Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2446

Investigating African population structure using genomes and proteomes

IMKE TARA LANKHEET

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6214
ISBN 978-91-513-2224-7
urn:nbn:se:uu:diva-537831

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Friessalen, Evolutionsbiologiskt centrum (EBC), Norbyvägen 14, Uppsala, Friday, 25 October 2024 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Kristiina Tambets (University of Tartu, Estonia).

Abstract
Lankheet, I. T. 2024. Investigating African population structure using genomes and proteomes. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2446. 75 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2224-7.

Modern humans originated in Africa and this continent is home to the highest genetic diversity in the world. Despite this, Africa remains an understudied region. In this thesis, I use genetic and proteomic data to unravel the demographic history of modern humans in Africa. By employing various methods on different data types, I explore both broad genetic patterns, as well as fine-scale structure in African populations. In Paper I, I focus on the methodological aspects of researching maternal histories in sub-Saharan Africa. I explore mitochondrial haplogroup assignment in African populations when genome-wide SNP array data is available and conclude that mitochondrial haplogroup frequencies inferred from most common SNP arrays used for human population analysis should be considered with caution. In Paper II, I developed a long-range sequence assay to amplify full-sequence mitochondrial genomes and used this novel method to generate complete mitogenomes from underrepresented regions in Africa. Combined with published mitogenomes, these data provide an overview of African mitochondrial haplogroups and give insights into the maternal background of the African continent. By analyzing female effective population sizes over time, I discovered that two population expansions happened earlier than previously thought. In the next paper, Paper III, I utilized target enrichment to sequence portions of the Y chromosome from sub-Saharan African men. This study gives an overview of the paternal lineages of the studied populations and identifies three types of geographical distributions across the Y haplogroups in our dataset. As DNA preservation and survival are limited, I shift the focus from DNA to proteins in Paper IV. I explored the use of proteomics in ancient individuals to decipher deep population structure in Africa by investigating the potential differences in protein sequences between the two groups representing the deepest population divergence in the tree of all modern humans. I identify amino acid variation between Southern African hunter-gatherer Khoe-San groups on the one hand, and the rest of humanity on the other hand and show that these amino acids can be investigated in ancient individuals from Africa. In Paper V, I investigate the demographic histories of one population specifically; the South African Coloured (SAC), the most admixed population of South Africa. Using genotype data from these individuals, I identify geographical differences in ancestry proportions and sex-biased admixture in the SAC. Taken altogether, my work has deepened our understanding of continental and regional genetic structure in African populations.

Keywords: human evolutionary genetics, African population structure, proteomics, long-read sequencing, uniparental markers, SNP array data, NGS sequencing

Imke Tara Lankheet, Human Evolution, Department of Organismal Biology, Norbyvägen 18 A, Uppsala University, SE-752 36 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-2224-7
URN urn:nbn:se:uu:diva-537831 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-537831)

Deze thesis is opgedragen aan mijn familie, wiens liefde en steun mij hebben geïnspireerd en gemotiveerd

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Lankheet, I.T.**, Vicente, M., Barbieri, C., and Schlebusch, C.M. (2021). The performance of common SNP arrays in assigning African mitochondrial haplogroups. *BMC genomic data, 22(1), 43.*

II **Lankheet, I.T.**, Chowdhury, A., Tellgren-Roth, C., Jolly, C., Soares, A., de Navascués, M., Pacchiarotti, S., Maselli, L., Kouarata, G., Donzo, J., Coetzee, V., de Castro, M., Ebbesen, P., Priehodová, E., Podgorná, E., Černý, V., Green, S.T., Harena, P., Bakrobena, L., Fomine, F.L.M., Tolesa, Z., Mengesh, W.A., de Jongh, M., Soodyall, H., Bostoen, K., Barbieri, C., Larena, M., Malmström, H., and Schlebusch, C.M. Maternal histories of sub-Saharan Africa revealed through sequencing of complete mitochondrial genomes. *Manuscript*

III **Lankheet, I.T.**, Chowdhury, A., Jolly, C., Green, S.T., Harena, P., Bakrobena, L.,Maselli, L., Kouarata, G., Donzo, J., Pacchiarotti, S., Coetzee, V., de Castro, M., Ebbesen, P., Soodyall, H., Bostoen, K., Larena, M., Malmström, H., and Schlebusch, C.M. Y chromosome diversity and geographical patterns in sub-Saharan Africa. *Manuscript*

IV **Lankheet, I.T.**, Oskolov, N., Mirabello C., Mtetwa, E., Burgarella, C., Jolly, C., Bergquist, J., Lombard, M., Meyer, A., Steyn, M., and Schlebusch, C.M. Ancient proteins and the deepest population divergence among modern human populations. *Manuscript*

V **Lankheet, I.T.***, Hammarén R.*, Alva Caballero, L.X.*, Larena, M., Malmström, H., Jolly, C., Soodyall, H., de Jongh, M., and Schlebusch, C.M. Wide-scale Geographical Analysis of Genetic Ancestry in the South African Coloured Population. *bioRxiv*, 2024.06.13.598620. *Under Review*

* indicates equal contribution

Reprints were made with permission from the publishers.

Additional papers

The following paper was published during the course of my doctoral studies but is not part of the present dissertation.

> Vicente, M., **Lankheet, I.**, Russell, T., Hollfelder, N., Coetzee, V., Soodyall, H., Jongh, M., & Schlebusch, C. M. (2021). Male-biased migration from East Africa introduced pastoralism into southern Africa. *BMC biology*, 19(1), 259.

Contents

Abbreviations

aDNA	Ancient DNA
AMH	Anatomically Modern Human
BP	Before Present
bp	Base pairs
BSP	Bayesian skyline plot
CCS	Circular Consensus Sequencing
DEIC	Dutch East India Company
DNA	Deoxyribonucleic acid
DRC	Democratic Republic of the Congo
eRHG	Eastern Rainforest Hunter-Gatherers
ESI	Electrospray Ionization
ESSs	Effective Sample Sizes
HiFi	High-Fidelity
kb	Kilo base pairs
Kya	Thousand Years Ago
LC-MS/MS	Liquid Chromatography-Tandem Mass Spectrometry
MALDI	Matrix-assisted laser desorption/ionization
MALDI-TOF	MALDI-time-of-flight
Mb	Megabases
MCMC	Markov Chain Monte Carlo
MRCA	Most Recent Common Ancestor
mRNA	Messenger RNA
MS	Mass Spectrometry
MSY	Male specific portion of the Y chromosome
mtDNA	Mitochondrial DNA
mya	Million years ago
N_e	Effective population size
NRY	Non-recombining region of Y chromosome
PCA	Principal Component Analysis
PTMs	Post-translational modifications
rCRS	Revised Cambridge Reference Sequence
RHG	Rainforest Hunter-Gatherers
RSRS	Reconstructed Sapiens Reference Sequence
SAC	South African Coloured
SAK	Southern African Khoisan
SAP	Single Amino Acid Polymorphism
SBS	Sequencing-by-synthesis
SMRT	Single Molecule, Real-Time
SNP	Single Nucleotide Polymorphism
TMRCA	Time to the Most Recent Common ancestor
WGS	Whole-genome sequencing
wRHG	Western Rainforest Hunter-Gatherers
ZMWs	Zero-Mode Waveguides
ZooMS	Zooarchaeology by Mass Spectrometry

1. Introduction

1.1 Human Evolutionary Genetics

The field of human evolutionary genetics investigates human genomes in order to shed light on the human evolutionary past. The first disciplines to answer major questions in the field of human evolution were archaeology, paleontology, anthropology and linguistics. Whereas these disciplines are still contributing valuable information to human evolutionary questions, the arrival of DNA sequencing in the 1970'ies brought about a new perspective.

One valuable aspect of studying human evolution from a genetics' point of view is the fact that DNA, unlike many other traits, is transmitted from parents to offspring. This allows for direct investigation of genetic patterns, which can then reveal aspects about human evolution without confounding factors. Thus, studying the genomes of people living today can provide insight into past demographic events. Moreover, ancient DNA (aDNA), DNA from ancient specimens, can give a direct insight into the genetics of people living in the past.

1.1.1 Human evolution in short

The closest living relatives to us humans are chimpanzees and bonobos (Jobling et al., 2014). It is estimated from genome data that the split between humans on the one hand and chimpanzees and bonobos on the other hand happened around 7–6 million years ago (mya). Fossils from this time period are scarce, but start to become more abundant around 2.6 mya. Somewhere between 765 thousand years ago (kya) and 550 kya, modern humans split from Neanderthals and Denisovans (Prüfer et al., 2014). These two archaic hominin species themselves split from each other around 640–381 kya (Prüfer et al., 2014; Reich et al., 2010; Schlebusch et al., 2017). The first split between modern human populations is estimated to between 350 kya and 260 kya (Schlebusch et al., 2017). The earliest remains of a fully modern human are dated to $\sim$233 kya and were found in Ethiopia (White et al., 2003; Vidal et al., 2022), and fossils displaying some of the features of early anatomical modernity originate from Morocco are dated to $\sim$315 kya (Hublin et al., 2017).

It is widely accepted that modern humans evolved in Africa and subsequently spread from Africa to the rest of the world (Tishkoff and Verrelli, 2003). This is referred to as the Out-Of-Africa event. There are indications that there were smaller migrations out-of-Africa 130–90 kya, which were followed by larger migrations 70–50 kya (Jobling et al., 2014). As modern humans migrated out of Africa, they admixed with Neanderthals and Denisovans (Evans et al., 2006; Green et al., 2010; Prüfer et al., 2014; Reich et al., 2010; Huerta-Sánchez et al., 2014; Wall and Hammer, 2006). This has resulted in all modern-day non-Africans carrying 1.5–2.1% Neanderthal-derived DNA, all modern day Papuans and Australians carrying roughly 3% of Denisovan-derived DNA, and all modern people from mainland Asia and Native Americans carrying 0.2% of Denisovan-derived DNA (Prüfer et al., 2014). As a result of a long evolution of our species on the African continent, most genetic variation can be found in African individuals. The genetic variation of non-African individuals is only a subset of diversity found among Africans. In 1987, Cann et al. were the first to show this using one of the uniparental markers, mitochondrial DNA (mtDNA) (Cann et al., 1987).

However, there is much debate about the location of the origin of anatomically modern humans (AMH). AMH are members of the species *Homo sapiens* who exhibit a range of physical characteristics that are essentially identical to those found in contemporary human populations. Both East Africa and Southern Africa have been proposed as locations for the origin of our species, based on archaeological findings and genetic diversity (Barham et al., 2008; Jakobsson et al., 2008; Ramachandran et al., 2005). More recently, archaeological and genetic studies have proposed and supported a multiregional model, which entails that AMH originated and evolved in multiple places on the African continent (Henn et al., 2018; Hublin et al., 2017; Schlebusch et al., 2012).

The current geological epoch is termed the Holocene and it started 11 650 years ago. It marked the end of the last glacial period and is characterized by a warmer and more stable climate. These conditions invoked changes in the human lifestyle; in at least seven places around the world, agriculture was independently invented. Agriculture became a widespread subsistence strategy (Brown, 1999) and led to many communities living a sedentary lifestyle. However, it is not the only subsistence strategy known today. Hunter-gathering is still practiced by populations, mostly in areas where agriculture is impossible, or nearly impossible.

Throughout our species' evolution on the African continent, human populations did not stay isolated. Especially in the Holocene, we know of many populations movements that influenced the genetic landscape. In the following sections, I will describe the most important populations and migrations during the Holocene in Africa.

1.2 African population history

1.2.1 The Southern African Khoe-San populations

The earliest population divergence within modern humans has been dated to 350–260 kya (Schlebusch et al., 2017). This deepest split is between the Khoe-San people and all other modern humans. The Khoe-San people are hunter-gatherers and pastoralists and currently live in Southern Africa. The Khoe-San consist of various groups that share similar cultural, linguistic and genetic characteristics. As the name suggests, it encompasses both Khoekhoe pastoralists, as well as San hunter-gatherers (Schlebusch, 2010). Nowadays, Khoe-San populations inhabit regions of Angola, Botswana, Namibia and South Africa. The Khoe-San people have consistently been found to represent the first population divergence considering autosomal data and to carry deeply diverging mitochondrial DNA and Y chromosome lineages (Barbieri et al., 2013; Behar et al., 2008; Henn et al., 2011; Pickrell et al., 2012; Salas et al., 2002; Schlebusch et al., 2013, 2017; Tishkoff et al., 2009; Veeramah et al., 2012; Schlebusch et al., 2020; Breton et al., 2024b). aDNA from three 2 000-year-old Southern African Stone Age hunter-gatherers has shown that all modern Khoe-San groups are admixed; roughly 9-30% of their genetic ancestry is from East Africans or Europeans (Schlebusch et al., 2017).

The Khoe-San people speak Khoisan languages. The five language groups that fall under the category of Khoisan languages are unrelated, and are grouped together because of specific shared features, such as the use of click consonants. These five language groups are Hadza, Sandawe, Kx'a, Tuu, and Khoe-Kwadi (Table 1.1). Hadza and Sandawe are often referred to as the East African Khoisan languages, and Kx'a, Tuu, and Khoe-Kwadi are referred to as the Southern African Khoisan (SAK) languages, as they are spoken in the southern parts of the African continent. Within the SAK languages, one can distinguish between K'xa (or Ju, referred to as Northern Khoisan), Khoe-Kwadi (Central Khoisan) and Tuu (Southern Khoisan). Kwadi is included in the Central Khoisan languages, and was spoken in certain parts of Angola. However, it is not spoken anymore. While the K'xa and Tuu groupings only contain San hunter-gatherers, the Khoe-Kwadi language group contains both herders (Khoekhoe) and hunter-gatherers (San).

The San

Currently living San hunter-gatherers can be classified in the following linguistic subgroups: there are the Kx'a-speaking San, the Kalahari Khoe-speaking San, and the Tuu-speaking San (Figure 1). The Kx'a-speaking San comprise three main etho-linguistic groups; the !Xun, the Ju|'hoansi and the ǂHoan. In the 1980s, size estimations of these three Kx'a-speaking San populations were up to 25 000 to 30 000 individuals (Marshall and Ritchie, 1984; Lorna Marshall, 1986). The current number of San individuals is unknown. The !Xun inhabit the largest area of the three San groups and live in forested regions in northern Namibia and Southern Angola. The Ju|'hoansi is the largest ethnolinguistic group of the San and they are located centrally compared to the other two groups, in northeastern Namibia and the northwestern Kalahari Desert region of Botswana (Barnard, 1992). The ǂHoan reside in the central part of Botswana.

The Khoe-speaking San speak Kalahari Khoe, a language belonging to Khoe-Kwadi. They are distributed over a large part of Botswana and can be divided into the Tshwa, Shua, the Khwe, the G|ui and G||ana, and the Naro. The Tshwa and Shua inhabit the northeastern part of Botswana, where the Tshwa occupy a region more to the south than that of the Shua. These groups are known to live in close contact with Bantu-speaking groups (Barnard, 1992). Northwest of the Tshwa and Shua, in the Okavango swamp area, live various Khwe groups. They show high phenotypic and genetic similarity with Bantu-speaking populations (Cashdan, 1986; Nurse and Jenkins, 1977; Schlebusch et al., 2012). The origin of this similarity is unclear. These Khwe groups could have admixed with Bantu-speaking people, or they could be Bantu-speakers that lost their cattle and subsequently mixed with San groups. The G|ui and G||ana live in central Botswana and are among the most isolated groups of Khoe-San people. The Naro live in western Botswana, where there is a richness of water resources (Barnard, 1992).

