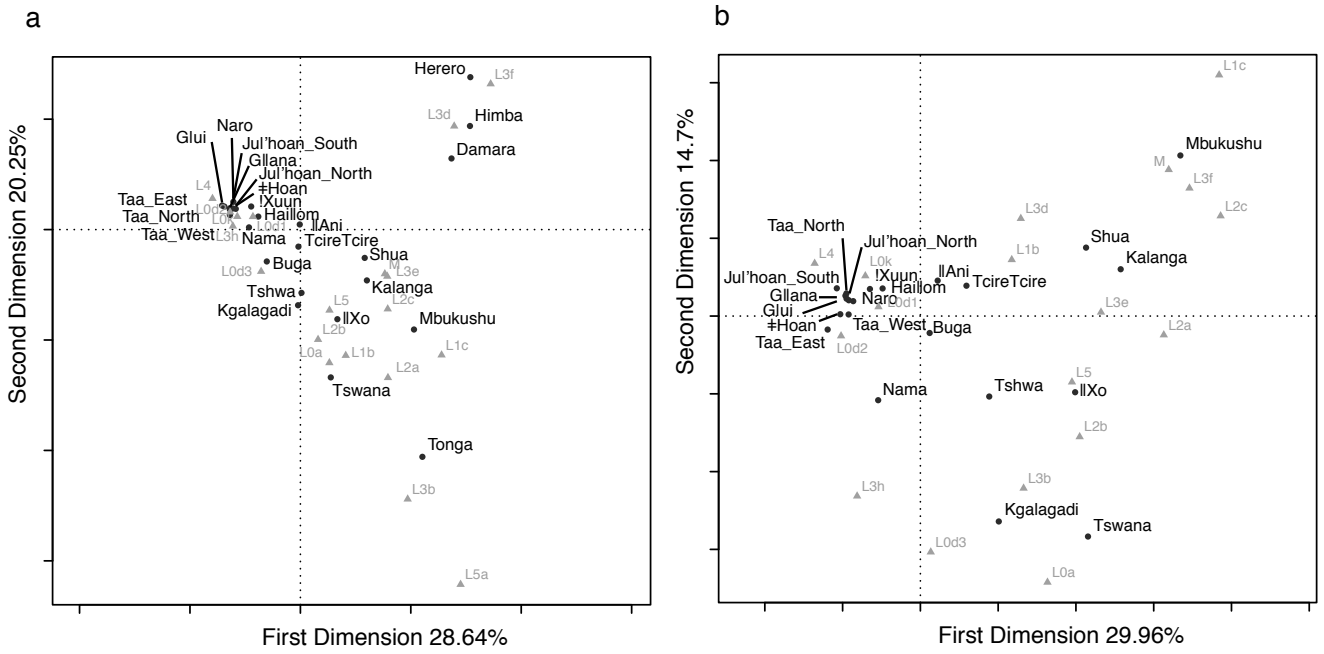
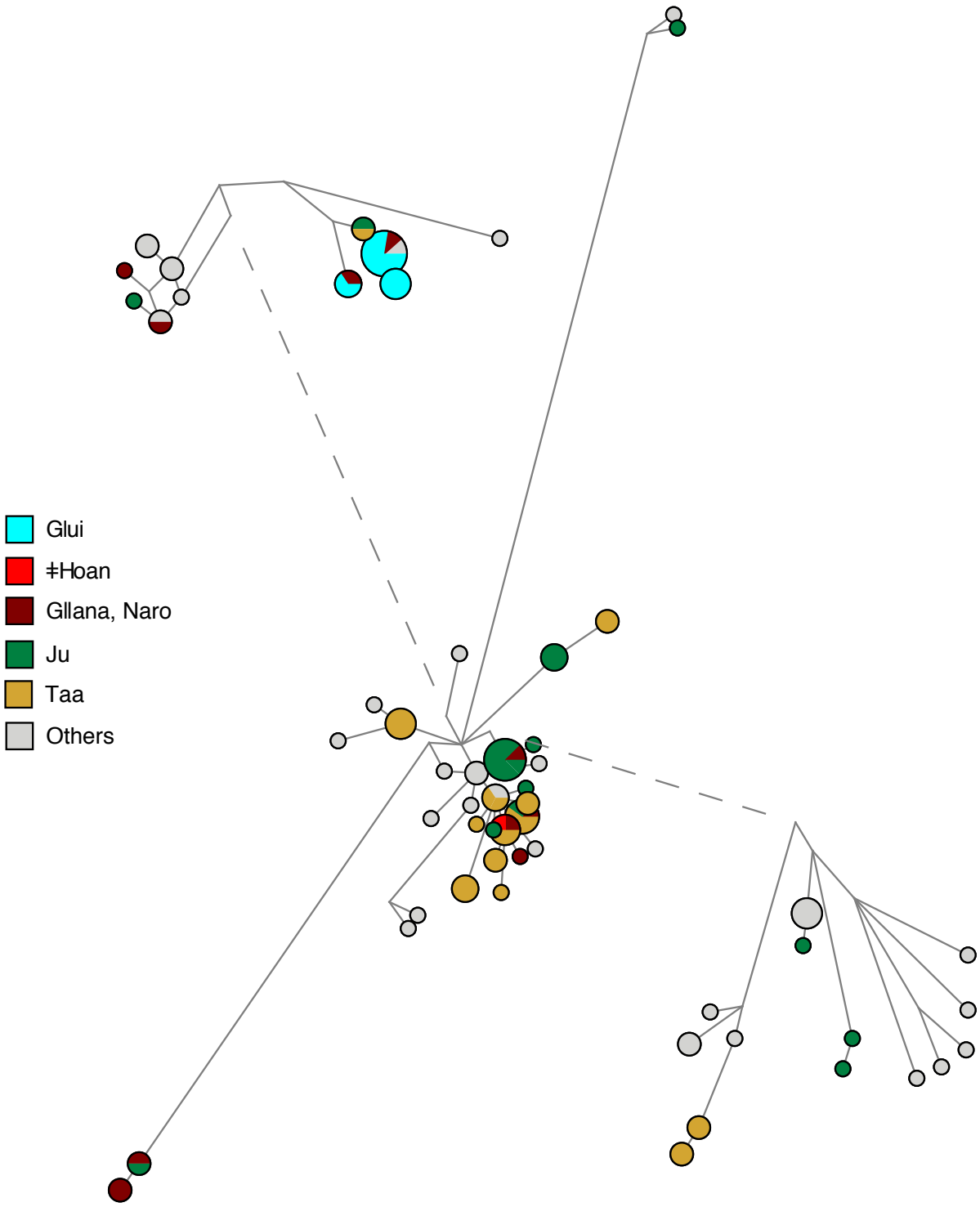


Figure 7: Network of L3d haplotypes. The dashed line indicates a branch that has been shortened for graphic purposes.