The majority of the people belonging to the Tuu language division have lost their cultural identity and their language. Their descendants were taken up in other populations, such as the South African Coloured (SAC) population. There are a few San groups that speak Tuu languages; San speaking Ui! are distributed throughout South Africa, and San speaking Taa, living in the south central Kalahari of Botswana.

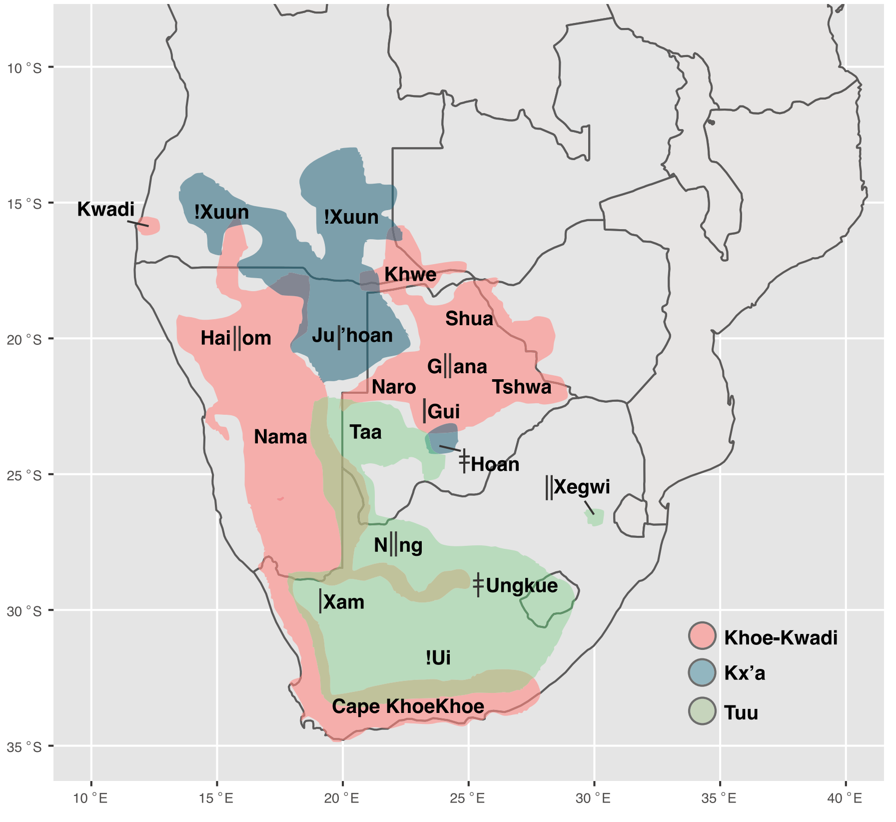


Figure 1. Distribution of the Southern African Khoisan languages on a map. Red = Khoe-Kwadi, blue = Kx'a and green = Tuu. Redrawn from Guldemann 2014 (Güldemann, 2014).

Table 1.1. *The five Khoisan languages and their sublanguages.*

<table>
<tr><th colspan="3">Linguistic lineages and subbranches</th><th>Languages and dialects</th></tr>
<tr><td>Hadza</td><td></td><td></td><td>Single language</td></tr>
<tr><td>Sandawe</td><td></td><td></td><td>Single language</td></tr>
<tr><td>Khoe-Kwadi</td><td></td><td></td><td></td></tr>
<tr><td></td><td>Kwadi</td><td></td><td>Single language (extinct)</td></tr>
<tr><td></td><td>Khoe</td><td></td><td></td></tr>
<tr><td></td><td></td><td>Kalahari Khoe</td><td></td></tr>
<tr><td></td><td></td><td></td><td>Shua, Tshwa (East)
Khwe, G||ana, Naro (West)</td></tr>
<tr><td></td><td></td><td>Khoekhoe</td><td>Cape Khoe (extinct),
!Ora-Xiri, Nama-Damara,
Eini (extinct),
Hai||om, ǂAakhoe</td></tr>
<tr><td>Kx’a</td><td></td><td></td><td></td></tr>
<tr><td></td><td>Ju</td><td></td><td></td></tr>
<tr><td></td><td></td><td>North</td><td>Angolan !Xuun varieties</td></tr>
<tr><td></td><td></td><td>North-central</td><td>Ekoka !Xuun, Okongo !Xuun, etc.</td></tr>
<tr><td></td><td></td><td>Central</td><td>Grootfontein !Xuun, etc.</td></tr>
<tr><td></td><td></td><td>Southeast</td><td>Ju|’hoansi, ǂKx’au||’e</td></tr>
<tr><td></td><td>ǂ’Amkoe</td><td></td><td>ǂHoan, N!aqriaxe, Sasi</td></tr>
<tr><td>Tuu</td><td></td><td></td><td></td></tr>
<tr><td></td><td>Taa-Lower Nossob</td><td></td><td></td></tr>
<tr><td></td><td></td><td>Taa</td><td>West: West !Xoon, N|u||’en
East: East !Xoon, ’N|oha,
N|amani, Kakia</td></tr>
<tr><td></td><td></td><td>Lower Nossob</td><td>|’Auni, |Haasi (both extinct)</td></tr>
<tr><td></td><td>Ui!</td><td></td><td>N||ng, Xam (extinct),
ǂUngkue (extinct),
|Xegwi</td></tr>
</table>

The migration of East African pastoralists to Southern Africa

During the late Holocene, East African pastoralists migrated from Eastern Africa to Southern Africa. The origins of pastoralist practices in East Africa can be found outside the African continent, in the Levant (Tishkoff et al., 2009; Pagani et al., 2012; Pickrell et al., 2014; Prendergast et al., 2019). Although the earliest findings of domestic cattle in Eastern Africa date to around 5 000 before present (BP) (Marshall and Hildebrand, 2002), the introduction of the Eurasian genetic component into northeastern Africa is complex and difficult to assess from modern-day genetic data alone (Hammarén et al., 2023); dates vary from 1 027–73 BP. aDNA data detected the non African component as early as 4 000 BP in East Africa (Prendergast et al., 2019). Pastoralist practices spread from East Africa to Southern Africa, accompanied by a Eurasian genetic component. The people spreading pastoralism from East Africa mixed with the local Southern African hunter-gatherer groups, which gave rise to the ancestors of the Khoekhoe herders. It has been shown that the population that admixed with the Southern African hunter-gatherer groups carried a Eurasian genetic component of around 31%, and an East African genetic component of

69% (Schlebusch et al., 2017). Skoglund *et al.* also found Levantine ancestry ($\sim$38%) in a 3 000 year old individual from Tanzania (Skoglund et al., 2017). Using modern genomes, it was estimated that this (West-) Eurasian ancestry arrives in Southern Africa around 1 800–900 BP (Pickrell et al., 2014). This is slightly later than the archaeological evidence in Southern Africa, which suggests the presence of pastoralism in Southern Africa around 2 100 BP (Sealy and Yates, 1994; Coutu et al., 2021). aDNA data detected the pastoralist genetic component in Southern Africa as early as 1 200 BP, in an individual that was found in a pastoralist context, and carried around 40% mixed East African/Eurasian ancestry and 60% San ancestry.

Khoekhoe settlements were very common in Southern Africa before the arrival of the colonists in the 16th and 17th century. Several Khoekhoe languages (!Ora, Eini and various Cape Khoekhoe languages) were spoken in what is now Namibia, Botswana and South Africa. The Khoekhoe people that used to live in the Eastern and Western Cape provinces of South Africa are referred to as the Cape Khoekhoe, but they gradually disappeared, and lost their culture and their language. Their descendants were taken up in a group with mixed ancestries, which is referred to as the South African Coloured population (De Wit et al., 2010). Currently, the only remaining Khoekhoe pastoralist group that retained their language and cultural identity is the Nama, who mainly live in current-day central and southern Namibia (Figure 1). They can be divided into two groups; the Great Nama and the Little Nama (Barnard, 1992). Both groups probably lived in what is now known as the Northern Cape province of South Africa. The Great Nama already moved to Namibia before the arrival of the Europeans, and the Little Nama moved to Namibia in the 19th century. The Nama show the highest proportion (23–30%) of East African ancestry among all Southern African Khoe-San groups (Schlebusch et al., 2017). The Hai||om and Damara also speak Khoekhoe languages, but the Damara most likely acquired it through interaction with the Nama, as their genetic background is mainly Bantu-speaker-related (Barnard, 1992; Barbieri et al., 2014; Pickrell et al., 2012; Vicente et al., 2019). The Hai||om live in northern Namibia and speak Khoekhoe languages. Genetically, the Hai||om resemble the Kx'a speaking !Xun and Ju|'hoansi. Therefore, it has been suggested that the Hai||om could have been !Xun, who shifted their language and subsistence practice due to contact with the Nama (Barnard, 1992; Vicente et al., 2019).

1.2.2 The expansion of Bantu-speaking people

An event that had a major influence on the current-day genetic composition of sub-Saharan African populations is the migration of Bantu-speaking people. This spread of people is often referred to as the Bantu Expansion and its success at the later stages is connected to farming practices and the use of iron working. Farming practices originated independently in different parts of the world (Jobling et al., 2014). It is generally considered that the African continent hosts at least three regions in which agriculture has independently developed: the Ethiopian highlands around 7 000 to 4 000 BP, the Sahel belt around 7 000 BP, and Western Africa around 5 000 to 3 000 BP.

Western Africa, more specifically eastern Nigeria and Western Cameroon, is the homeland of the Bantu Expansion (Diamond and Bellwood, 2003). New subsistence techniques and changes in climatic conditions allowed populations to grow and expand far beyond Western Africa, starting around 5 000 to 3 000 BP. With this expansion, specific cultural packages, genetics, language, and probably iron metallurgy spread across most of sub-Saharan Africa (Li et al., 2014; Mitchell and Lane, 2013; Patin et al., 2017). At the later stages of the expansion, adoption of farming practices had a big impact on the environment, and also changed human societies in terms of culture, demographics and possibly diseases prevalent among humans. It has been shown that the Bantu Expansion was the result of a so-called demic diffusion; the movement of people, rather than just the dispersal of language and material culture (cultural diffusion) (De Filippo et al., 2012). Around 1 800–1 300 BP, Bantu-speaking people arrived in Southern Africa (Ribot et al., 2010), reaching as far as northern Namibia and the east of South Africa.

Joseph Greenberg originally proposed that there are four linguistic phyla in Africa (Greenberg, 1963): Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Khoisan. However, the genealogical validity of some of these phyla has been questioned by several linguists (Dimmendaal, 2008; Güldemann, 2018). Despite the linguistic obsolescence of the linguistic phyla proposed by Greenberg, genome-wide genetic analyses by Tishkoff et al. (2009) have demonstrated that populations grouped under these four linguistic labels proposed by Greenberg, form distinct genetic clusters. Therefore, these genetic clusters are often referred to using the linguistic labels. Bantu languages are part of the Niger-Congo language phyla (Mitchell and Lane, 2013). Currently, there are more than 300 million Africans speaking one of the 500 Bantu languages. The Bantu languages are closely related, despite their wide distribution over an area of roughly 500 000 km^2. Four major groups can be distinguished among the Bantu languages: North-Western Bantu, Eastern Bantu, West-Western and South-Western Bantu (Figure 2) (Van de Velde et al., 2019). The widest variety among the Bantu languages can be found among the North-Western Bantu languages, which are mainly spoken in and around the homeland of the Bantu Expansion (Mitchell and Lane, 2013). The split between Eastern and Western Bantu languages and their spread from the Bantu homeland to most of sub-Saharan Africa has been a topic of debate. One hypothesis, the "Early-split" hypothesis, suggests that there was an early split around 4 000 BP north of the rainforest, followed by a spread of Eastern Bantu speakers north of the rainforest and a spread of Western Bantu speakers southwards through the rainforest (Figure 3a). Another hypothesis, the "Late-split" hypothesis, proposes a common migration through the rainforest, after which the two groups separated around 2 000 BP (Figure 3b). Linguistic and genetic data mainly support the "Late-split" model (De Filippo et al., 2012; Fortes-Lima et al., 2024; Semo et al., 2019; Grollemund et al., 2023).

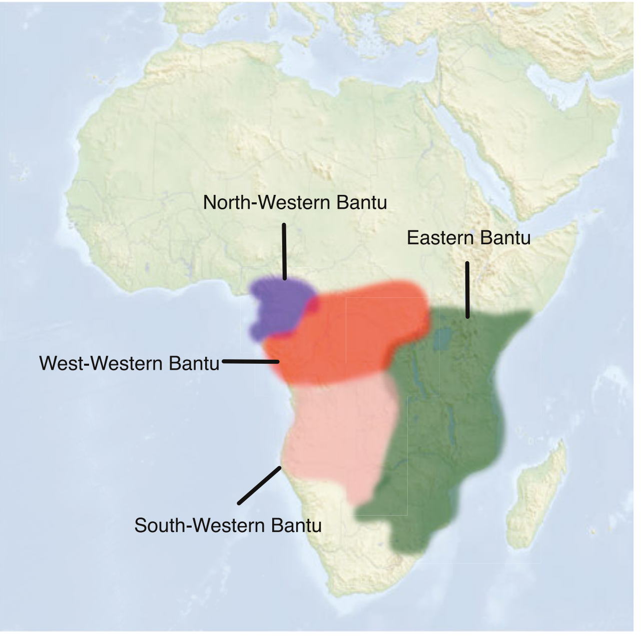
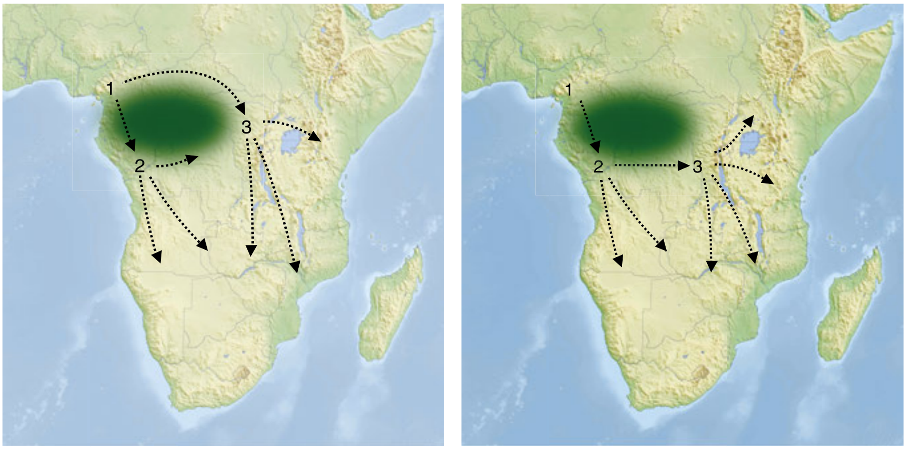


Figure 2. Rough geographic distribution of the four major groups of Bantu languages; North-Western Bantu (purple), West-Western Bantu (red), South-Western Bantu (pink) and Eastern Bantu (green). Geographic language distribution is based on the classification and location on Glottolog (https://glottolog.org).