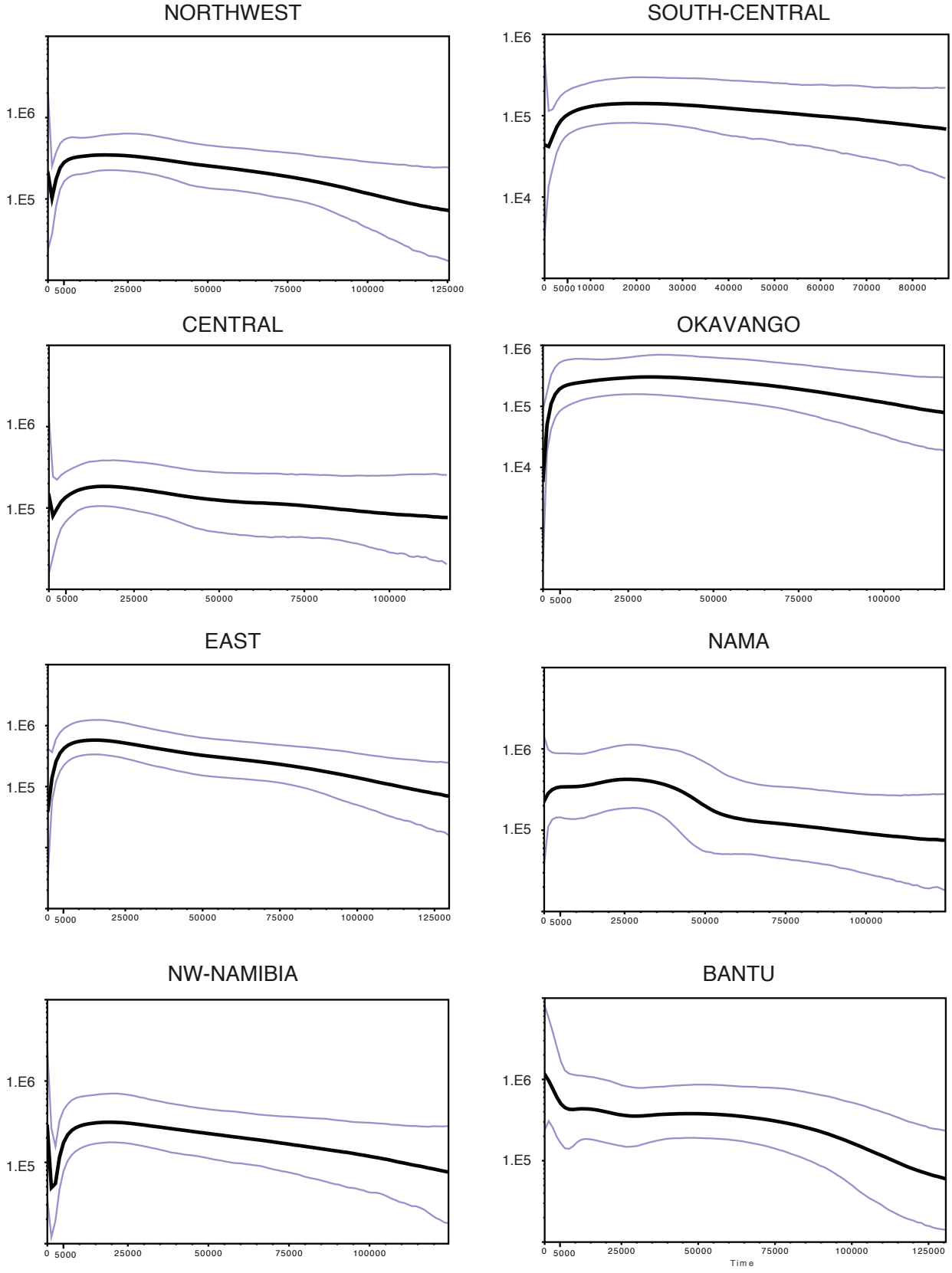
Supplementary Figure 1: Correspondence Analysis based on haplogroup frequencies.
a) entire dataset of 26 populations; b) excluding outliers Himba, Herero, Damara and Tonga.



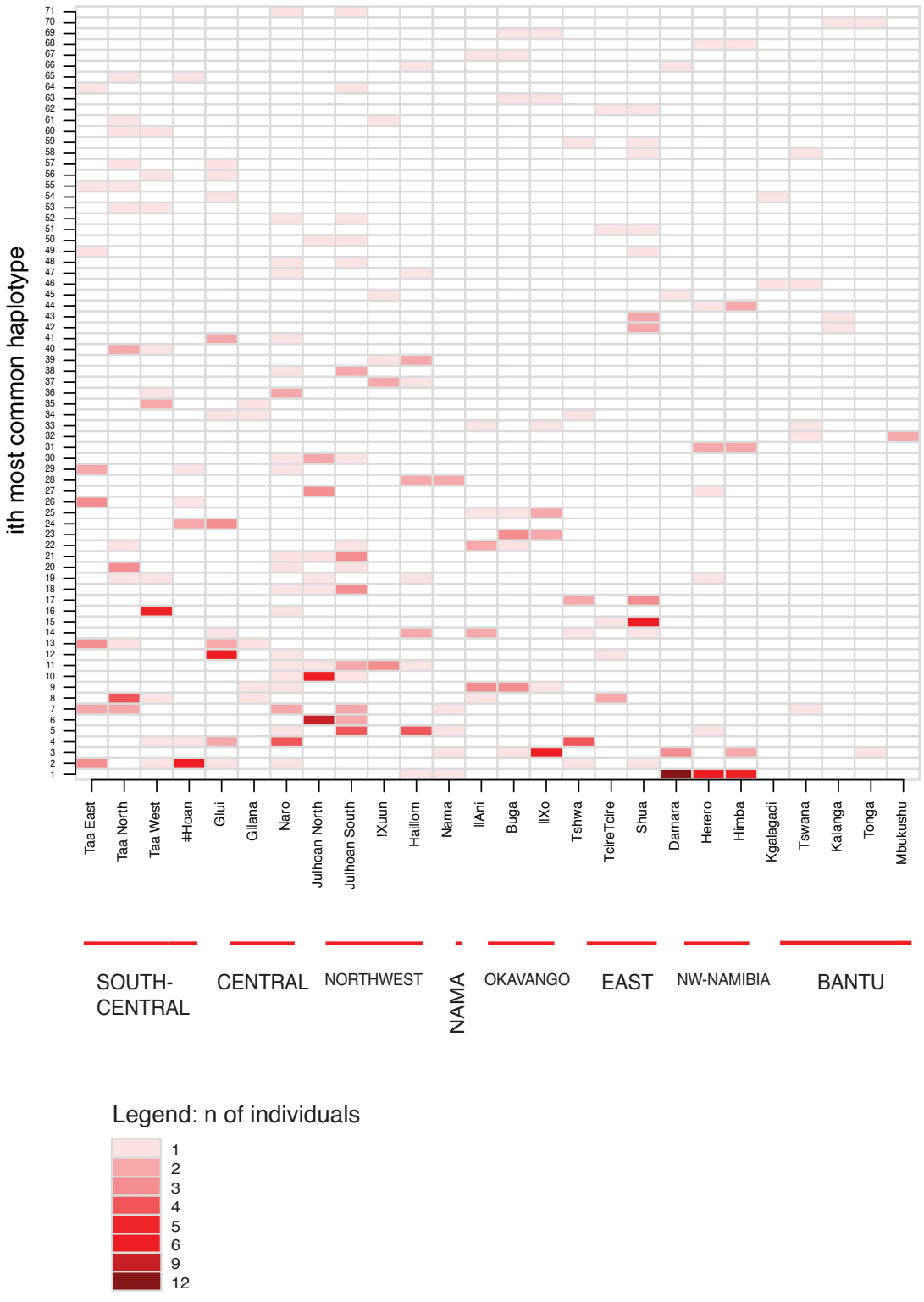
Supplementary Figure 2: Network of L0d2. The dashed lines indicate branches that have been shortened for graphic purposes.