Archaeological evidence for the spread of Bantu-speaking people is slightly more complicated (Mitchell and Lane, 2013). The spread of South-Western Bantu-speakers has been associated with the Naviundu tradition, a ceramic tradition that spread from West Africa directly southwards (Figure 4). The Urewe pottery tradition is associated with Eastern Bantu-speakers and can be found in the Great Lakes region. However, the origin of the Urewe tradition is unclear (Bostoen, 2018). From Lake Victoria in East Africa, two migrations southwards can be observed; one migration is associated with the Nkope tradition and follows an inland route to current-day Zambia, Zimbabwe and Malawi. The other migration followed the East African coast and is associated with the Kwale tradition. Moreover, the Kalundu tradition has been associated with a spread from West Africa across Zambia towards Zimbabwe and Mozambique (Huffman, 2007), but the validity of this pottery tradition has been questioned (Bostoen, 2018). The southeastern Bantu language speakers currently living in Zimbabwe, Mozambique and South Africa could be considered a composite of the Kalundu, Nkope and Kwale traditions.



(a) Early-Split model

(b) Late-split model

Figure 3. Two linguistic hypotheses for the spread of Bantu-speaking people. The first hypothesis, the so-called Early-split hypothesis, is shown on the left (a) and states that there was an initial split among Bantu-speakers in the Bantu homeland (indicated by the number 1), resulting in an Eastern Bantu speaker group and a Western Bantu speaker group. In this hypothesis, Eastern Bantu speakers spread eastwards north of the rainforest, and the Western Bantu speakers spread southwards through the rainforest. The other hypothesis, the Late-split hypothesis, is shown on the right (b) and proposes a common migration through the rainforest, followed by a split only after the rainforest.

Shum Laka is an archaeological rock shelter site located in the homeland of the Bantu Expansion and the archaeological findings are associated with Bantu material culture (Lavachery, 2001). This site from Cameroon has yielded genomes from four children, two from 8 000 BP and two from 3 000 BP (Lipson et al., 2020). These Shum Laka ancient individuals were genetically more similar to rainforest hunter-gatherers (RHG) currently living in western Central Africa, than to modern-day Bantu-speakers. This indicates that Bantu-speakers across the continent, as well as most of the people currently living in Cameroon, are not direct descendants of the population represented by these four Shum Laka individuals. However, this does not necessarily indicate that there were no Bantu-speaker ancestors living in the region at that time. It has been proposed that the ancestors of RHG and Bantu-speakers lived in the Shum Laka region at the same time (Lipson et al., 2020).

All of these different patterns pointed out by linguistic and archaeological hypotheses show that the Bantu Expansion was most likely a very complex expansion and migration, possibly comprised of multiple smaller waves of migrations. A genetics study confirmed that the Bantu Expansion followed a serial-founder migration model, with possible spread-over-spread events (Fortes-Lima et al., 2024). It has also been shown that the genetic diversity amongst

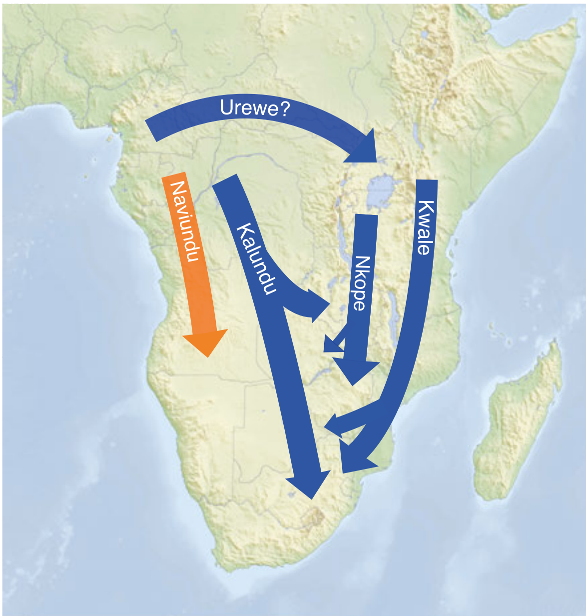


Figure 4. Archaeologically proposed spread of Bantu-speaking people and their associated pottery traditions. The spread of South-Western Bantu-speakers has been associated with the Naviundu tradition (orange). The spread of Bantu speakers towards eastern Africa is characterized by the Urewe pottery tradition, that later splits into two migrations southwards: the Nkope tradition (inland) and the Kwale tradition (coastal). Moreover, the Kalundu tradition has been associated with a spread from West Africa to southeastern Africa. The spread of Urewe, Kalundu, Nkope and Kwale traditions are depicted with blue arrows. Redrawn from Mitchell & Lane (Mitchell and Lane, 2013), with the question mark after the Urewe tradition referring to the uncertainty around the origin of the Urewe tradition (Bostoen, 2018).

Bantu-speaking populations declines with distance from Western Africa, where Zambia and DRC were crossroads of interaction.

1.2.3 Rainforest hunter-gatherers

After the first population divergence between Khoe-San people and all other modern humans, we can distinguish a second divergence event. This divergence has been dated to 202–135 kya (Breton et al., 2024b) and separates a lineage containing RHG from a lineage containing other modern humans (except the Khoe-San). There are more than 20 RHG groups, living across the Congo Basin in Central Africa. Based on their geographical location, they are grouped into eastern RHG (eRHG) and western RHG (wRHG). The estimates for the divergence time between eastern and western RHG range from $\sim$44 kya to $\sim$16 kya (Patin et al., 2009; Veeramah et al., 2012; Batini et al., 2010; Breton et al., 2024b). The wRHG live in the western part of the Congo Basin and

include the Baka, Bakola, Bakoya, Bezan, Biaka, and some Babongo groups. The eRHG are the Mbuti, Asua, Efe, and Twa. The Mbuti, Asua and Efe live around the Ituri rainforest in the eastern part of the DRC. The Twa have a slightly more southern distribution, in the direction of lake Victoria.
Prior to the migration of Bantu-speakers and East African pastoralists, there was a genetic cline of hunter-gatherer populations stretching from East to Southern Africa (Skoglund et al., 2017). Genomic analysis of BaTwa populations in Zambia reveals that they possess a hunter-gatherer-like genetic ancestry, comprising approximately 19-31% of their genetic ancestry (Breton et al., 2024a). This ancestry is intermediate between the Khoe-San ancestry found in present-day southern Africa and the central African rainforest hunter-gatherer ancestry, and most likely comes from local hunter-gatherer populations that inhabited Zambia in the past.

1.2.4 East African populations

East Africa holds a significant place in the study of human evolution, evidenced by its rich hominin fossil record. Among these fossils are Australopithecus afarensis (famously known as "Lucy," dating to 3.9–3.0 mya) (Johanson and White, 1979), the Omo I fossil (255–211 kya) (Vidal et al., 2022), and the Herto fossils (160–154 kya) (McCarthy and Lucas, 2014). These discoveries have led to the proposal that East Africa is one of the regions where modern humans evolved.

In addition to its role in human evolution, East Africa is a key region in the narrative of human migrations. It is both the region from where modern humans left Africa and a region that experienced significant back-migrations. Notably, the Mota individual, a 4 500-year-old individual found in Ethiopia, represents an instance of an East African genome without any evidence of back-migrations (Llorente et al., 2015; Lipson et al., 2022). This contrasts with genetic studies of modern-day Ethiopians, which reveal approximately 40–50% Levant ancestry (Pagani et al., 2012). Studies on modern and ancient DNA have shown that the formation of East African populations was a complex process, with multiple phases and admixture events (Prendergast et al., 2019; Hammarén et al., 2023).

This genetic complexity is reflected in the region's linguistic diversity. East Africa mainly hosts two large language families; Afro-Asiatic (AA) and Nilo-Saharan (NS). Pagani *et al.* have highlighted a strong link between genetics and linguistics in this region (Pagani et al., 2012). Specifically, they identify genetic clusters related to Omotic (AA), Nilotic (NS), and Semitic-Cushitic (AA) speakers.

1.2.5 The genetic influence of European colonists and the slave trade in South Africa

In previous sections, we have discussed several important ancestries that can be found in Southern Africa, including South Africa. South Africans show, apart from Khoe-San ancestry, Bantu ancestry and East African ancestry also ancestry components from Europeans and slaves that arrived from Madagascar, Asia and the African coasts. This is true for both the Afrikaners (Hollfelder et al., 2020) and the South African Coloured (SAC) population (Lankheet et al., 2024; Vicente et al., 2021; De Wit et al., 2010; Quintana-Murci et al., 2010; Petersen et al., 2013; Patterson et al., 2010).

The history of the Afrikaners, that speak Afrikaans, begins with the arrival of the Dutch East India Company (DEIC)(in old Dutch Vereenigde Oostindische Compagnie, or VOC) at the Cape of Good Hope in 1652. The Cape of Good Hope started as a small station to resupply ships and has evolved into what we know as Cape Town. The first immigrants were mainly men, and later some women as well. Six years after the arrival of settlers in the Cape of Good Hope, an independent settlement was established at the cape. In 1688, French Huguenots arrived. European ancestry among the Afikaners is high. It is estimated from genealogical records that the Dutch contributed 34–37%, Germans 27–34%, and French 13–26% (De Bruyn, 1976; Greeff, 2007; Heese, 1971). From the same records, non-European contributions have been estimated to be between 5.5 and 7.2%. Genetic research has shown that these estimations are not far off; 95.3% of the Afrikaners ancestry is from European populations, with more association to northwestern European populations (Hollfelder et al., 2020). The 4.7% of non-European ancestry is mainly from people who were brought from Asia and West Africa to the Cape as slaves (3.4%), whereas a much smaller proportion comes from Khoe-San people (1.3%).

At some point after the year 1700, the term "Cape coloureds" became established to refer to people of mixed ancestry. Nowadays, they are often referred to as the South African Coloured population. They are descendants of Khoe-San, Bantu-speaking populations, European settlers and slaves from Madagascar, Asia and the African coasts. The SAC shows sex-biased admixture patterns (Quintana-Murci et al., 2010; Vicente et al., 2021). Maternal ancestry is mostly from Khoe-San women, whereas the paternal ancestry is mostly from European men. The Coloured people today mostly live in the Western, Northern, and Eastern Cape provinces of South Africa.

1.3 Genetics and proteomics

1.3.1 General human genetics

Deoxyribonucleic acid (DNA) is the genetic material carried by all organisms (except some viruses). DNA is a polymer of nucleotides in the form of a double helix. There are four different nucleotides bases that make up the DNA: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The order in which these bases occur, is the genetic information that is being passed down from parent to offspring. Most of the human genome can be found in the nucleus, where the $\sim$6.4 x 10^9 base pairs (bp) are stored in the form of chromosomes. A small part of the human genome, the mitochondrial genome, can be found in organelles called mitochondria. Humans have 46 chromosomes, consisting of 23 pairs. The haploid size of the genome (the 23 "unique chromosomes") is $\sim$3.2 x 10^9 bp. Twenty-two of these chromosome pairs are known as the autosomes, and the last pair is the sex chromosomes. The composition of the sex chromosome pair determines the sex of an individual; men carry one X chromosome and one Y chromosome, whereas women carry two X chromosomes. As you get one chromosome of each pair from your father, and the other chromosome of that pair from your mother, females will always pass on an X chromosome to their offspring, and men pass on either an X chromosome or a Y chromosome, which will determine the sex of the offspring. Not all humans have the exact same bases at each of the 6.4 billion bp of the human genome. Changes can occur through mutations. A change in one of the bases is called a Single Nucleotide Polymorphism (SNP) and if these mutations occur in the germline, they can be passed on from parent to offspring.

1.3.2 The uniparental markers

The uniparental genetic markers, the mitochondrial genome and the non-recombining part of the Y chromosome (NRY), are two popular tools that have been employed to study genetic diversity in human populations. They can be used to obtain information about the maternal and paternal lineages respectively (Underhill and Kivisild, 2007). The fact that the uniparental genetic markers do not undergo recombination, allows for tracking of genetic mutations through lineages. These lineages that share the same mutations inherited from a common ancestor are called haplotypes. Haplotypes can be grouped into groups called haplogroups, that share the same clade-defining mutations.

Mitochondrial DNA

The mitochondrial genome (or mitochondrial DNA (mtDNA)) is the circular double stranded DNA located in the mitochondria, the organelles producing energy for the cell. The size of the mtDNA is very small compared to the complete genome, the mtDNA is only 16 569 base pairs. Each mtDNA molecule encodes 13 proteins, 22 transfer RNAs and two ribosomal RNAs (Anderson et al., 1981). Aside from some exceptional cases of paternal inheritance of the mitochondrial genome (Luo et al., 2018), mtDNA is inherited from the mother (Giles et al., 1980). The average rate of base-substitutions in the mtDNA is ten times higher than the rate in the nuclear DNA (Jobling et al., 2014). This is caused by the higher concentration of mutagenic, oxygen free radicals in the mitochondria, the longer period in which the mtDNA is single-stranded, the lack of histones, and the faster DNA replication caused by the high turnover rate of mitochondria. Because of this high mutation rate, a wide variety of mitochondrial sequences can be found among humans. That, as well as the small size of the mitochondrial genome and the fact that is almost exclusively maternally inherited, has made the mitochondrial genome a popular tool for investigating population structure, genetic diversity, and demographic events in the human population. Moreover, mitochondrial genomes offer the opportunity to investigate maternal population history and together with the data from the Y chromosome, it can be used to study sex-biased processes.

Mitochondrial DNA among humans is, as described before, structured into haplogroups. These haplogroups are collections of sequences that have been inherited from the same common ancestor (van Oven and Kayser, 2009). All modern humans have a matrilineal most recent common ancestor (MRCA) that lived about 210–120 kya in sub-Saharan Africa (Atkinson et al., 2008; Behar et al., 2008; Fu et al., 2013; Mishmar et al., 2003). African mitochondrial haplogroups can be subdivided into seven haplogroups (L0, L1, L2, L3, L4, L5, and L6), while all other, non-African mitochondrial haplogroups belong to haplogroup M, N and R, which are subgroups of haplogroup L3. A recent paper has described an eighth haplogroup; L7, a sister group to L5 (Maier et al., 2022). A schematic and simplified overview of the general structure of the human mitochondrial phylogeny is shown in Figure 5. The haplogroups are named using letters and numbers, adding a digit or letter every time a subdivision is made. The first split in the human mitochondrial tree separates L0 sequences from all other sequences (L1–L7) (Behar et al., 2008; Maier et al., 2022). The revised Cambridge Reference Sequence (rCRS) (GenBank sequence NC_012920) and the reconstructed sapiens reference sequence (RSRS) are often used as a reference when working with mitochondrial genomes (Andrews et al., 1999; Behar et al., 2012). Table 1.2 provides information on the age estimates of the different African haplogroups. It has been shown that when different secondary mitochondrial DNA groupings are applied, inconsistent results are found in frequency-based analyses (Bajić et al., 2023).