Supplementary Figure 3: Bayesian Skyline Plots for geo-linguistic clusters. The x axis corresponds to the time from present in years; the y-axis corresponds to the effective population size x generation time. The thick black line is the median estimate and the blue lines show the 95% confidence intervals.



Supplementary Figure 4: Heatplot of haplotype sharing between individual populations



Supplementary Table: frequency of single haplogroups for each population

Populations	Linguistic affiliation	Geo-linguistic cluster	n	L0a	L0d1	L0d2	L0d3	L0k	L1b	L1c	L2a	L2b	L2c	L3b	L3d	L3e	L3f	L3h	L4	L5	M
Taa East	Tuu	SOUTH-CENTRAL	30	0.00	0.47	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Taa North	Tuu	SOUTH-CENTRAL	25	0.00	0.68	0.16	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Taa West	Tuu	SOUTH-CENTRAL	31	0.03	0.52	0.23	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ǂHoan	Kx'a	SOUTH-CENTRAL	13	0.00	0.92	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Glui	Khoe	CENTRAL	31	0.00	0.52	0.42	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Gllana	Khoe	CENTRAL	15	0.00	0.80	0.13	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Naro	Khoe	CENTRAL	35	0.00	0.51	0.26	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00
Jul'hoan North	Kx'a	NORTHWEST	40	0.00	0.50	0.23	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00
Jul'hoan South	Kx'a	NORTHWEST	44	0.00	0.50	0.20	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
!Xuun	Kx'a	NORTHWEST	27	0.04	0.44	0.11	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00
Haillom	Khoe	NORTHWEST	51	0.00	0.39	0.27	0.02	0.14	0.04	0.02	0.00	0.00	0.00	0.00	0.06	0.04	0.00	0.00	0.02	0.00	0.00
Nama	Khoe	NAMA	29	0.07	0.38	0.34	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.03	0.00	0.00	0.00
ǁAni	Khoe	OKAVANGO	18	0.06	0.44	0.00	0.00	0.22	0.06	0.00	0.06	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00
Buga	Khoe	OKAVANGO	14	0.07	0.43	0.00	0.00	0.29	0.07	0.00	0.00	0.07	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00
ǁXo	Khoe	OKAVANGO	17	0.12	0.18	0.00	0.00	0.12	0.06	0.00	0.06	0.12	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00
Tshwa	Khoe	EAST	22	0.09	0.50	0.05	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.18	0.00
Tcire Tcire	Khoe	EAST	12	0.00	0.42	0.08	0.00	0.17	0.00	0.00	0.17	0.00	0.00	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.00
Shua	Khoe	EAST	42	0.00	0.36	0.00	0.00	0.02	0.00	0.05	0.17	0.00	0.00	0.00	0.10	0.17	0.07	0.00	0.00	0.05	0.02
Damara	Khoe	NW-NAMIBIA	38	0.03	0.11	0.03	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.63	0.11	0.00	0.00	0.00	0.00	0.00
Herero	Bantu	NW-NAMIBIA	30	0.00	0.13	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.03	0.33	0.00	0.00	0.00	0.00
Himba	Bantu	NW-NAMIBIA	21	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.38	0.19	0.29	0.00	0.00	0.00	0.00
Kgalagadi	Bantu	BANTU	19	0.26	0.26	0.16	0.11	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00
Tswana	Bantu	BANTU	17	0.35	0.12	0.18	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kalanga	Bantu	BANTU	17	0.06	0.29	0.00	0.00	0.06	0.00	0.06	0.18	0.00	0.06	0.00	0.00	0.18	0.12	0.00	0.00	0.00	0.00
Tonga	Bantu	BANTU	22	0.14	0.00	0.00	0.00	0.00	0.09	0.23	0.27	0.00	0.00	0.09	0.00	0.14	0.00	0.00	0.00	0.05	0.00
Mbukushu	Bantu	BANTU	20	0.00	0.05	0.05	0.00	0.05	0.05	0.25	0.25	0.00	0.00	0.00	0.15	0.15	0.00	0.00	0.00	0.00	0.00
TOTAL			680	0.04	0.38	0.16	0.01	0.11	0.01	0.03	0.05	0.00	0.00	0.00	0.10	0.06	0.03	0.00	0.01	0.01	0.00

Chapter 10

CONCLUSIONS

Human variability in the African continent is like a multilayered puzzle: to grasp the overall picture, we have to position the pieces next to each other in the right orientation, as well as in the right depth, or time scale. In fact, the journey of our species started in Africa in the Pleistocene, and through time the vastness of this continent was covered by a tapestry of multiple population migrations, assimilations, and replacements. This intriguing scenario, far from being completely deciphered, is being progressively revealed by recent advances in genetic research, made possible by the availability of more sampled populations and fine-grained analysis.

This dissertation combines the newest analytical techniques on a considerable number of samples to investigate key questions of African human variability. Some of these questions address i) the amount of genetic variation on a continental scale and the effects of the widespread migration of Bantu speakers, ii) the extent of ancient population structure, which has been lost in present day populations, iii) the colonization of the southern edge of the continent together with the degree of population contact/replacement, and iv) the prehistory of the diverse Khoisan ethnolinguistic groups, who were traditionally understudied in spite of representing one of the most ancient divergences of modern human phylogeny.

The research questions are contextualized in a dialogue between multiple disciplines, which allows for a broader perspective. Looking at the present, we

evaluated primarily the linguistic environment, but also the social settings and the anthropological factors that play a role in each area of study. Looking at the past, we tried to collect early ethnographies and historical data, and merged it with the information retrieved from the archeological and paleoclimatic record. This challenging approach was realized thanks to the collaboration of scholars from different disciplines and of local specialists, dedicated to individual languages or populations and often involved in direct fieldwork experience.

The most relevant technical advance of this dissertation was the analysis of full mtDNA sequences from large databases. This vast amount of data was generated with the so-called "next generation" Illumina sequencing technology, and with an extensive amount of teamwork in optimizing the methodology from labwork to computational analysis. At present, cutting-edge technologies allow us to generate an increasing amount of data for each run and, in a few years, complete high-coverage genomes will be available at an affordable cost.

For a future perspective, Y chromosome sequence data should be compared to the available mtDNA sequences to complete the picture of the population prehistory, and to define the extent of sex-bias in demographic phenomena. The availability of sequence data for both uniparental markers will allow us to perform the same analysis on the maternal and paternal backgrounds. For the Y chromosome, SNPs and STR haplotypes are often the only available data; however, they do not provide enough resolution for some analyses, such as Bayesian Skyline Plots and simulations. Furthermore, the inclusion of full genomic data and the discussion of functional variants associated with phenotypes (which can be subjected to selective pressures or environmental constraints) will provide a deeper understanding of the populations of interest.

Certainly, the African continent will still be a major focus of investigation: many pieces of this complex puzzle are still waiting to be discovered. I think that the studies presented in this dissertation represent an interesting contribution to this debate and possibly open up further research threads.

Bibliography

Atkinson QD, RD Gray, AJ Drummond. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Mol Biol Evol. 25(2):468.

Alves, I, M Coelho, C Gignoux, A Damasceno, A Prista, J Rocha. 2011. Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations. Hum Biol 83:13-38.

Anderson, S, AT Bankier, BG Barrel et al. 1981. Sequence and organization of the human mitochondrial genome. Nature 290(5806): 457:465

Armitage, SJ, SA Jasim, AE Marks, AG Parker, VI Usik, H-P Uerpmann. 2011. The southern route "out of Africa": Evidence for an early expansion of modern humans into Arabia. Science 331:453-456.

Barbieri, C, A Butthof, K Bostoen, B Pakendorf. 2012a. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet. doi: 10.1038/ejhg.2012.192. Aug 29. [Epub ahead of print]

Barbieri, C, M Whitten, K Beyer, H Schreiber, M Li, B Pakendorf. 2012b. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. Mol Biol Evol 29:1213-1223.

Barbieri, C, M Vicente, J Rocha, Sununguko W Mpoloka, M Stoneking, B Pakendorf. 2013. Ancient Substructure in Early mtDNA Lineages of Southern Africa. Am J Hum Genet 92:285-292.

Barbujani, G, G Bertorelle, L Chikhi. 1998. Evidence for Paleolithic and Neolithic gene flow in Europe. American journal of human genetics 62:488-492.

Barham, L, P Mitchell. 2008. The First Africans: African Archaeology from the Earliest Toolmakers to Most Recent Foragers: Cambridge University Press.

Barnard, A. 1976. Nharo Bushmen kinship and the transformation of the Khoe kin categories. Department of Anthropology. London: University College London.

Barnard, A. 1992. Hunters and herders of southern Africa: a comparative ethnography of the Khoisan peoples. Cambridge ; New York: Cambridge University Press.

Barnard, A. 2008. Ethnographic analogy and the reconstruction of early Khoekhoe society. Southern African Humanities 20:61-75.

Bar-Yosef, O. 2000. The Middle and early Upper Paleolithic in Southwest Asia and neighboring regions. In O Bar-Yosef, Pibeam DR, editors. The geography of Neandertals and modern humans in Europe and the Greater Mediterranean Harvard, Peabody museum of Archaeology and Ethnology, Harvard University p107-156.

Bastin, Y, A Coupez, M Mann. 1999. Continuity and divergence in the Bantu languages: perspectives from a lexicostatistic study. Musée royal de l'Afrique centrale.

Batibo, H. 2005. Language decline and death in Africa: Causes, consequences, and challenges: Multilingual Matters Limited.

Batini, C, V Coia, C Battaggia, J Rocha, MM Pilkington, G Spedini, D Comas, G Destro-Bisol, F Calafell. 2007. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol 43:635-644.

Batini, C, J Lopes, DM Behar, F Calafell, LB Jorde, L van der Veen, L Quintana-Murci, G Spedini, G Destro-Bisol, D Comas. 2011. Insights into the Demographic History of African Pygmies from Complete Mitochondrial Genomes. Mol Biol Evol 28:1099-1110.

Batini, C, MA Jobling. 2011. The jigsaw puzzle of our African ancestry: unsolved, or unsolvable? Genome Biology 12:118

Behar, DM, R Villems, H Soodyall, et al. 2008. The Dawn of Human Matrilineal Diversity. Am J Hum Genet 82:1130-1140.

Beleza, S, L Gusmao, A Amorim, A Carracedo, A Salas. 2005. The genetic legacy of western Bantu migrations. Hum Genet 117:366-375.

Bender, ML. 1996. The Nilo-Saharan languages: a comparative essay: Lincom Europa.

Bender, ML. 2000. Nilo-Saharan. In: B Heine, D Nurse, editors. African languages, an introduction. Cambridge, UK: Cambridge University Press. p. 43-73.

Berniell-Lee, G, F Calafell, E Bosch, E Heyer, L Sica, P Mouguiama-Daouda, L Van Der Veen, JM Hombert, L Quintana-Murci, D Comas. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Mol Biol Evol 26:1581-1581.

Bleek, WHI. 1862. A comparative grammar of South African languages. London: Trübner. Bleek, WHI, LC Lloyd. 1911. Specimens of Bushmen Folklore. London: George Allen.

Blench, R. 2006. Archaeology, language, and the African past. Lanham, New York, London: AltaMira Press.

Blench, RM. 1992. Is Niger-Congo simply a branch of Nilo-Saharan? In Proceedings of the Fifth Nilo-Saharan Linguistic Colloquium, Nice p. 68-118.

Blench, R, M Spriggs. 1997. Archaeology and language. London ; New York: Routledge.

Boattini, A, L Castrì, S Sarno, A Useli, M Cioffi, M Sazzini, P Garagnani, S De Fanti, D Pettener, D Luiselli. 2013. mtDNA variation in East Africa unravels the history of afro-asiatic groups. American Journal of Physical Anthropology 150:375-385.

Bolnick, DA, DI Bolnick, DG Smith. 2006. Asymmetric Male and Female Genetic Histories among Native Americans from Eastern North America. Mol Biol Evol 23:2161-2174.

Bostoen, K. 2007. Pots, words and the Bantu problem: On lexical reconstruction and early African history. Journal of African History 48.

Bräuer, G. 2008. The origin of modern anatomy: By speciation or intraspecific evolution? Evolutionary Anthropology 17:22-37.

Brooks, N, I Chiapello, S Di Lernia, N Drake, M Legrand, C Moulin, J Prospero. 2005. The climate-environment-society nexus in the Sahara from prehistoric times to the present day. The Journal of North African Studies 10:253-292.

Brunet, M, F Guy, D Pilbeam, DE Lieberman, A Likius, HT Mackaye, MS Ponce de León, CPE Zollikofer, P Vignaud. 2005. New material of the earliest hominid from the Upper Miocene of Chad. Nature 434:752-755.

Campbell, MC, SA Tishkoff. 2008. African genetic diversity: implications for human demographic history, modern human origins and complex disease mapping. Annu Rev Genom Human Genet 9:403-433

Campbell, MC, SA Tishkoff. 2010. The evolution of human genetic and phenotypic variation in Africa. Curr Biol 20:R166-173.

Cann, RL, M Stoneking, AC Wilson. 1987. Mitochondrial DNA and human evolution. Nature 325:31-36.

Cashdan, EA. 1984. G//ana territorial organization. Human Ecology 12:443-463.

Cavalli-Sforza, LL, P Menozzi, A Piazza. 1994. The history and geography of human genes: Princeton university press.

Cavalli-Sforza, LL. 2001. Genes, peoples, and languages. University of California Press.

Černỳ, V, A Salas, M Hajek, M Žaloudková, R Brdička. 2007. A Bidirectional Corridor in the Sahel-Sudan Belt and the Distinctive Features of the Chad Basin Populations: A History Revealed by the Mitochondrial DNA Genome. Annals of human genetics 71:433-452.

Černỳ, V, V Fernandes, MD Costa, M Hájek, CJ Mulligan, L Pereira. 2009. Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. BMC Evolutionary Biology 9.