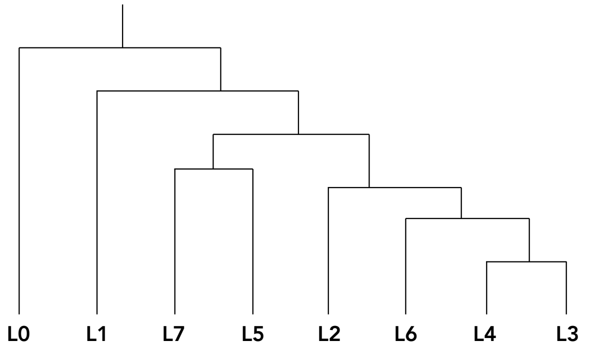


Figure 5. Schematic and simplified overview of the general structure of the human mitochondrial phylogeny.

Khoe-San associated haplogroups

The various mitochondrial haplogroups from Africa have distinct geographical distributions. L0d occurs mostly in South Africa, Namibia and Botswana (Barbieri et al., 2013; Schlebusch et al., 2013), whereas L0k generally has a more northern distribution, compared to L0d, and is mostly found in Namibia, Angola, Botswana (Barbieri et al., 2013) and Zambia (Schlebusch et al., 2013). Mitochondrial haplogroups L0d and L0k show high prevalence in Khoe-San groups and in the SAC population. Ninety-six percent of the Ju|'hoansi (Vigilant et al., 1991), 51–65% of the !Xun, and 11–16% of the Khwe (Chen et al., 2000; Schlebusch et al., 2013) carry haplogroup L0d. L0d1b, L0d2a and L0d3 have higher frequencies in the south of Southern Africa, whereas L0d1a and L0d2c show a more central distribution (Schlebusch et al., 2013). L0d2a shows a very strong recent expansion (Schlebusch et al., 2013). The distribution of L0k is more restricted, mainly to the Kalahari desert region (Barbieri et al., 2013). Most of the L0k sequences belong to L0k1, which shows higher frequency in the north of Southern Africa (Zambia and northern Botswana) (Schlebusch et al., 2013). The frequencies of L0k in the Ju|'hoansi are rather low (4%) (Vigilant et al., 1991), but intermediate in the !Xun and the Khwe (26–33% and 22% respectively)(Barbieri et al., 2013; Chen et al., 2000). Barbieri *et al.* found that, in contrast to other L0d and L0k haplogroups, two ancient sublineages (L0k1b and L0k2) were present almost exclusively in Bantu-speaking populations from Zambia. They suggest that these lineages might be remnants of what used to be a richer diversity of these two haplogroups in Southern Africa.

Mitochondrial haplogroup L0d and L0k split around 145 kya (Barbieri et al., 2013). All variation within L0d coalesces 102–48 kya (Behar et al., 2008, 2012), and all L0k lineages coalesce around 40–36 kya (Barbieri et al., 2013; Behar et al., 2008, 2012; Heinz et al., 2017; Rito et al., 2013). All L0d1 haplogroups show signs of recent expansion (within the last 30 000 years), where L0d1a and L0d1b had stronger expansion signals than the L0d1c haplogroup (which is associated to northern San) (Schlebusch et al., 2013). L0k also shows an expansion of effective population size starting around 5 000 years ago (Barbieri et al., 2013).

Bantu-speaker associated mitochondrial haplogroups
The expansion of Bantu-speaking people across sub-Saharan Africa, as discussed in section 1.2.2, is associated with certain mitochondrial haplogroups: L0a (Bandelt et al., 1995; Chen et al., 1995; Pereira et al., 2001), specific clades of L1c (Beleza et al., 2005; Rando et al., 1998), L2a (Pereira et al., 2001; Salas et al., 2002), L3b (Watson et al., 1997), and L3e (Alves-Silva et al., 2000; Bandelt et al., 2001). L0a is very common and wide-spread on the African continent (Rito et al., 2013). It originated most likely in Eastern Africa around 42 kya, but the highest frequencies of L0a are currently found in Mozambique and Zimbabwe (around 25%). Most likely, the eastern branch of Bantu-speakers picked up L0a in eastern Africa and brought it to Southern Africa (Salas et al., 2002). It is now common in Central, Eastern, and Southeastern Africa, but it is nearly absent in Western, Northern, and Southern Africa. L1c is most prevalent in Central Africa (Cameroon and Gabon), and more specifically, it has been proposed that L1c1b, L1c1c and L1c2 originated in Bantu ancestors and these haplogroups show signs of recent expansion (Batini et al., 2007). Other L1c haplogroups have been associated with western RHG (L1c1a, L1c4 and L1c5) (Batini et al., 2007; Quintana-Murci et al., 2008; Gonder et al., 2006). The age of the L1c clade has been estimated to 103 kya (Quintana-Murci et al., 2008). Mitochondrial haplogroup L2a is the most frequent haplogroup in Africa (Rosa and Brehm, 2011) and is prevalent in many parts of the African continent, but is specifically abundant in Central Africa (Cerezo et al., 2011; Rosa and Brehm, 2011). Coalescence estimates for L2a lie between 98 and 46 kya (Behar et al., 2008; Heinz et al., 2017; Salas et al., 2002). Mitochondrial haplogroup L3 encompasses both African-specific haplogroups, as well as non-African haplogroups. The age of African L3 haplogroups is around 70 kya. L3e is the most widespread L3 haplogroup, and a subclade L3e1 is common among southeastern Bantu speakers (Soares et al., 2012). L3b and L3d are found mainly in Western Africa.

Other mitochondrial haplogroups
Apart from the Khoe-San and Bantu-speaker-associated mitochondrial haplogroups, the African continent harbours more mitochondrial haplogroups. Among these haplogroups are lineages associated to Eastern Africa, as well as lineages that originated outside Africa, but are currently mainly found in Africa.

Several L haplogroup lineages occur most frequently in Eastern Africa. Examples of these haplogroups are L0a, L0f, and L4 to L7 (Gonder et al., 2006; Maier et al., 2022; Rosa and Brehm, 2011). As discussed in the previous section, L0a originated in eastern Africa and was brought to large parts of sub-Saharan Africa by the Bantu Expansion. L0f is a mitochondrial haplogroup that occurs at low frequencies, and only in Eastern Africa (Salas et al., 2002). Haplogroup L4 is common in Kenya and Tanzania, where it reaches frequencies of 60% in the Hadza population and 48% in the Sandawe population (Tishkoff et al., 2007a). Haplogroup L5 reaches a maximum frequency of 9.3% in the Mbuti population of the Democratic Republic of the Congo (DRC) (Maier et al., 2022). L6 occurs mostly in the Horn of Africa; in Ethiopia and Somalia (Fernandes et al., 2015). A recent paper expanded the number of basal lineages of the mitochondrial tree from seven to eight and described haplogroup L7. This haplogroup is rather rare, as it is found at a maximum frequency of 1.4% in the Sandawe population of Tanzania (Maier et al., 2022).

Even though mitochondrial haplogroups M1 and U6 originated outside the African continent, they are generally considered African haplogroups, as they are mainly found in Africa and were reintroduced by back-migrations (Maca-Meyer et al., 2003; Pennarun et al., 2012; González et al., 2007). M1 is mainly found in northeast Africa, and U6 in northwest Africa (Rosa and Brehm, 2011).

Table 1.2. *Mitochondrial haplogroups and their estimated ages (years BP). Estimates are based on the following publications: (Behar et al., 2012; Heinz et al., 2017; Rito et al., 2013; Salas et al., 2002; Soares et al., 2009, 2012; Maier et al., 2022).*

Mitochondrial haplogroup	Age estimation (kya)	(Behar et al., 2012)	(Soares et al., 2009, 2012)	(Heinz et al., 2017)	(Rito et al., 2013)	(Salas et al., 2002)	(Behar et al., 2008)	(Maier et al., 2022)
L0	128-169	136289	149700	168500	128200	-	153971	-
L0a	40.3-63.3	46741	44800	63300	42400	40350	53176	-
L0d	48.4-102	48422	-	-	-	49600	101589	-
L0f	68.9-88.0	68988	-	-	69300	-	88097	-
L0k	11.2-41.7	36763	11200	41700	35700	-	39683	-
L1	124-142	128521	140600	142000	123800	-	-	-
L1b	4.3-39.2	4334	9700	39200	31000	30550	29366	-
L1c	59.6-102	77932	85400	90000	81900	59650	102383	-
L2	70.1-119	88453	89300	119000	82800	70100	104764	-
L2a	46.0-97.9	79305	48300	97900	-	55150	46033	-
L2b	14.3-31.6	21259	14300	30400	-	31600	26985	-
L2c	17.7-27.5	17762	25200	22800	-	27500	23810	-
L2d	17.0-122	17002	-	-	-	121900	23810	-
L2e	46.8	-	-	-	-	-	46826	-
L3	61.3-76.1	67262	71600	70800	-	61300	76192	-
L3a	49.6	49689	-	-	-	-	-	-
L3b	16.4-28.5	21538	16400	25700	-	21600	28572	-
L3c	9.2	9243	-	-	-	-	-	-
L3d	25.7-34.0	25739	31000	34000	-	-	-	-
L3e	36.1-44.3	36189	39000	44300	-	-	-	-
L3f	46.3-56.2	46305	53200	56200	-	-	-	-
L3h	59.4-67.2	59422	66700	67200	-	-	-	-
L3i	34.8-51.1	41231	34800	51100	-	-	-	-
L3k	23.9	23983	-	-	-	-	-	-
L3x	36.1-39.1	36750	36100	39100	-	-	-	-
L4	79.0	78996	-	-	-	-	-	-
L5	90.6-151	110805	120200	151000	109400	-	138098	90636
L6	22.2-22.3	22331	-	-	-	-	22223	-
L7	91.4	-	-	-	-	-	-	91476

The non-recombining part of the Y chromosome

As described before, females have two X chromosomes and males have one X chromosome and one Y chromosome. Only small regions at the end of the X and Y chromosomes are homologous and most of the Y chromosome does not recombine, i.e. they do not undergo the process in which pieces of DNA are exchanged and reshuffled to produce new combinations of genetic variants. This part of the Y chromosome that does not undergo recombination, is called the non-recombining part of the Y chromosome, or NRY for short. Sometimes it is also referred to as the male specific portion of the Y chromosome (MSY). All the variants on the NRY are thus inherited together as an intact haplotype. The only changes that occur arise through mutation and these can be traced backwards in time. As the Y chromosome is passed on from father to son over generations, the NRY can be used to study paternal ancestry of an individual or a population.

Figure 6 shows a schematic and simplified overview of the general structure of the human Y chromosome phylogeny. The classification of Y haplogroups follows two different systems (Y Chromosome, 2002). The first system uses a combination of letters and numbers, such as A1b1a1. The second system identifies the lineage by the specific mutation, for example, A-M14, where M14 indicates a particular mutation on the Y chromosome that defines the clade. While the first approach is currently in use, older classifications often rely on the second method. The International Society of Genetic Genealogy (ISOGG) keeps an updated version of the Y phylogeny (https://isogg.org/tree/index.html). Y haplogroups A, B and E are most frequent in Africa, where the other Y haplogroups are mainly occurring outside Africa. Y chromosomal haplogroups A and B are deeply rooted branches of the Y chromosome phylogeny. Haplogroup A is found in the Khoe-San people of Southern Africa, as well as in Eastern Africa (Naidoo et al., 2010; Jobling et al., 2014). Haplogroup A00, a paragroup to haplogroup A, is the most deeply rooted Y haplogroup in humans and is found in Western Africa (Mendez et al., 2013). The finding of 11 A00 sequences in the Mbo population of western Cameroon re-rooted the Y chromosome tree and estimated the time to the most recent common ancestor (TMRCA) for the Y chromosome tree to be 338 kya (Mendez et al., 2013). Notably, A00 has also been found in one ancient individual from Shum Laka, Cameroon, that was dated to 8 000 BP (Lipson et al., 2020).

Haplogroup A1b1b2a (A-M51) is the most common haplogroup A subclade in Southern Africa and is found mostly among the Khoe-San people, or in populations with significant levels of Khoe-San admixture (Barbieri et al., 2016). Y haplogroup A1b1a1 (A-M14) has been found in both Khoe-San populations, as well as in central African rainforest hunter-gatherers (Naidoo et al., 2020; Wood et al., 2005). Y haplogroup B2a (B-M150) shows frequencies in both Khoe-San populations, as well as Bantu-speaking populations, thereby making it unclear if this haplogroup is autochthonous to the Khoe-San population, or that it was brought to Southern Africa by Bantu-speakers (Barbieri et al., 2016). Haplogroup B2b (B-M112) has been linked to ancient gene flow from East to Southern Africa (Naidoo et al., 2020). A subgroup of Y haplogroup E1b1b1 (E-M35), known as E1b1b1b2b2a1 (E-M293), has been associated with the migration of pastoralist groups from East Africa to Southern Africa approximately 2 000 years ago (Naidoo et al., 2020; Henn et al., 2008), though some studies have not confirmed this connection (Bajić et al., 2018). Additionally, haplogroups E1b1a1 (E-M2), E2 (E-M75), and B2a (B-M150) are thought to be related to the spread of Bantu-speaking peoples (Beleza et al., 2005; Berniell-Lee et al., 2009; Naidoo et al., 2010).

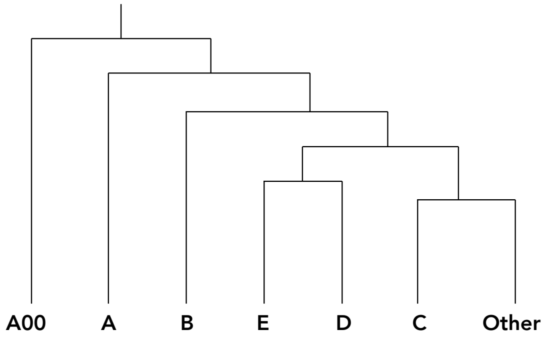


Figure 6. Schematic and simplified overview of the general structure of the human Y chromosome phylogeny.

1.3.3 Palaeoproteomics

Although the field of genetics and aDNA has significantly enhanced our knowledge of human history and human evolution, it also has its limitations. One of the largest limitations is the survival and preservation of aDNA. The oldest aDNA findings recovered from direct animal remains are from a mammoth that lived 1.65 mya in Siberia and was preserved in the permafrost (van der Valk et al., 2021). However, finding DNA from very old specimens is quite the exception. In places with a warm and more humid climate, aDNA preservation is much less optimal, as DNA degrades faster under hot and moist conditions (Allentoft et al., 2012; Lindahl, 1993; Smith et al., 2003). As an illustration, the oldest, directly dated human remains that have yielded aDNA from the African continent are found in Taforalt, Eastern Morocco, and are 15 000 years old (van de Loosdrecht et al., 2018). Although 15 000 years ago is a long time and these specimens provide valuable information, aDNA from the African continent remains rare (Vicente and Schlebusch, 2020; Ávila-Arcos et al., 2023).

Fortunately, DNA is not the only biomolecule that can help us provide insight into the human past. Proteins also contain precious molecular biological information and they generally preserve longer than DNA. This is exhibited by the oldest remains that have yielded protein results; ostrich eggshells dated to 3.8 mya (Demarchi et al., 2016). These remains are not only more than twice as old as the remains providing the oldest aDNA, they are also from a warm and humid climate, as they have been found in northern Tanzania. This provides perspective on the information that could potentially be obtained from human remains from deeper time periods and from climatically warmer regions such as Africa, compared to DNA sequencing.