Černỳ, V, L Pereira, E Musilová, M Kujanová, A Vasková, P Blasi, L Garofalo, P Soares, I Diallo, R Brdička. 2011. Genetic structure of pastoral and farmer populations in the African Sahel. Mol Biol Evol 28:2491-2500.

Chandrasekar, A, SY Saheb, P Gangopadyaya, et al. 2007. YAP insertion signature in South Asia. Annals of human biology 34:582-586.

Childs, GT. 2003. An introduction to African languages: John Benjamins Publishing Company.

Conrad, DF, M Jakobsson, G Coop, X Wen, JD Wall, NA Rosenberg, JK Pritchard. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38:1251-1260.

Coppens, Y. 1983. Les plus anciens fossiles d'Hominidés. Recent advances in the evolution of Primates:1-9.

Cruciani, F, B Trombetta, A Massaia, G Destro-Bisol, D Sellitto, R Scozzari. 2011. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. The American Journal of Human Genetics 88:814-818.

Darwin, C. 1859. On the origin of species by means of natural selection. London, J. Murray.

Deacon, HJ, J Deacon. 1999. Human beginnings in South Africa: uncovering the secrets of the Stone Age. Walnut Creek, CA: Altamira Press.

Deacon, J. 1984. Later Stone Age people and their descendants in southern Africa. Southern African Prehistory and Paleoenvironments. Klein R G. Rotterdam, A. A. Balkema: 221-328.

de Filippo, C, P Heyn, L Barham, M Stoneking, B Pakendorf. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. Am J Phys Anthropol 141:382-394.

de Filippo, C, C Barbieri, M Whitten, et al. 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol 28:1255-1269.

de Filippo, C, K Bostoen, M Stoneking, B Pakendorf. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proceedings of the Royal Society B-Biological Sciences 279:3256-3263.

De Jongh, M. 2002. No fixed abode: the poorest of the poor and elusive identities in rural South Africa. Journal of Southern African Studies 28:441-460.

Denbow, J. 1984. Prehistoric herders and foragers of the Kalahari: the evidence for 1500 years of interaction. In: C Schrire, editor. Past and Present in Hunter Gatherer Studies. Orlando: Academic Press. p. pp.175-193.

Destro-Bisol, G, F Donati, V Coia, I Boschi, F Verginelli, A Caglià, S Tofanelli, G Spedini, C Capelli. 2004. Variation of Female and Male Lineages in Sub-Saharan Populations: the Importance of Sociocultural Factors. Mol Biol Evol 21:1673-1682.

Diamond, J, P Bellwood. 2003. Farmers and Their Languages: The First Expansions. Science 300:597-603.

Dimmendaal, GJ. 2008. Language ecology and linguistic diversity on the African continent. Language and Linguistics Compass 2:840-858.

Doke, CM, DT Cole. 1984. Contributions to the History of Bantu Linguistics: Papers. Witwatersrand University Press.

Dunn, M, A Terrill, G Reesink, RA Foley, SC Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309:2072-2075.

Dyen, I. 1965. A lexicostatistical classification of the Austronesian languages. International Journal of American Linguistics Memoir l9, 3l (l). Baltimore: Waverly Press.:285-305.

Dyen, I, JB Kruskal, P Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. Transactions of the American Philosophical Society.

Eggert, MKH. 2005. The Bantu problem and African archaeology. Ann Brower Stahl (ed.), African Archaeology: A Critical Introduction:301-326.

Ehret, C. 1974. Ethiopians and East Africans: The Problem of Contacts. Nairobi, East African Publishing House.

Ehret, C. 1979. On the antiquity of agriculture in Ethiopia. Journal of African History 20.

Ehret, C. 1983. In: Culture History in the Southern Sudan. Mack J, Robertshaw P, editors. Nairobi, British Institute in Eastern Africa; pp. 19-48.

Ehret, C. 2000. Language and history. African languages: An introduction:272-297.

Ehret, C. 2001. Bantu expansions: re-envisioning a central problem of early African history. The International Journal of African Historical Studies 34:5-41.

Estermann, C. 1976. The Ethnography of Southwest Angola. New York: Africana Publishing Company.

Excoffier, L, RM Harding, RR Sokal, B Pellegrini, A Sanchez-Mazas. 1991. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum Biol 63:273-307.

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

Forster, P, C Renfrew. 2011. Mother Tongue and Y Chromosomes. Science 333:1390-1391.

Forster, P, S Matsumura. 2005. Did early humans go north or south? Science 308:965-966.

Garcea, EAA. 2012. Successes and failures of human dispersals from North Africa. Quaternary International 270:119-128.

Garrigan, D, SB Kingan, MM Pilkington, JA Wilder, MP Cox, H Soodyall, B Strassmann, G Destro-Bisol, P De Knijff, A Novelletto. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177:2195-2207.

Gignoux, CR, BM Henn, JL Mountain. 2011. Rapid, global demographic expansions after the origins of agriculture. Proc Natl Acad Sci U S A 108:6044-6049.

Gil, D. 2002. Clicks in space and time. Paper presented to the Annual Conference of the North West Centre for Linguistics (NWCL) "Linguistic areas, convergence and language change", Manchester, 2nd November, 2002.

Gonder, MK, HM Mortensen, FA Reed, A de Sousa, SA Tishkoff. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24:757-768.

Gray, R. 2005. Pushing the time barrier in the quest for language roots. Science 309.5743: 2007-2008.

Greenberg, JH. 1963. The Languages of Africa. Indiana University Research Center in Anthropology. Folklore, and Linguistics, Publication.

Green, RE, J Krause, AW Briggs, T Maricic, U Stenzel, M Kircher, N Patterson, H Li, W Zhai, MH-Y Fritz. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Gregersen, EA. 1972. Kongo-Saharan. Journal of African Languages 11:69-89.

Grün, R, C Stringer, F McDermott, R Nathan, N Porat, S Robertson, L Taylor, G Mortimer, S Eggins, M McCulloch. 2005. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. Journal of Human Evolution 49:316-334.

Grün, R. 2006. Direct dating of human fossils. American journal of physical anthropology 131.S43: 2-48.

Guenther, MG. 1986. Acculturation and Assimilation of the Bushmen of Botswana and Namibia. in R Vossen and K Keuthmann, editors. Contemporary Studies on Khoisan l, :347-373.

Güldemann, T. 1998. The Kalahari basin as an object of areal typology — a first approach. In: M Schladt, editor. Language, identity and conceptualization among the Khoisan. Köln: Rüdiger Köppe. p. 137-169.

Güldemann, T, R Vossen. 2000. Khoisan. In: B Heine, D Nurse, editors. African languages, an introduction. Cambridge, UK: Cambridge University Press. p. 99-122.

Güldemann, T. 2004. Reconstruction through de-construction: The marking of person, gender, and number in the Khoe family and Kwadi. Diachronica 21:251-306.

Güldemann, T. 2005. Studies in Tuu (Southern Khoisan). Papers on Africa, Languages and Literatures 23. Leipzig: Institut für Afrikanistik, Universität Leipzig

Güldemann, T. 2007. Clicks, Genetics, and "proto-world" from a Linguistic Perspective. Leipzig: Inst. für Afrikanistik, Universität Leipzig.

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

Güldemann, T. 2008b. Greenberg's "case" for Khoisan: the morphological evidence. In: D Ibriszimow, editor. Problems of linguistic-historical reconstruction in Africa. Köln: Rüdiger Köppe. p. 123-153.

Güldemann, T, M Stoneking. 2008. A Historical Appraisal of Clicks: A Linguistic and Genetic Population Perspective. Annu Rev Anthropol 37:93-109.

Güldemann, T, ED Elderkin. 2010. On external genealogical relationships of the Khoe family. In: M Brenzinger, C König, editors. Khoisan languages and linguistics: proceedings of the 1st International Symposium January 4-8, 2003: Riezlern/Kleinwalsertal. Köln: Rüdiger Köppe,. p. 15-52.

Güldemann, T, R Loughnane. 2012. Are there "Khoisan" roots in body-part vocabulary? On linguistic inheritance and contact in the Kalahari Basin. Language Dynamics & Change 2: 1-44

Güldemann, T. forthcoming. Introduction. In Güldemann, T and A-M Fehn, editors. Beyond Khoisan: historical relations in the Kalahari Basin. Current Issues in Linguistic Theory. Amsterdam: Benjamins.

Gunnarsdottir, ED, MR Nandineni, M Li, S Myles, D Gil, B Pakendorf, M Stoneking. 2011. Larger mitochondrial DNA than Y-chromosome differences

between matrilocal and patrilocal groups from Sumatra. Nature Communications 2:228.

Gunz, P, FL Bookstein, P Mitteroecker, A Stadlmayr, H Seidler, GW Weber. 2009. Early modern human diversity suggests subdivided population structure and a complex out-of-Africa scenario. Proceedings of the National Academy of Sciences 106:6094-6098.

Gusinde, M. 1966. Von Gelben und Schwarzen Buschmännern: Eine untergehende Allkultur im Süden Afrikas. Graz: Akademische Druckund Verlagsanstalt

Guthrie, M. 1948. The classification of the Bantu languages: Pub. for the International African Institute by the Oxford Univ. Press.

Guthrie, M. 1967. Comparative Bantu: V. 1. The Comparative Linguistics of the Bantu Languages. Farnborough, UK: Gregg International.

Guthrie, M. 1971. Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages. Farnborough, UK: Gregg International.

Haile-Selassie, Y. 2001. Late Miocene hominids from the Middle Awash, Ethiopia. Nature 412:178-181.

Hammer, MF, AB Spurdle, T Karafet, MR Bonner, ET Wood, A Novelletto, P Malaspina, RJ Mitchell, S Horai, T Jenkins. 1997. The geographic distribution of human Y chromosome variation. Genetics 145.

Hammer, MF, AE Woerner, FL Mendez, JC Watkins, JD Wall. 2011. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences 108:15123-15128.

Harding, RM, G McVean. 2004. A structured ancestral population for the evolution of modern humans. Current opinion in genetics & development 14:667-674.

Harich, N, MD Costa, V Fernandes, M Kandil, JB Pereira, NM Silva, L Pereira. 2010. The trans-Saharan slave trade — clues from interpolation analyses and high-resolution characterization of mitochondrial DNA lineages. BMC Evolutionary Biology 10.

Hassan, HY, PA Underhill, LL Cavalli-Sforza, ME Ibrahim. 2008. Y-chromosome variation among Sudanese: Restricted gene flow, concordance with language, geography, and history. American Journal of Physical Anthropology 137:316-323.

Hayward, RJ. 2000. Afroasiatic. In: B Heine, D Nurse, editors. African languages: An introduction. Cambridge, UK: Cambridge University Press. p. 74-98.

Heine, B, D Nurse, editors.. 2000. African languages: An introduction: Cambridge University Press.

Heine, B, H Honken. 2010. The Kx'a Family: A New Khoisan Genealogy. J Asian Afr Stud 79:5-36.

Heinz, HJ. 1994. Social organization of the! Kõ Bushmen. Köln: R. Köppe.

Henn, BM, C Gignoux, AA Lin, PJ Oefner, P Shen, R Scozzari, F Cruciani, SA Tishkoff, JL Mountain, PA Underhill. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci U S A 105:10693-10698.

Henn, BM, CR Gignoux, M Jobin, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A 108:5154-5162.

Henn, BM, LL Cavalli-Sforza, MW Feldman. 2012. The great human expansion. Proc Natl Acad Sci U S A 109:17758-17764.

Heyer, E, R Chaix, S Pavard, F Austerlitz. 2012. Sex-specific demographic behaviours that shape human genomic variation. Molecular Ecology 21:597-612.

Hoernlé, W. 1985. The social organization of the Nama and other essays. Johannesburg: Witwatersrand Univ Pr.

Holden, CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. Proceedings of the Royal Society of London. Series B: Biological Sciences 269:793-799.

Holman, EW, C Schulze, D Stauffer, S Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations. Linguistic Typology 11:393-421.

Hublin, JJ, SP McPherron, editors. 2012. Modern Origins: A North African Perspective. Dordrecht: Springer. 244 pp.

Huxley, TH. 1870. On the geographical distribution of the chief modifications of mankind. Journal of the Ethnological Society of London (1869-1870):404-412.

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

Jakobsson, M, SW Scholz, P Scheet, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998-1003.

Jenkins, T. 1986. Genetic variation and disease in southern African. Variation, culture, and evolution in African populations: papers in honour of Dr. Hertha de Villiers.

Jenkins, T. 1986. The prehistory of the San and the Khoikhoi as recorded in their blood. Voen and Keuthmann (eds.):51-77.

Jenkins, T, GT Nurse. 1972. Blood group gene frequencies. South African medical journal 46:560.

Jenkins, T, HC Harpending, H Gordon, MM Keraan, S Johnston. 1971. Red-cell-enzyme polymorphisms in the Khoisan peoples of Southern Africa. American journal of human genetics 23.

Jenkins, T. 1986. The prehistory of the San and Khoikhoi as recorded in their blood. In: R Vossen and K Keuthmann. Contemporary Studies on Khoisan. Hamburg, Helmut Buske Verlag. 2: 51-77.

Johanson, DC, TD White. 1979. A systematic assessment of early African hominids. Science 203:321-330.

Kaufman, T, SG Thomason. 1988. Language contact, creolization and genetic linguistics. Berkeley CA: University of California.

Kinahan, J. 1991. Pastoral Nomads of the central Namib Desert: the people history forgot. Windhoek: Namibia Archaeological Trust.

Kinahan, J. 2011. From the beginning: the archaeological evidence. In: M Wallace. A History of Namibia: From the Beginning to 1990. London: Hurst and Company. p. 15-43.

Kuper, R, S Kröpelin. 2006. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. Science 313:803-807.

Lachance, J, B Vernot, CC Elbers, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150:457-469.

Ladefoged, P, A Traill. 1984. Linguistic Phonetic Descriptions of Clicks. Language 60:1-20.

Lancaster, A. 2009. Y Haplogroups, archaeological cultures and language families: a review of the multidisciplinary comparisons using the case of E-M35. J Genet Geneal 5: 35-65.

Leakey, MG, CS Feibel, I McDougall, A Walker. 1995. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. Nature 376:565-571.

Lee, RB. 1979. The !Kung San: men, women, and work in a foraging society. Cambridge, Cambridge University Press.