Proteins are biomolecules built up of amino acids. The name amino acid comes from its structure; they contain both an amino (-NH2) and a carboxylic acid (-COOH) functional group. The side-chain group determines the unique properties of each amino acid. In nature, there are about 500 amino acids, but only 20 are found in the human body (see Table 1.3). The order of nucleotides in a gene determines the order of amino acids in a protein, as well as the length of a protein. The combination of three nucleotides corresponds to one amino acid (see Table 1.4). Through a process called transcription, messenger RNA (mRNA) is created with DNA as a template, and this serves as a blueprint for translation, the actual synthesis of proteins. During translation, amino acids are coupled through a peptide bond, which is the connection of an amino group of one amino acid to the carboxylic acid group of another. The resulting protein fulfills a particular function in or outside the cell. Examples of these functions are providing structure to a cell or extracellular matrix, enzymatic functions, and signalling functions. After translation, amino acids can undergo chemical changes, which are termed post-translational modifications (PTMs). Phosphorylation (the addition of a phosphate), methylation (the addition of a methyl group), and acetylation (the addition of an acetyl group to lysine or to the amino-end of a protein) are some of the well-known PTMs. The complete set of proteins expressed by an organism is termed a proteome and the study of the proteome is termed proteomics.

Table 1.3. *List of the twenty amino acids that are encoded by the human genome. The full name, as well as a three letter code and a single letter code is listed for each amino acid.*

Amino acid	Three letter code	Single letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

As described above, proteins contain valuable biomolecular information and investigating ancient proteomes can provide insight into ancient diets (Bleasdale et al., 2021; Solazzo et al., 2008), ancient clothing (Hollemeyer et al., 2008), ancient pathogens (Warinner et al., 2014), and it can facilitate species identification (Buckley et al., 2014; Coutu et al., 2021; Le Meillour et al., 2020). Moreover, it has been shown that ancient proteins can be informative for the distinction between mammalian species (Welker et al., 2015), and also between hominins (a term used for humans and all of our extinct bipedal ancestors) (Chen et al., 2019; Welker et al., 2016, 2020). As the genetic code is redundant, meaning certain amino acids are encoded by multiple codons (Table 1.4), many SNPs do not result in amino acid changes, or single amino acid polymorphisms (SAPs). We differentiate between synonymous SNPs, SNPs that do not cause a change in amino acid, and nonsynonymous SNPs, SNPs that cause an alteration in the amino acid sequence. Due to synonymous SNPs, and due to purifying selection preventing mutations to accumulate in protein-coding regions, the rate of accumulation of SAPs during evolution is

Table 1.4. *Codon table showing the relationship between the genetic code and the amino acids. Each three-nucleotide combination (codon) corresponds to an amino acid or start- or stop-codon. The start-codon (or initiation codon) and the stop-codons (or chain-terminating codon) are indicated. There are several codons that encode certain amino acids, for example Leucine (Leu), a phenomenon that is termed redundancy of the genetic code.*

First Position (5' end)	**Second Position**	**Third Position**			
		U	**C**	**A**	**G**
U	**U**	UUU Phe	UUC Phe	UUA Leu	UUG Leu
	C	UCU Ser	UCC Ser	UCA Ser	UCG Ser
	A	UAU Tyr	UAC Tyr	UAA **Stop**	UAG **Stop**
	G	UGU Cys	UGC Cys	UGA **Stop**	UGG Trp
C	**U**	CUU Leu	CUC Leu	CUA Leu	CUG Leu
	C	CCU Pro	CCC Pro	CCA Pro	CCG Pro
	A	CAU His	CAC His	CAA Gln	CAG Gln
	G	CGU Arg	CGC Arg	CGA Arg	CGG Arg
A	**U**	AUU Ile	AUC Ile	AUA Ile	AUG **Met**
	C	ACU Thr	ACC Thr	ACA Thr	ACG Thr
	A	AAU Asn	AAC Asn	AAA Lys	AAG Lys
	G	AGU Ser	AGC Ser	AGA Arg	AGG Arg
G	**U**	GUU Val	GUC Val	GUA Val	GUG Val
	C	GCU Ala	GCC Ala	GCA Ala	GCG Ala
	A	GAU Asp	GAC Asp	GAA Glu	GAG Glu
	G	GGU Gly	GGC Gly	GGA Gly	GGG Gly

much lower than the rate of SNP accumulation. Moreover, proteins are expressed in a tissue-specific manner, while whole genomes are present in all tissues. Thus only those proteins expressed in bone and teeth can be investigated in ancient specimens. In spite of that, ancient proteomes carry enough resolution to phylogenetically differentiate between ancient hominins (Welker et al., 2016).

The field that investigates ancient proteomes is termed palaeoproteomics. An important aspect when working with either ancient DNA or ancient proteins is contamination of the specimen. Contamination of the specimen with modern proteins can occur at many times during the investigation process; during the excavation, the storage, and the analysis. Measures have to be taken during all these steps to minimize the contamination with modern proteins. One of the measures taken is to work in a clean lab. Moreover, during the bioinformatics analysis, there are ways to identify contiminating proteins. This concept is based on the degradation of proteins, which will be described further in Section 2.4.2.

2. Methods

2.1 Genomic data generation techniques

In order to work with genetic data, data generation is required. Various types of methods for generating genetic data exist and what method you choose is based on your research questions and your budget. If one is interested in the bases at certain predetermined positions of the genome, genotyping your DNA of interest on a SNP array (sometimes referred to as SNP-chips) is a good choice. However, if there is a broader interest for the whole genome, whole-genome sequencing (WGS) is a better option. Long-read sequencing is the last sequencing technique that will be discussed here. It is sometimes referred to as third generation sequencing, as it arose after the short-read sequencing techniques of the next-generation sequencing wave. Long-read sequencing generates much longer reads (>10 000 bp) than conventional short read sequencing techniques (75–300 bp). These techniques will be described further in the following sections.

2.1.1 Genotyping

A SNP array is a powerful and affordable approach for genotyping large numbers of predetermined SNPs across the genome. To detect these SNPs on the array, short DNA probes specific to each SNP of interest are attached onto the array surface. Genomic DNA from the samples is fragmented, labeled, and then hybridized to these probes. The probes and DNA fragments hybridize based on complementary sequences, enabling the array to identify the SNPs present in the sample. A high-throughput array scanner reads the fluorescence signals from the hybridized probes, determining the genotype at each SNP. These SNPs are selected because they exhibit variation between different populations. Since SNP arrays are often designed using data from specific populations, they tend to be most effective for those populations, which can introduce bias when applied to other groups. There is a wide variety of SNP arrays available, each containing a unique set of SNPs. Thus, the choice of a specific SNP array should be tailored to the population under study. For example, many SNP arrays are based on the variation in Europeans. However, genotyping African populations on these SNP arrays creates a bias. The Infinium H3Africa Consortium Array v2 from Illumina has been designed for the study of African populations and has included 2.2 million SNPs, of which 2528 are located on the Y chromosome and 234 on the mitochondrial genome

(Mulder et al., 2018). Moreover, SNP arrays are designed to handle multiple samples simultaneously. Depending on the number of SNPs being investigated, one plate can accommodate 96, 48, or 24 individuals, allowing for efficient high-throughput analysis of many samples at once.

2.1.2 Whole-genome sequencing

Whole-genome sequencing is the sequencing of the entire genome of an organism. A commonly used technique is next-generation sequencing. In short, DNA is fragmented and adapters are added during DNA library preparation. During the library preparation, unique indices can also be added to the primers for each library to facilitate simultaneous sequencing of multiple samples (Meyer and Kircher, 2010). The DNA library is then loaded onto a flowcell, that contains oligonucleotides that are complementary to the adapters. This allows the fragments to bind to the flowcell. Subsequently, fluorescent nucleotides (each base carries a different fluorescence) are added to the flowcell and their incorporation is measured. This sequencing approach that relies on polymerases to incoporate bases while the signal is being read, is referred to as sequencing-by-synthesis (SBS). There are also strategies to perform targeted sequencing, in which you first enrich for genomic regions of interest, after which the previously mentioned next-generation sequencing techniques are applied. For targeted sequencing, it is for example common to investigate the protein-coding regions of the genome (the exome), or certain chromosomes only, for example the Y chromosome.

2.1.3 Long-read sequencing of mitochondrial genomes and haplogroup assignment

By sequencing the complete mitochondrial DNA or sections of it, it is possible to answer questions about ancient maternal population structure in Africa. In the past, studies have focused on one or more of the hypervariable regions of the mitochondrial genome, sometimes also investigating some additional SNPs (Černý et al., 2007; Schlebusch et al., 2013, 2011; Tishkoff et al., 2007b; Pereira et al., 2010). Since 1977, it has been possible to sequence mitochondrial genomes (Sanger et al., 1977) and the methods to do this have constantly been becoming less complex and faster. In 1981, the complete sequence of the human mitochondrial genome was published (Anderson et al., 1981). The first full mitochondrial genomes were amplified using a PCR of multiple overlapping fragments, which was recently reduced to an amplification in two fragments (Vossen and Buermans, 2017). Working with these large fragments of at least 7 kilo base pairs (kb) requires longer reads than conventional Sanger sequencing and next-generation sequencing techniques deal with. As described before, and also suggested by its name, long-read sequencing produces long reads. Typically, the reads are over 10 000 bp long. There are several companies producing long-read sequencers such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT or Nanopore) (Logsdon et al., 2020). PacBio works with a principle called Single Molecule, Real-Time (SMRT) sequencing. SMRTbell® libraries are created by ligating adaptors to double-stranded DNA. This creates a circular, single-stranded DNA molecule. These molecules are then loaded onto a SMRT® cell, that contains many small wells called zero-mode waveguides (ZMWs). In every ZMW, one circular DNA molecule is immobilized. The DNA sequence is then read by measuring the fluorescence signals of the different nucleotides being incorporated, in a similar way to how it is performed for other next-generation sequencing techniques. One addition is that the same circular sequence is read multiple times, to provide more accuracy. This method is called circular consensus sequencing (CCS), and it creates very accurate long reads, known as high-fidelity (HiFi) reads. Nanopore sequencing is based on a slightly different principle. The Nanopore flow cell contains a synthetic membrane with small pores embedded in it. The DNA molecules are unwound and an electric current is applied to pass the single-stranded DNA molecule through the pore. As the DNA molecule passes through the pore, the electric current flowing through the nanopore changes, from which the DNA sequence can be deciphered.

PacBio sequencing has proven to be very useful in deciphering full mitochondrial genomes. HiFi read lengths for PacBio are typically between 10 and 25 kb, with accuracies higher than 99.5% (Hon et al., 2020). As opposed to variant calling for short reads, variant calling for long reads is not as standardized. Variant calling from the HiFi reads can be done with GATK HaplotypeCaller (Van der Auwera and O'Connor, 2020) or DeepVariant (Poplin et al., 2018). The obtained mitochondrial sequences can be used to determine mitochondrial haplogroups. Automated mitochondrial haplogroup assignment can be performed with HaploGrep (Weissensteiner et al., 2016), either as an online graphical interface, or on the command-line. Sequences must be aligned to the rCRS or the RSRS. Alignment can be performed with MUSCLE (Edgar, 2004). Input file formats for running HaploGrep2 are FASTA, VCF or hsd.

2.2 Inferring population structure and gene flow

Understanding the genetic landscape of human populations involves exploring the complex patterns of population structure and gene flow. In the field of genetics, population structure looks at how genetic differences are spread and arranged within and between populations. In humans, this structure is shaped by various factors, such as geographical and cultural barriers. Population structure is influenced by migration and admixture events. Admixture events involve the mixing of genetic material when individuals from different ancestral populations meet and produce offspring. Studying genetic variation within and between populations can provide insights into historical demographic processes. Modern genetic research often relies on extensive datasets, featuring millions of SNPs. To interpret these highly complex datasets, various methods can provide useful insights into genetic structure. In the following sections, I will discuss some of these methods.

2.2.1 Principal component analysis

Principal Component Analysis (PCA) is a statistical technique commonly used in human population genetics to analyze and visualize genetic variation within and between populations. This method simplifies complex genetic data, such as SNP array data, making it easier to identify patterns and relationships among individuals based on their genetic makeup. PCA reduces the complexity of data from numerous genetic variants by transforming it into a smaller set of variables known as principal components. These components capture the most significant variation in the data, allowing researchers to see how individuals and populations cluster genetically. Typically, two or three principal components are plotted against each other. In a PCA plot individuals who are genetically closer tend to cluster near each other in the principal component space, whereas those who are genetically more distant appear further apart. This plot can give clues about the genetic structure of individuals and populations, providing valuable insights into their ancestral origins. In Figure 7, a PCA plot shows individuals from various geographic regions. The South African Coloured individuals, represented in blue, are positioned between individuals from different geographic areas. This intermediate position reflects their mixed ancestry, showcasing how PCA can effectively give a first impression of genetic backgrounds.