Lee, RB, L Marshall. 1984. The Dobe! Kung: Holt, Rinehart and Winston New York.

Lieberman, DE, BM McBratney, G Krovitz. 2002. The evolution and development of cranial form in Homo sapiens. Proc Natl Acad Sci U S A 99:1134-1139.

Li, JZ, DM Absher, H Tang, AM Southwick, AM Casto, S Ramachandran, HM Cann, GS Barsh, M Feldman, LL Cavalli-Sforza. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Liu, H, F Prugnolle, A Manica, F Balloux. 2006. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79:230-237.

Macaulay, V, C Hill, A Achilli, C Rengo, D Clarke, W Meehan, J Blackburn, O Semino, R Scozzari, F Cruciani. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034-1036.

Maho, J. 2003. A classification of the Bantu languages: an update of Guthrie's referential system. The Bantu Languages. London: Routledge. p. 639-651.

Mann, M, D Dalby, P Baker. 1987. Thesaurus of African languages: a classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation. London: Hans Zell.

Marshall, L. 1959. Marriage among !Kung Bushmen. Africa 29: 335-365.

Marshall, L. 1976. The !Kung of Nyae Nyae. Cambridge, Harvard University Press

McDougall, I, FH Brown, JG Fleagle. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733-736.

Mellars, P. 2006. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. Nature 439:931-935.

Mendez, Fernando L, T Krahn, B Schrack, et al. 2013. An African American Paternal Lineage Adds an Extremely Ancient Root to the Human Y Chromosome Phylogenetic Tree. The American Journal of Human Genetics. online http://dx.doi.org/10.1016/j.ajhg.2013.02.002.

Mitchell, P. 2002. The Archaeology of Southern Africa. Cambridge: Cambridge University Press.

Mitchell, P. 2010. Genetics and southern African prehistory: an archaeological view. Journal of Anthropological Sciences 88:73-92.

Mukarovsky, HG. 1966. Zur stellung der mandesprachen. Anthropos:679-688.

Nettle, D. 1999. Is the rate of linguistic change constant? Lingua 108:119-136.

Newman, JL. 1997. The Peopling of Africa. Yale Univ. Press; New Haven.

Nichols, J. 1997. Modeling ancient population structures and movement in linguistics. Annual review of anthropology:359-384.

Nichols, J. 2007. What, if anything, is typology? Linguistic Typology 11:231-238.

Nichols, J. 2008. Diversity and stability in language. The handbook of historical linguistics:283-310.

Naidoo, T, CM Schlebusch, H Makkan, P Patel, R Mahabeer, JC Erasmus, H Soodyall. 2010. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investig Genet 1:6.

Nurse, D, G Philippson. 2003. Towards a historical classification of the Bantu languages. The Bantu Languages:164-181.

Nurse, GT, AB Lane, T Jenkins. 1976. Sero-genetic studies on the Dama of South West Africa. Annals of human biology 3:33-50.

Nurse, GT, T Jenkins. 1977. Serogenetic studies on the Kavango peoples of South West Africa. Annals of human biology 4:465-478.

Oota, H, W Settheetham-Ishida, D Tiwawech, T Ishida, M Stoneking. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:20-21.

Pagani, L, T Kivisild, A Tarekegn, et al. 2012. Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. The American Journal of Human Genetics 91:83-96.

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

Patin, E, G Laval, LB Barreiro, et al. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5:e1000448.

Phillipson, DW. 2005. African archaeology. Cambridge: Cambridge University Press.

Pickrell, JK, N Patterson, C Barbieri, et al. 2012. The genetic prehistory of southern Africa. Nature Communications 3. doi:10.1038/ncomms2140

Pijper, A. 1932. Blood groups of Bushmen. S. Afr. Med. J 6:35-37.

Pleurdeau, D, E Imalwa, F Detroit, J Lesur, A Veldman, JJ Bahain, E Marais. 2012. "Of sheep and men": earliest direct evidence of caprine domestication in southern Africa at leopard cave (Erongo, Namibia). PLoS One 7:e40340.

Potgieter, EF. 1955. The disappearing Bushmen of Lake Chrissie: a preliminary survey. Pretoria: JL van Schaik.

Pour, NA, CA Plaster, N Bradman. 2012. Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people. European Journal of Human Genetics. doi: 10.1038/ejhg.2012.176. [Epub ahead of print].

Prugnolle, F, A Manica, F Balloux. 2005. Geography predicts neutral genetic diversity of human populations. Current Biology 15:R159-R160.

Quintana-Murci, L, C Harmant, H Quach, O Balanovsky, V Zaporozhchenko, C Bormans, PD van Helden, EG Hoal, DM Behar. 2010. Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. Am J Hum Genet 86:611-620.

Ramachandran, S, O Deshpande, CC Roseman, NA Rosenberg, MW Feldman, LL Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic

distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102:15942-15947.

Reed, FA, SA Tishkoff. 2006. African human diversity, origins and migrations. Current opinion in genetics & development 16:597-605.

Reid, A, K Sadr, N Hanson-James. 1998. Herding traditions. In: P Lane, A Reid, A Segobye, editors. Ditswa MMung: The Archaeology of Botswana. Gaborone: Pula Press and The Botswana Society. p. 81-100.

Reich, D, N Patterson, M Kircher, F Delfin, MR Nandineni, I Pugach, AM-S Ko, Y-C Ko, TA Jinam, ME Phipps. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. The American Journal of Human Genetics 89:516-528.

Renfrew, C. 2010. Archaeogenetics-towards a'new synthesis'? Current biology: CB 20.

Rexová, Kv, Y Bastin, D Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. Naturwissenschaften 93:189-194.

Richter, D, J Moser, M Nami, J Eiwanger, A Mikdad. 2010. New chronometric data from Ifri nAmmar (Morocco) and the chronostratigraphy of the Middle Palaeolithic in the Western Maghreb. Journal of Human Evolution 59:672-679.

Ruhlen, M. 1991. A Guide to the World's Languages: Volume I, Classification: Stanford University Press.

Salas, A, M Richards, T De la Fe, MV Lareu, B Sobrino, P Sanchez-Diz, V Macaulay, A Carracedo. 2002. The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111.

Sanchez-Mazas, A, ES Poloni. 2008 Genetic diversity in Africa. In: Encyclopedia of Life Sciences (ELS), eds. John Wiley & Sons , Ltd: Chichester,

Sánchez-Quinto, F, LR Botigué, S Civit, C Arenas, MC ávila-Arcos, CD Bustamante, D Comas, C Lalueza-Fox. 2012. North African Populations Carry the Signature of Admixture with Neandertals. PLoS One 7.

Sands, BE. 1998. Eastern and southern African Khoisan: evaluating claims of distant linguistic relationships. Kln: Rdiger Kppe.

Sands, B. 2009. Africa's linguistic diversity. Language and Linguistics Compass 3:559-580.

Schapera, I. 1930. The Khoisan Peoples of South Africa: Bushmen and Hottentots. London, George Routledge and Sons

Schlebusch, CM, T Naidoo, H Soodyall. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis 30:3657-3664.

Schlebusch, CM. 2010. Genetic variation in Khoisan-speaking populations from southern Africa. Johannesburg: University of the Witwatersrand.

Schlebusch, CM, M de Jongh, H Soodyall. 2011. Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. J Hum Genet 56:623-630.

Schlebusch, CM, P Skoglund, P Sjodin, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374-379.

Schultze, LE. 1928. Zur Kenntnis des Korpers der Hottentotten und Buschmanner. Zoologische und Anthropologische Ergebnisse einer Forschungsreise im westlichen und zentralen Sudafrika.

Schuster, SC, W Miller, A Ratan, LP Tomsho, B Giardine, LR Kasson, RS Harris, DC Petersen, F Zhao, J Qi. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943-947.

Scozzari, R, A Massaia, E DAtanasio, NM Myres, UA Perego, B Trombetta, F Cruciani. 2012. Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree. PLoS One 7.

Shimada, MK, K Panchapakesan, SA Tishkoff, AQ Nato, J Hey. 2007. Divergent Haplotypes and Human History as Revealed in a Worldwide Survey of X-Linked DNA Sequence Variation. Mol Biol Evol 24:687-698.

Silberbauer, GB. 1981. Hunter and habitat in the central Kalahari Desert. Cambridge: Cambridge University Press.

Silberbauer, GB, B Protectorate. 1965. Report to the Government of Bechuanaland on the Bushman Survey: Bechuanaland Government Gaberones.

Smith, AB. 1990. On becoming herders: Khoikhoi and San ethnicity in southern Africa. African Studies 49:51-73.

Soares, P, L Ermini, N Thomson, M Mormina, T Rito, A Rohl, A Salas, S Oppenheimer, V Macaulay, MB Richards. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740-759.

Soares, P, F Alshamali, JB Pereira, V Fernandes, NM Silva, C Afonso, MD Costa, E Musilová, V Macaulay, MB Richards. 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. Mol Biol Evol 29:915-927.

Soodyall, H, H Makkan, P Haycock, T Naidoo. 2008. The genetic prehistory of the Khoe and San. Southern African Humanities 20:37-48.

Steyn, HP. 1984. Southern Kalahari San Subsistence Ecology: A Reconstruction. The South African Archaeological Bulletin 39: 117-124

Stringer, C. 2002. Modern human origins: progress and prospects. Philosophical Transactions of the Royal Society B: Biological Sciences 357:563-579.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. Proceedings of the American philosophical society:452-463.

Swadesh, M, J Sherzer, D Hymes. 1971. The origin and diversification of language: Aldine De Gruyter.

Tishkoff, SA, E Dietzsch, W Speed, AJ Pakstis, JR Kidd, K Cheung, B Bonne-Tamir, AS Santachiara-Benerecetti, P Moral, M Krings. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380-1387.

Tishkoff, SA, SM Williams. 2002. Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3:611-621.

Tishkoff, SA, BC Verrelli. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340.

Tishkoff, SA, MK Gonder, BM Henn, et al. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180-2180.

Tishkoff, SA, FA Reed, FR Friedlaender, et al. 2009. The genetic structure and history of Africans and African Americans. Science 324:1035-1044.

Tömöry, G, B Csányi, E Bogácsi-Szabó, et al. 2007. Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. American Journal of Physical Anthropology 134:354-368.

Torroni, A, A Achilli, V Macaulay, M Richards, H-J Bandelt. 2006. Harvesting the fruit of the human mtDNA tree. TRENDS in Genetics 22:339-345.

Traill, A. 1973. "N4 or S7": another Bushman language. African Studies 32:25-32.

Traill, A. 1986. Do the Khoi have a place in the San?: new data on Khoisan linguistic relationships. In Rottland, Franz and Rainer Vossen (eds.), Tagungsberichte des Internationalen Symposions "Afrikanische Wildbeuter", Sankt Augustin, Januar 3-5, 1985. Sprache und Geschichte in Afrika 7,1: 407-430.

Traill, A. 1994. A !Xóõ dictionary. Cologne: R. Köppe.

Traill, A, H Nakagawa. 2000. A historical !Xóõ |Gui contact zone: linguistic and other relations. In: H Batibo, J Tsonope, editors. The state of Khoesan languages in Botswana. Gaborone: Basarwa Languages Project. p. 1-17.

Trinkaus, E. 2005. Early Modern Humans. Annual review of anthropology 34:207-230.

Underhill, PA, P Shen, AA Lin, et al. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361.

Underhill, PA, T Kivisild. 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu. Rev. Genet. 41:539-564.

van Oven, M, M Kayser. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386E394.

Vansina, J. 1995. New linguistic evidence and 'the Bantu expansion'. Journal of African History 36:173-173.

Veeramah, KR, D Wegmann, A Woerner, FL Mendez, JC Watkins, G Destro-Bisol, H Soodyall, L Louie, MF Hammer. 2011. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal re-sequencing data. Mol Biol Evol. 29(2), 617-630.

Verdu, P, NSA Becker, A Froment, et al. 2013. Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies. Mol Biol Evol, doi: 10.1093/molbev/mss328.

Wall, JD, KE Lohmueller, V Plagnol. 2009. Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations. Mol Biol Evol 26:1823-1827.

Watson, E, P Forster, M Richards, H-J Bandelt. 1997. Mitochondrial footprints of human expansions in Africa. American journal of human genetics 61.

Weiner, JS, R Harris, GA Harrison, R Singer, W Jopp. 1964. Skin Colour in Southern Africa. Hum Biol 36:294.

Wei, W, Q Ayub, Y Chen, S McCarthy, Y Hou, I Carbone, Y Xue, C Tyler-Smith. 2012. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Research.

Westphal, EOJ. 1963. The linguistic prehistory of southern Africa: Bush, Kwadi, Hottentot, and Bantu linguistic relationships. Africa:237-265.

Westphal, EOJ. 1971. The click languages of Southern and Eastern Africa. In J Berry and JH Greenberg, editors. Linguistics in Sub-Saharan Africa. Current Trends in Linguistics 7. The Hague/ Paris: Mouton, 367-420.

Whaley, L. J. 1997. Introduction to typology: The unity and diversity of language. Thousand Oaks, London and New Delhi: Sage.

White, TD, B Asfaw, Y Beyene, Y Haile-Selassie, CO Lovejoy, G Suwa, G Woldegabriel. 2009. Ardipithecus ramidus and the paleobiology of early hominids. Science 326:75-86.

Wichmann, S. 2008. The emerging field of language dynamics. Language and Linguistics Compass 2:442-455.

Wichmann, S, A Saunders. 2007. How to use typological databases in historical linguistic research. Diachronica 24.2: 373-404.

Wichmann, S, EW Holman. 2009. Assessing temporal stability for linguistic typological features. München: LINCOM Europa.

Williamson, K. 1989. Niger-congo overview. In J Bendor-Samuel and LR Hartell, editors. The Niger-Congo languages: A classification and description of Africas largest language family, 1-45.

Williamson, K, R Blench. 2000. Niger-Congo. In: B Heine, D Nurse, editors. African languages: An introduction. Cambridge, UK: Cambridge University Press p. 11-42.

Wood, ET, DA Stover, C Ehret, et al. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet 13:867-876.

Yotova, V, J-F Lefebvre, O Kohany, J Jurka, R Michalski, D Modiano, G Utermann, SM Williams, D Labuda. 2007. Tracing genetic history of modern humans using X-chromosome lineages. Human Genetics 122:431-443.