2.2.2 Partial least squares discriminant analysis

Partial Least-Squares Discriminant Analysis (PLS-DA) is, like PCA, a dimensionality reduction technique (Barker and Rayens, 2003). Where PCA reduces dimensionality without regard to class labels, PLS-DA aims to find a set of latent variables that maximize the separation between predefined classes in the data. As input, it requires two matrices: a predictor matrix (X) containing feature data and a response matrix (Y) representing class labels, with X being any type of numeric data and Y typically being categorical or binary. PLS-DA is often thought of as a "supervised" version of PCA, as it performs dimensionality reduction, but takes into consideration the class labels. As a machine learning tool, PLS-DA is used for classification and predictive modeling. It combines features of Partial Least Squares (PLS) regression and Discriminant Analysis to handle complex data with many variables. In essence, PLS-DA works by finding a set of latent variables that best discriminate be-

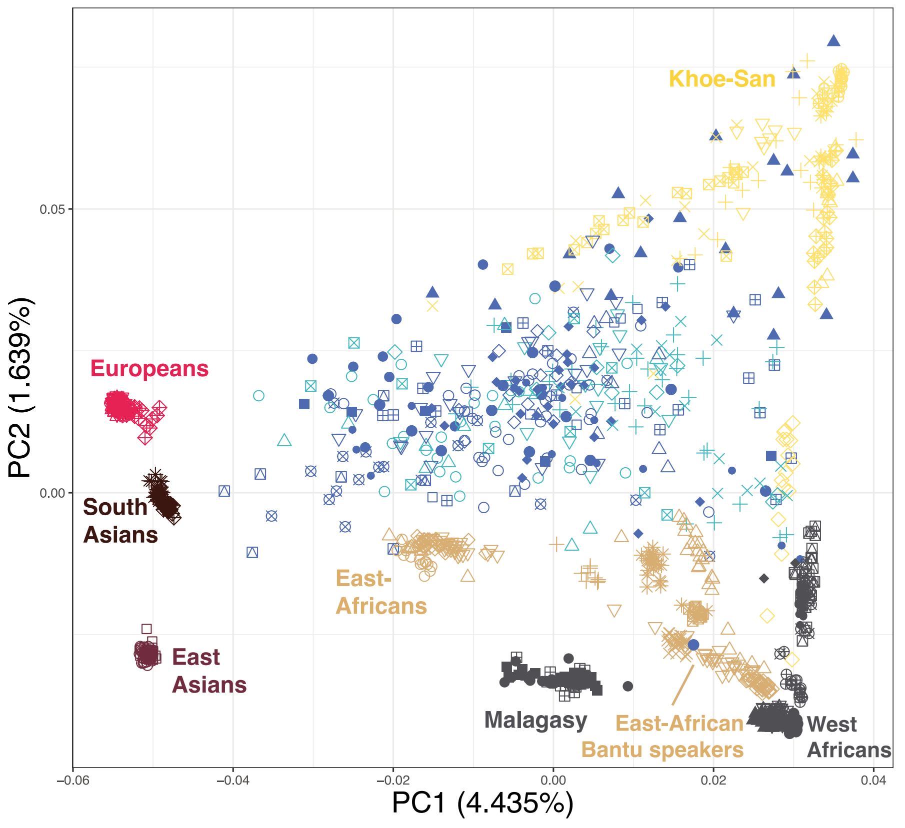


Figure 7. Principal component analysis showing individuals from Africa, Europe and Asia. The first two components are shown, and the amount of variation explained by each component is indicated in parentheses. Data from **Paper V**. The individuals are coloured based on the geographic region they are from. In dark and light blue, the South African Coloured (SAC) individuals. SAC individuals have genetic ancestry from Europeans, the Southern African Khoe-San, South- and East-Asians, West-Africans and East-Africans. Due to their mixed heritage, SAC individuals are situated between these populations in the PCA plot, reflecting their diverse genetic background.

tween classes in a dataset. These latent variables are linear combinations of the original predictors, which helps to reduce the dimensionality. By projecting the data onto these latent variables, PLS-DA maximizes the separation between classes, making it easier to build predictive models. It provides a robust approach for classification tasks, especially when traditional methods might struggle with the complexity of the data. In Figure 8, we see a PLS-DA plot based on Single Amino Acid Polymorphisms in bone-related proteins, where the separation between Khoe-San individuals and non-Khoe-San individuals is maximized.

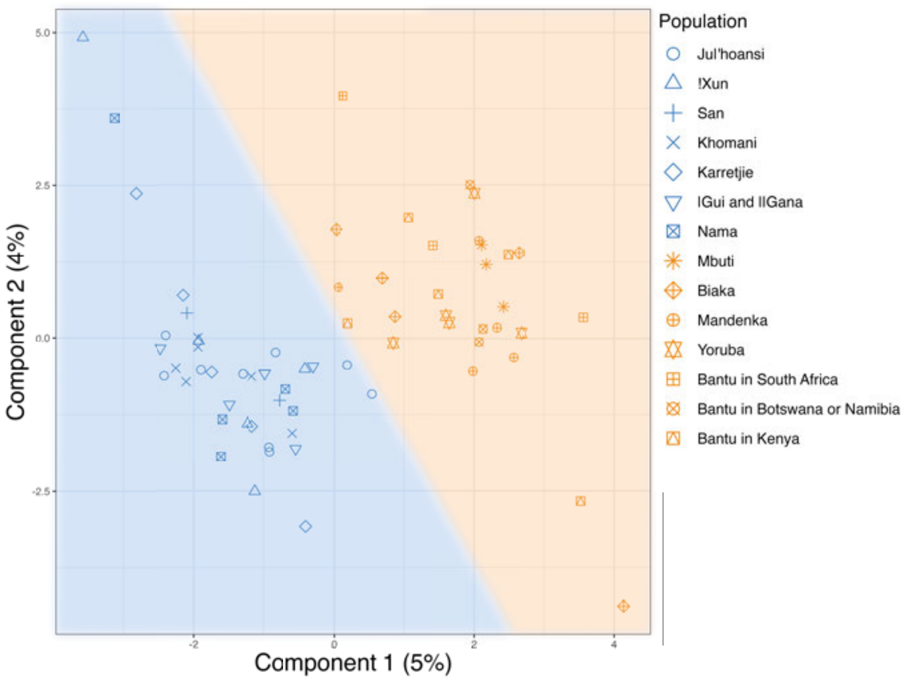


Figure 8. Partial least squares discriminant analysis (PLS-DA) on the classes Khoe-San individuals and non-Khoe-San individuals, based on 88 Single Amino Acid Polymorphisms in bone- and teeth-related proteins. Khoe-San individuals are shown in blue, and non-Khoe-San individuals in orange. The blue space in the PLS-DA plot indicates the area of the plot where the Khoe-San individuals cluster, the orange part indicates where the non-Khoe-San individuals cluster. The data from **Paper IV** was used.

2.2.3 Model-based estimation of ancestries

Model-based clustering methods like STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander and Lange, 2011) offer an alternative way to gather information about the relationships between individuals and populations. Both methods provide estimates of the genetic composition of individuals using likelihood models. Likelihood models calculate the probability of observing the data given various parameter values of the model. ADMIXTURE came out more recently and runs considerably faster. The model assumes a certain number of clusters, or K's, to which individuals' ancestries are assigned. Pong is often used to visualize the results (Behr et al., 2016). It is important to note that these methods explain the variation between individuals, and should be used as exploratory analyses, not for inferring divergence and admixture events (Lawson et al., 2018). Figure 9 shows the ADMIXTURE results for individuals from Africa, Europe and Asia, visualized using Pong for K=6. Each vertical bar, representing an individual, is divided into colored segments, which represent the estimated proportion of ancestry that individual has from each genetic cluster. The individuals from the same geographic region are grouped together posteriorly, making it clear that the major ancestries correspond to the continental regions and specific African populations. The South African Coloured individuals, shown on the left side of the figure, show ancestry proportions related to Khoe-San, West-African groups, Europeans, and people from other regions in Africa and Asia.

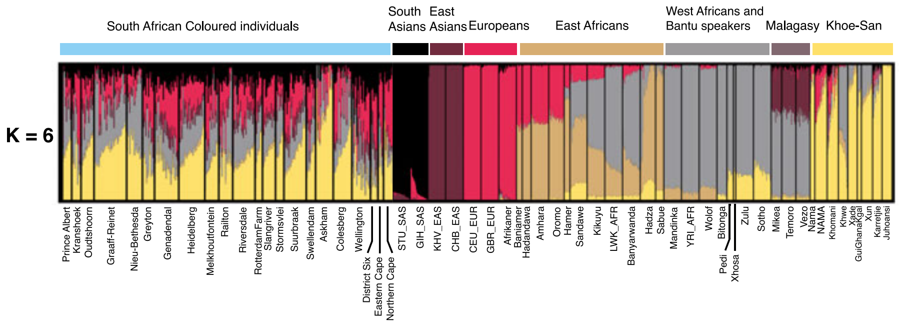


Figure 9. ADMIXTURE results for populations from Africa, Europe and Asia, visualized using pong for K = 6. Data from **Paper V** was used. Each vertical bar represents an individual, the colours represent the different genetic clusters.

2.2.4 Formal tests of admixture

F-statistics are a set of statistical tools used primarily in population genetics to analyze genetic relationships between populations and test hypotheses about their evolutionary history (Peter, 2016). The amount of populations involved in the statistical test is indicated by their name; there is f_2, f_3, and f_4, which use two, three, and four populations respectively. f_2 measures genetic drift between two populations, and higher values indicate more genetic differentiation, implying more drift or less recent common ancestry. f_3 has two variations; outgroup f_3 and admixture f_3. Outgroup f_3 measures the shared genetic drift between two test populations relative to an outgroup, and higher values indicate more genetic affinity between the two test populations. Admixture f_3 tests for admixture in a target population (X) from two source populations (A and B). If the admixture f_3 value is negative, this is evidence for admixture between populations A and B contributing to population X. A positive admixture f_3 value, however, does not prove the absence of an admixture event. The f_4 statistic can be used to test whether one population X is more admixed with population A, or with population B. The fourth population is an outgroup. If the f_4 value is zero, no admixture occurred between any of the populations. Assuming the tree topology is correct, significant deviations from zero in the f_4 statistic can only be attributed to admixture. Whether the f4 value is negative or positive offers insight into the potential pairwise combinations involved in the admixture event.

2.2.5 Local ancestry estimation and admixture dating

When two populations meet and produce offspring, their genetic material mixes, creating individuals with a combination of genetic backgrounds. This process is referred to as admixture. Chromosomes in the first generation offspring are a complete copy of the chromosome from either parent. This non-random association of alleles at different loci is known as linkage disequilibrium (LD). However, over generations, the mixed genomes undergo recombination, which is the reshuffling of genetic material between two homologous chromosomes. The more generations have passed since the initial admixture, the more recombination has happened and the shorter the size of the fragments with the same ancestry. In other words, LD decreases. When we investigate the ancestries of two interbreeding populations in an admixed population, we can infer which part of the genome came from which parent population, which is called local ancestry estimation. The size of these ancestry segments can provide valuable insights into the timing of admixture events. The longer ago an admixture event happened, the shorter the fragments. This information is used to estimate the timing of the admixture. In **Paper V**, I use the program MOSAIC (Salter-Townshend and Myers, 2019) to infer which genomic fragments in the South African Coloured population came from which source population, and eventually determine the timing of admixture between the different source populations.

2.2.6 Sex-biased demographic patterns

In section 2.2.5, the process known as admixture was introduced. In the special case of sex-biased admixture, the source populations of an admixture event contribute different proportions of males and females to the resulting population (Goldberg and Rosenberg, 2015). Reasons for sex-biased admixture are varied and can reflect certain cultural practices, mating practices or sex-biased migration events. Sex-biased admixture can be detected in various ways. The uniparental markers, the mitochondrial genome and the Y chromosome, are inherited through the female and male line respectively, and can thus be used to study sex-biased admixture events. The drawback from working with the uniparental markers is that they provide information from only a single locus, and thus only from a single lineage. Another way to detect sex-biased admixture is by comparing ancestries found on the X chromosome to those found on the autosomes. Since females inherit two copies of the X chromosome, and males inherit one copy, the X chromosome spends twice as much time in females than in males. Employing this knowledge, we can compare the pattern of genetic variation found on X chromosomes to that found on the autosomes by computing X-to-autosomal difference ratios (Rishishwar et al., 2015) per investigated ancestry follows:

$$\overline{\Delta Admix} = F_{anc,total} * (F_{anc,X} - F_{anc,auto}) / (F_{anc,X} + F_{anc,auto})$$

where $F_{anc,total}$ is the genome-wide admixture proportion for a given ancestry, $F_{anc,X}$ is the X chromosome admixture proportion for a given ancestry and $F_{anc,auto}$ is the autosomal admixture proportion for a given ancestry. Negative X-to-autosomal difference ratios are indicative of male-biased admixture for that ancestry, positive X-to-autosomal difference ratios are indicative of a female-biased admixture for that ancestry.

2.2.7 Detecting population expansions and contractions

An important aspect of human evolutionary research and population history is the investigation of the population size changes over time with the use of genetic data. Effective population size (N_e) is often used in these cases, and this is the number of breeding individuals in a population, in a simplified scenario. For most populations, the effective population size is lower than the actual number of individuals in the population, the census size. The effective population size can also be inferred going backwards in time, with specific methods like Pairwise Sequentially Markovian Coalescent (PSMC) (Li and Durbin, 2011), Multiple Sequentially Markovian Coalescent (MSMC) (Schiffels and Durbin, 2014), and GONE (Santiago et al., 2020). However, these methods employ information about heterozygous sites and recombination events and thus require diploid genomes. Bayesian skyline analyses can be performed on haploid genomes, such as mitochondrial DNA. I use this method on mitochondrial genomes in **Paper II** to provide insight into past population dynamics of females from various sub-Saharan African populations. Bayesian skyline analyses employ Markov chain Monte Carlo (MCMC) methods (Drummond et al., 2005). MCMC methods sample systematically from a probability distribution, where each new sample is dependent on the previous sample. In the case of the Bayesian skyline analyses, the effective population sizes through time are estimated from a distribution of genealogies. These genealogies have been generated using MCMC algorithms. The results are often represented in the Bayesian skyline plot (BSP), where the effective population size is plotted against time. An example of the effective population size through time of African individuals speaking Afro-Asiatic languages is shown in Figure 10. This allows for detection of population expansions or contractions. For example, expansions in the effective population size of Afro-Asiatic speakers seem to have occurred twice, once after the time period of the out-of-Africa migration and once after the Last Glacial Maximum. It should be noted that population sizes and haplogroup-specific sizes are not the same, and that, when investigating the effective population size of a specific population, all haplogroups occurring in that population should be included in the BSP. The software package BEAST is frequently used (Suchard et al., 2018) to perform Bayesian skyline analyses, and the program BEAUTi is often used to prepare the input file for BEAST. BEAST is a Bayesian approach with Markov chains searching the likelihood space. It is run for a certain amount of iterations and the output files have to be checked with Tracer, a software package that can visualize and analyze the MCMC trace files from BEAST. Specifically, Tracer can tell you how many iterations should be discarded (burnin phase), and what the Effective Sample Sizes (ESSs) are. These indicate whether or not the analysis produced a sufficient number of independent samples from the posterior distribution for a variety of parameters.

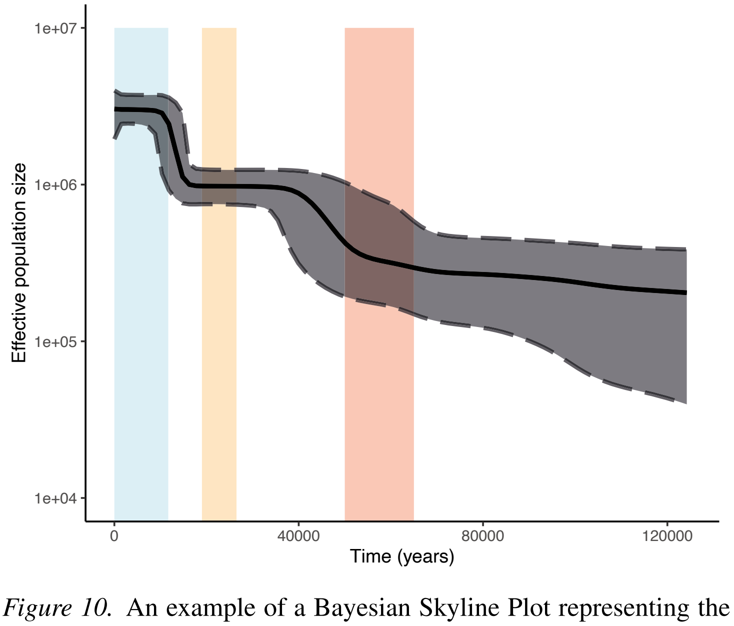


Figure 10. An example of a Bayesian Skyline Plot representing the female effective population size of individuals speaking Afro-Asiatic languages. On the Y-axis, the effective population size is shown with a log-scale, on the X-axis the time in the past (current-day on the left and going backwards in time to the right). This plot is based on mitochondrial sequences from 400 African individuals speaking Afro-Asiatic languages. The thick middle line represents the mean estimates and the two dashed lines represent the 95% highest posterior density (HPD) intervals. The red area denotes the time of the out-of-Africa migration (65–50 kya), the yellow area the Last Glacial Maximum (LGM)(26.5–19 kya), and the blue the Holocene (last 11.7 ky).

2.3 Spatial representation of data

When working with genetic data from a variety of regions, investigating different genetic patterns between those regions is of great interest. To represent the genetic data spatially, genetic clines and patterns are frequently plotted on maps. In **Paper II**, I have used the kriging method to represent the mitochondrial haplogroup frequencies across the African continent. This method predicts the values at an unknown point by calculating a weighted average of the surrounding data points (not taking into account physical barriers such as mountains). This way, even though no data was obtained from a particular area, inferences about occurrences of genetic patterns can be made. This should be done with caution, as these inferences do not necessarily represent true occurrences. An example of a frequency map that was plotted using the kriging method is shown in Figure 11.

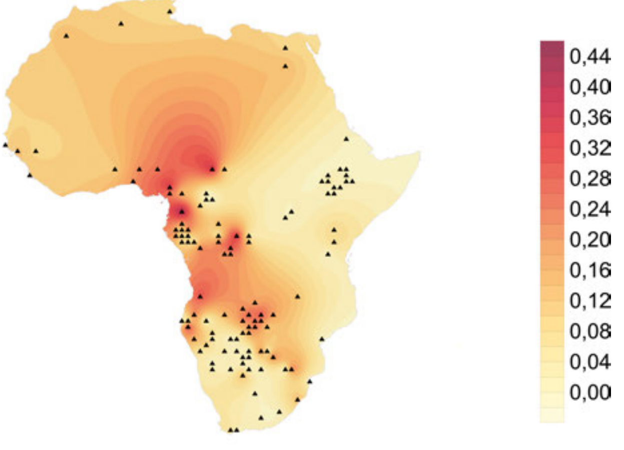


Figure 11. An example of a frequency map, which was plotted using the kriging method. It shows the frequency of the individuals carrying mitochondrial sequences that belong to mitochondrial haplogroup L3e. Darker red indicates higher frequencies of haplogroup L3e in that region. The triangles indicate the sampling locations on which this frequency map is based.

2.4 Analyzing ancient proteins

2.4.1 Methods to infer ancient protein sequences

Section 1.3.3 introduced the field of palaeoproteomics and focused on its applications. In this section, we will introduce some of the methods employed in ancient protein studies. The investigation of protein sequences has a long and diverse history. The first method to determine the amino acid sequence of a peptide was Edman degradation (Edman et al., 1950). Edman degradation, or Edman sequencing, relies on the chemical labeling of the amino-terminal of a protein, after which this amino-acid is cleaved and identified through chromatography. In 1953, the first amino acid sequence, part of the hormone insulin, was published (Sanger and Thompson, 1953). Unfortunately, Edman sequencing is not suited for ancient proteins, as it relies on purified, concentrated, and unmodified proteins (Robbins et al., 1993). In the 1980s, mass spectrometry (MS) became widely used to determine the sequence of proteins from modern-day humans. However, it is not until the advent of soft-ionization mass spectrometry in the year 2000 that mass spectrometry can be used to determine the sequence of ancient proteins (Ostrom et al., 2000).

Mass spectrometry is a technique that separates ions based on their mass-to-charge ratio (m/z), after which they are detected and this results in a specific MS spectrum. From the MS spectrum, a peptide can be identified or the the amino acid sequence of a protein can be deduced. In general, there are four parts to the process of MS for proteins: preparation of the sample, an ionization source, a mass analyzer and a system for ion detection. The process starts with the preparation of the sample, where proteins are extracted and proteases are added. These proteases cleave the proteins into smaller peptides, which are in the range of the mass spectrometer. Often, the protease trypsin is used, but more proteases are available. The sample is now ready for ionization. There is a variety of ways to ionize a sample, but matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are the two most commonly used techniques to ionize proteins and peptides for MS (Aebersold and Mann, 2003). At the same time, the samples are turned into a gas phase. In the third part, the peptides are separated based on their m/z values. In the last part of the mass spectrometer, these m/z values are detected and converted into a digital output. This digital output is often represented as a spectrum, with the measured intensities for each of the detected m/z values. These spectra can either be compared to a known database, or can, in some cases, be used to deduce the amino acid sequence of a peptide.

Here, I describe two methods used widely in the analysis of ancient proteins; peptide mass fingerprinting using MALDI-time-of-flight (MALDI-TOF) (Buckley et al., 2009) and liquid chromatography-tandem mass spectrometry (LC-MS/MS). The application of peptide mass fingerprinting using MALDI-TOF mass spectrometry is often referred to as zooarchaeology by mass spectrometry (ZooMS). It focuses on collagen proteins, a group of structural proteins abundant in connective tissues such as bone, teeth and skin. ZooMS allows you to identify species through matching the resulting MS spectra against a database, and it is therefore often used to determine species when morphological data is inconclusive. As opposed to ZooMS, LC-MS/MS actually allows for direct identification of a peptide sequence. LC-MS/MS is special in the sense that two mass analyzers are coupled. After the first detection, a mass spectrum (MS1) is generated, after which the peptides are subjected to another round of fragmentation, mass analysis and detection. This generates the MS/MS spectrum. Both the MS1 and MS/MS spectra are used to search against a protein sequence database, allowing for amino acid sequence determination (Aebersold and Mann, 2003). LC-MS/MS can investigate entire proteomes, and it is sometimes referred to as shotgun proteomics. Database searches can be performed with several programs, such as Proteome Discoverer (Thermo Fisher Scientific), MaxQuant, Mascot and Bionic. UniProt is often used as the actual database against which the spectra are searched.

2.4.2 Deamidation as a way to authenticate ancient proteins

When working with ancient remains, it is important to minimize contamination and make sure that the biomolecules you are working with are endogenous to the sample. When working with ancient DNA, precautions to minimize contamination with modern DNA are taken and various metrics can provide information about the authenticity of the DNA. For example, ancient DNA read length is often shorter than 50 base pairs (Hofreiter and Hartmann, 2020) and shows DNA damage in the form of deamination near the ends of the DNA fragments (Sawyer et al., 2012). Within the field of palaeoproteomics, a phenomenon called deamidation can be used to authenticate the ancient proteins. Deamidation is a a post-translational modification where amide groups in asparagine (N) or glutamine (Q) residues are converted into acidic residues, resulting in aspartic acid or glutamic acid. As deamidation increases over time, it serves as a valuable metric for determining the authenticity of ancient proteins. Deamidation rate of glutamine is more than 60 times slower than that of asparagine (Robinson and Robinson, 2001). In **Paper IV**, I use the program deamiDATE (Ramsøe et al., 2020) to analyze the deamidation patterns of ancient proteins. It infers the bulk deamidation (the total rate of deamidation), as well as site specific deamidation. Site specific deamidation looks at the deamidation of asparagin and glutamine in the context of the flanking amino acids. Deamidation rates are affected by the two surrounding amino acids.

3. Ethical considerations

Genetic research on humans, particularly in the context of evolutionary studies, holds great potential for deepening our understanding of human history and the processes that have shaped our species. However, it is important to be aware of the ethical considerations when conducting research on humans, both modern and ancient. The necessity arises from the need to respect individual rights, cultural sensitivities, and the potential effects that the findings may have on society. It has been shown that most genetic variation (94.6%) is found between individuals within groups, and only a small amount of genetic variation exists between groups (5.4%) (Rosenberg et al., 2002). Unfortunately, identifying differences between groups has the potential to be used in a discriminatory way, which is the basis for racism. Moreover, all cultural values are different. Following and applying western cultural values when working with other populations' history is also discriminatory. Thus, the explanation of the research beforehand and the presentation of these results should be performed with care. Here, I expand on the ethical considerations that come into play when conducting research involving modern populations, and research involving ancient humans. Moreover, I briefly touch on ethical considerations when performing research on the African continent, as the colonial history of Africa has large implications for current research. This should be taken into consideration and all means of exploitation should be avoided.

Most of the work in this thesis has been conducted on modern humans and involves many ethical considerations. First, the sampling protocol for modern samples has to be reviewed by an ethical board. Local ethics committees in the country of the investigated population have to approve the sampling protocol and the project as a whole. The projects, involving the sequencing and analyses of samples in Sweden, also need to be approved by the Swedish ethical review board. Informed consent is a cornerstone of ethical research involving modern human populations, and is obtained in a written form where possible. Researchers must ensure that participants fully understand the purpose, scope, potential implications, and risks of the study before agreeing to participate. To achieve this, it is important to involve community-based organizations in the field work. The guidelines set by the Swedish Ethics Board are specifically designed to safeguard participants. All documents containing personal data are securely stored, typically in a locked cabinet under the supervision of the project leader. Digital data is maintained in secure databases, such as the UPPMAX system, with access limited to authorized personnel. Likewise, collected samples and extracted DNA are kept in a secure freezer with controlled access. The samples are anonymized, with the key required to retrieve participant identities accessible only to authorized personnel. Study participants can always retract their participation. In that case, the sample and personal data obtained should be destroyed. When conducting research in Africa, or among other marginalized or underrepresented groups, additional care must be taken to ensure that consent is genuinely informed. For example, many hunter-gatherer languages lack specific words for 'science' or 'research' and these concepts are often merged into a single term synonymous with 'medicine' (Marshall and Marshall, 2007). This can pose significant risks when attempting to obtain informed consent in the context of genomic research that is not intended for any medical purposes. Moreover, as we live in a world mostly characterized by land ownership and boundaries, we have to be careful of the implications of genetic research, especially when we study pastoralist groups and hunter-gatherers (Bankoff and Perry, 2016; Marshall and Marshall, 2007). Local genetic ancestry can be and has been used to claim the right to land. Researchers should, therefore, be particularly mindful of the local political and historical contexts. It should be noted and explained that genetic research on modern populations does not provide definitive answers about where their ancestors lived. A final aspect to think about when studying population genetics of modern individuals, is the correct naming and referral of ethnicities and populations. Throughout this thesis, I use the names preferred by the communities and individuals themselves.

A smaller part of this thesis has been conducted using ancient human remains. My work in Paper IV, as well as my position in a research group with a focus on both modern and ancient humans, has made me aware of the ethical aspects of working with ancient human remains. Unlike working with modern individuals, laws and regulations are generally less refined when it comes to working with ancient human remains. As the sampling is destructive, i.e. the material is destroyed with the purpose of obtaining biomolecular information, it is critical to minimize this harm to the sample. A sampling permit from the local museum or heritage agency should also be obtained, and should include the permission to export the human remains to our clean lab in Uppsala, Sweden. It is also important to be aware of the implications of working with human remains. In the case of populations practicing ancestor religions, revealing the genetic ancestry of ancient human remains can have large implications for the communities that worship them. Thus, communication and collaboration between anthropologists, archaeologists, community representatives, and geneticists in this type of investigation is crucial.

Lastly, I want to reflect on the ethical considerations when performing research on the African continent. The colonial history of Africa has large implications for current research and it is important to be aware of this as researchers based in Europe. It is important for both work involving modern populations and ancient human remains to give feedback to the communities involved. This can be done by presenting the general findings of the research back to the community or, in the case of modern populations, in the form of an individual report. Also, the education and involvement of African researchers in our studies has been a focus point.

To conclude, the ethical complexity of genetic research on African populations arises from the need to balance scientific interest with respect for individual rights, cultural sensitivities, and the potential societal impacts of the findings. Researchers have to be aware of these aspects and carry out the research with utmost care.

4. Research aims

The aim of my thesis work is to investigate both modern and ancient population structure on the African continent to infer historical demographic processes. Specifically, my research focuses on the following objectives:

I Assess and develop new methods to analyze mitochondrial genomes (Paper I and II).

II Provide a new overview of the mitochondrial and Y chromosome diversity present on the African continent to infer maternal and paternal histories (Paper II and III).

III Investigate the potential differences in protein sequences between Khoe-San populations and all other modern human populations (Paper IV).

IV Shed light on the ancestral genetic components present in various South African Coloured populations and identify geographical differences in their genetic background (Paper V).

5. Summary of papers

Paper I: The performance of common SNP arrays in assigning African mitochondrial haplogroups

In my first paper I focus on the methodological aspect of researching maternal histories in sub-Saharan Africa (Lankheet et al., 2021). As the accurate assignment of mitochondrial haplogroups is important for the field of evolutionary genetics, I explore the mitochondrial haplogroup assignment in African populations when genome-wide SNP array data is available. We look specifically at African mitochondrial haplogroups, as African populations have high genetic diversity, including in their mitochondrial genomes. I investigate the performance of eight commonly used SNP arrays in assigning African mitochondrial haplogroups and show that two SNP arrays, the Illumina Infinium Multi-Ethnic Global-8 and the Illumina Infinium Multi-Ethnic AMR/AFR-8, outperform the other arrays when it comes to the correct assignment of African mitochondrial haplogroups. They were able to assign 59–81% of the African mitochondrial haplogroups. Other SNP arrays were able to assign 4–62% of the African mitochondrial haplogroups. We also assessed the performance of HaploGrep2 for assigning mitochondrial haplogroups from SNP array data. HaploGrep2 is a free tool for automated haplogroup assignment. We show that two SNP arrays show very low haplogroup assignment percentages. Leaving these two SNP arrays aside, we generally observe moderate to good haplogroup assignment percentages (58–86%). These results provide the first cross-checked quantification of mitochondrial haplogroup assignment performance from SNP array data, indicating that mitochondrial haplogroup frequencies inferred from most common SNP arrays used for human population analysis should be considered with caution. The testing of various SNP arrays presented in this study regarding their ability to assign mitochondrial haplogroups correctly will help researchers to decide whether they should include mitochondrial haplogroup assignment in their research based on SNP arrays.

Paper II: Maternal histories of sub-Saharan Africa revealed through sequencing of complete mitochondrial genomes

As mitochondrial genomes are inherited through the female line, they are an important part of our genome when it comes to studying maternal histories of an individual or a population. The African continent is especially interesting in this regard, as it is home to the highest diversity of mitochondrial DNA. In Paper I, I have shown that, generally, SNP arrays do not provide enough information for detailed mitochondrial haplogroup assigment. Therefore, we developed a long-range sequence assay to amplify full-sequence mitochondrial genomes, and used it to generate mitochondrial sequences from underrepresented regions in Africa, such as the DRC and Ethiopia. We generated data from a total of 1288 complete mitochondrial genomes from 14 different countries across sub-Saharan Africa. These genomes are used to delineate mitochondrial haplogroups and trace maternal ancestry. We merge the newly generated mitochondrial sequences with over 3600 available African mitochondrial genomes, and provide an overview of African mitochondrial haplogroups and give insights into the maternal background of the African continent. We show the frequency distribution of the ten most frequently occuring haplgroups and through the assessment of effective population sizes over time, we date the onset of the expansion of Niger-Congo and Mande speakers to 17 thousand years ago (kya) and the Bantu Expansion to 6 kya. Moreover, I investigate the mitochondrial haplogroups associated to the Bantu Expansion. By looking at nucleotide diversity patterns and female effective population sizes over time, I identify one haplogroup, haplogroup L3e, as one of the major haplogroups of the original expanding Bantu-speaking groups, and that the Bantu Expansion brought it to many parts of the continent. We confirm the significance of L3e as a crucial marker for the Bantu Expansion, unveiling its central role for the first time in this context. In summary, the results from this study not only enhance our understanding of African mitochondrial diversity and maternal lineage histories, it also sheds light on significant demographic events such as the Niger-Congo and Bantu Expansions.

Paper III: Y chromosome diversity and geographical patterns in sub-Saharan Africa

The non-recombining portion of the Y chromosome (NRY) provides a valuable resource for investigating paternal lineages due to its inheritance through the paternal line and lack of recombination. In this study, we employed target enrichment to sequence 2.36 Mb of the NRY in 369 males from 28 locations across 11 sub-Saharan African countries. We identified three distinct patterns of geographical distribution for Y haplogroups; those confined to southern Africa, such as A1b1b2a (A-M51); those widely distributed across sub-Saharan Africa, like the E1b1a1 (E-M2) haplogroups; and those of non-African origin, which exhibit highly localized occurrence. Additionally, our findings reveal that the Bantu-speaking Lemba and Remba populations possess a unique Y haplogroup composition, distinct from other Bantu-speaking groups, aligning with their oral history and highlighting their unique paternal lineage. In contrast, other Bantu-speaking populations display a high prevalence (57–100%) of the E1b1a1 (E-M2) haplogroup. Overall, this study provides a comprehensive overview of the paternal histories of the populations studied.

Paper IV: Ancient proteins and the deepest population divergence among modern human populations

The field of aDNA has taught us a wide variety of aspects about human evolution and human history. For Paper IV, we shift our focus from DNA to proteins as DNA preservation and survival are limited. In addition to DNA, proteins also contain valuable molecular biological information and they generally preserve longer than DNA. Especially for very old samples in hot and humid environments, protein data could provide unique information, which in turn could be extremely valuable for deciphering early human evolution in Africa. For this paper, I investigate the potential differences in protein sequences between the two groups representing the deepest population divergence in the tree of all modern humans; Southern African hunter-gatherer Khoe-San groups on the one hand, and the rest of humanity on the other hand. We make use of *in silico* translated genetic data, as well as protein data from ancient human remains from Africa. We developed a model based on 88 amino acids that exhibit frequency differences between Khoe-San and non-Khoe-San individuals. We show that 14 of these amino acids can be detected in ancient human remains. Additionally, we confirm the authenticity of ancient proteins by examining their deamidation patterns. These findings pave the way towards the careful exploration of protein analysis in older samples, with the aim of deciphering the affinity of ancient samples to either one of the branches representing the earliest split among modern humans. This will then shed light on population continuity and ancient population structure in sub-Saharan Africa.

Paper V: Wide-scale geographical analysis of genetic ancestry in the South African Coloured population

Whereas the papers up to this point in the thesis have focused on a methodology or wider view of genetics of African populations, in Paper V I investigate the demographic histories of one population specifically; the South African Coloured (SAC) population. The SAC population is the most admixed population in South Africa, resulting from centuries of complex migration patterns, admixture, and historical segregation. SAC individuals are the descendants of local Khoe-San and Bantu-speaking populations, European settlers, and enslaved individuals from Africa and Asia. In this paper, I investigate the genetic makeup of SAC individuals, utilizing autosomal genotypes alongside mitochondrial and Y chromosome data. We generated new genotype data for 125 individuals identifying as SAC from seven locations. We merge the genotype data from these individuals with genotype data from 231 SAC individuals from 15 other locations and investigate the mosaic of ancestries in these individuals. We focus specifically on geographical differences in ancestry and find variations observed across the sampled regions. Khoe-San ancestry predominates in 14 of the 22 locations, underscoring its enduring influence within the SAC population. Geographical analysis highlights distinct patterns of ancestry distribution: inland regions exhibit higher proportions of Khoe-San ancestry, eastern regions show more Bantu-speaker/West African ancestry, and western/coastal areas, particularly around Cape Town, have increased Asian ancestry. This geographic divergence mirrors historical migration routes and settlement patterns, offering insights into the complex interplay of factors shaping SAC genetic diversity. Moreover, we also investigate whether the admixture events shaping the SAC population were sex-biased by calculating sex-biased admixture ratios. We observe negative values for East African and European ancestries, indicating male-biased admixture from these populations. At the same time we observe a positive value for Khoe-San ancestry suggesting female-biased admixture. This sex-biased pattern was confirmed by results from mitochondrial DNA and Y chromosome data. Overall, this study shows the complex and varied genetic backgrounds of individuals identifying as SAC, thereby offering valuable insights into South Africa's diverse historical migrations and genetic landscape.

6. Conclusions and future prospects

Studying the genetics of humans living in Africa is a fascinating research field, given humanity evolved on this continent. Despite this fact, African populations have been relatively understudied due to the large focus on Europeans, and later also on Asian populations. African populations, as I show in this thesis, hold the key to answering many pivotal questions about human history. By exploring human proteomes, as well as investigating different regions of the human genome, such as the mitochondrial genome, Y chromosome and SNPs, I shed light on African population genetics from multiple perspectives. Furthermore, I do not just rely on existing methodologies; I also assess and develop new approaches to investigate and uncover ancient African population structure. In **Paper I**, I investigate mitochondrial haplogroup assignment in African populations using commonly used genome-wide SNP arrays. I show that two SNP arrays outperform the other six investigated arrays, the Illumina Infinium Multi-Ethnic AMR/AFR-8 and the Illumina Infinium Multi-Ethnic Global-8 array. However, I also caution that haplogroup assignment based on SNP array data should always be approached with care. Therefore, in **Paper II**, I develop a new assay to sequence full mitochondrial genomes and use it to generate mitochondrial sequences from underrepresented regions in Africa, Ethiopia and the DRC amongst others. Using this data, together with published mitochondrial genomes, I provide an overview of African mitochondrial haplogroups and offer new insights into the maternal background of the African continent. Our results show that the Niger-Congo and Bantu Expansions started earlier than previously thought, and that mitochondrial haplogroup L3e was one of the major haplogroups of the original expanding Bantu-speaking groups. Future research targeting various other genomic regions could provide new insights into the timing of the Niger-Congo and Bantu Expansions, potentially confirming or challenging the theories proposed here. In **Paper III**, I present an overview of the Y haplogroups found among 369 individuals from sub-Saharan Africa, as obtained by targeted enrichment of their Y chromosomes. I identify three different patterns of geographical distribution and show that two Bantu-speaking populations, the Lemba and Remba, have a distinct Y haplogroup composition from other Bantu-speaking groups. Future directions for this paper could focus on expansions of paternal lineages, as was done for the maternal lineages. Comparing maternal and paternal histories can provide insight into patrilocality and matrilocality patterns in Africa.

In the following paper, **Paper IV**, I investigate protein differences between the two groups representing the deepest population divergence in the tree of all modern humans; Khoe-San people and all other humans. I show that there are protein differences in bone- and teeth-related proteins and that these can be investigated in ancient human remains. However, this study also demonstrates the limitations of the current protein analysis methods in distinguishing ancient Khoe-San and non-Khoe-San individuals from ancient remains. Although the *in silico* analysis highlights the potential of investigating proteins in ancient specimens, protein coverage should be improved, before investigating older samples from Africa. Future research can focus on improving the protein coverage, to increase the chances of finding informative amino acids. Moreover, the robustness of the *in silico* model will need to be tested with additional modern sequences of Khoe-San and non-Khoe-San people. In the last paper, **Paper V**, I focus on the genetic make-up and demographic history of one population specifically, the SAC. I reveal geographical differences in ancestry proportions and show that this population contains patterns of sex-biased admixture.

Overall, my thesis explores various dimensions of African population genetics and proteomics. I have assessed and employed a variety of methods, studying both large-scale African population structure and smaller, more specific populations such as the SAC. I have shed light on key events in the Holocene, such as the Bantu Expansion and the expansion of Niger-Congo speakers. Altogether, my thesis contributes to a deeper understanding of African population genetics by revealing complex ancestral dynamics, refining genetic and proteomic methodologies, and offering new insights into the demographic history of diverse African populations. This thesis also emphasizes the complexity and diversity of African population history. One of the key unresolved questions in human evolution is identifying the specific region(s) of origin within Africa. Model testing methods could play a crucial role in addressing this question, offering tools to analyze genetic data, simulate evolutionary scenarios, and refine our understanding of population dynamics across the continent. Advancements in aDNA methods bring hope for increasing the number of samples from Africa, offering the potential to gather data from a wider range of time periods and geographically diverse regions. Human evolution research has historically been somewhat fragmented, with archaeologists, linguists, and geneticists partly working in isolation. However, increasing collaboration among these disciplines is fostering a more integrated approach, which is enriching our understanding and providing more holistic views of human evolution. Furthermore, research into pathogens and microbiomes in both modern and ancient populations can give insights into natural selection processes. This can enhance our understanding of how humans adapt to new environments, a topic of growing relevance in the context of climate change and future epidemics.

7. Svensk sammanfattning

Den anatomiskt moderna människan har sitt ursprung i Afrika. Trots detta förblir Afrika en tämligen outforskad region. I min avhandling använder jag genetiska och proteomiska data för att förstå människans evolution och historia. Jag avslöjar hur gamla populationsstrukturer i Afrika såg ut, och hur nuvarande populationer har bildats. En del av mina studier fokuserar på att testa och utveckla olika nya metoder för att generera genetiska och proteomiska data från människor som lever nu, men också från människor som levde förr. För övrigt fördjupar jag mig i både större genetiska mönster och i finskaliga populationstrukturer med hjälp av olika genetiska markörer.

I första delen av min avhandling fokuserar jag på en speciell typ av data som vi kallar för enbaspolymorfier, vilket är positioner i genomet som skiljer sig mellan olika människor. Man kan ta reda på vilka baser som finns i en enbaspolymorfi genom at sekvensera individer med hjälp av mikromatrisanalysteknik. Det finns olika typer av mikromatrisanalyser som har utvecklats av olika företag, och som inkluderar lite olika enbaspolymorfier. Man skulle kunna använda data från dessa mikromatrisanalyser för att forska på en specifik del av genomet; det mitokondriella DNAt (mtDNA), som både män och kvinnor ärver från sina mödrar och som därmed ger information om moderslinjen. I första artikeln har jag analyserat vilken mikromatrisanalys som fungerar bäst för att forska på mtDNA från Afrika, så vi kan dra de mest tillförlitliga slutsatserna om den afrikanska befolkningen. I detta arbete finner jag att moderslinjer som härleds med hjälp av alla olika mikromatrisanalystekniker bör betraktas med försiktighet, men att det finns vissa analyser som man skulle kunna lita på mer än andra.

I följande studie, studie II, fortsätter jag med att utforska mtDNA och modernlinjer hos olika populationer i Afrika. Istället för att undersöka enbaspolymofier, utvecklade jag en metod för att sekvensera hela mtDNA genom, och jag använder den nya metoden för att sekvensera nästan 1 300 nya mtDNA genom från människor i underrepresenterade regioner i Afrika. Genom att analysera dessa tillsammans med ungefär 3 600 publicerade mtDNA-sekvenser, ger jag en överblick över den maternella genetiska bakgrunden i Afrika. Genom att analysera populationsstorleken hos kvinnor över tid upptäckte jag att två expansioner inträffade tidigare än vad man tidigare trott.

I studien III har jag sekvenserat delar av Y-kromosomen från olika populationer i subsahariska Afrika. Vi identifierade tre olika geografiska spridningsmönster av Y-haplogrupper; de som begränsas till södra Afrika, de som är vitt spridda över hela subsahariska Afrika, och de med ett icke-afrikanskt ursprung som endast förekommer ytterst lokalt. Vidare fann vi att de Bantutalande Lemba- och Rembafolken förekommer i en distinkt sammansättning av Y-haplogrupper som skiljer sig från andra Bantutalande folkgrupper.

Eftersom bevarandet och överlevnaden av DNA från gamla ben är begränsad, flyttar jag fokus från DNA till proteiner i studie IV. Jag utforskar användningen av proteomik hos forntida individer för att uttyda djupgående populationsstrukturer i Afrika. Jag undersöker de potentiella skillnaderna i proteinsekvenser mellan de två grupper som representerar den djupaste förgreningen i det evolutionära trädet och som omfattar alla moderna människopopulationer. Jag identifierar variation i aminosyror mellan södra Afrikas jägar-samlande Khoe-San-grupperna å ena sidan, och resten av mänskligheten å andra sidan, och jag visar att dessa aminosyror kan undersökas i gamla prover från Afrika. Dessa fynd banar väg för en noggrann utforskning av proteinanalys i äldre prover, från vilka inget gammalt DNA kan utvinnas.

I den sista studien (V) undersöker jag genetiken hos en specifik population; de så kallade Coloured-folken (svenska ”färgade”) i Sydafrika. Det är den population som har mest blandad härkomst i Sydafrika, och troligtvis även i världen. De är ättlingar till såväl Khoe-Sanfolken, individer med västafrikansk härkomst, pastoralist-folken från Östafrika, som europeiska kolonisatörer samt slavar från Syd- och Ostasien. Genom att använda enbaspolymorfidata från dessa Coloured-individer identifierar jag geografiska skillnader i proportionerna av de olika genetiska härkomsterna och även vissa skillnader mellan män och kvinnor i dessa proportioner. Jag identifierade högre Khoe-San-bidrag i inlandsregioner och mot öster, och högre bidrag från personer med västafrikansk härkomst i östliga regioner, medan de asiatiska härkomsterna är högre i västra regioner.

Sammantaget har mitt arbete för denna avhandling fördjupat vår förståelse för kontinentala och regionala genetiska strukturer i afrikanska populationer. Genom att använda olika typer av genomdata såväl som proteomdata, belyser jag både metodologiska framsteg inom området, expansioner som inträffade tidigare än vad man tidigare trott, samt den unika och blandade bakgrund som innehas av Coloured-befolkingarna.

8. Declaration of generative AI in scientific writing

During the preparation of this work I used ChatGPT (version 4) in order to revise and refine my written text, making it more smooth, readable, and understandable. After using this tool, I reviewed and edited the content as needed and take full responsibility for the content of this thesis and the manuscripts it includes.

9. Acknowledgements

I would like to start by thanking **Carina** for all the knowledge that she has shared with me, and for being such a supportive supervisor. I am very grateful for the opportunity that she has given me by offering me the PhD position and guiding me throughout this journey, I have learned so much! Despite having many more tasks than being a researcher and my supervisor, Carina always kept her office door open for my questions, offering her time and expertise whenever I needed it. I also appreciate the warmth she has shown me, both in and out of the academic setting. Whether it was during our time at conferences in Cape Town and Puerto Vallarta, our squash matches, or the trust she placed in me by allowing me to house-sit for her, I truly value the kindness and support she has for me and others.
It felt quite exceptional to have three more PhD supervisors; Cécile, Helena and Max. I would like to start by thanking my colleague and dear friend, **Cécile**. Her support in the lab was invaluable, especially during the challenging moments of writing, extensive analyses, and collaboration. She was always there to lend an ear when I struggled. Beyond the professional help, Cécile has been a great friend. Whether it was taking me in her car to go for a swim at Wiks Slott, to go on climbing adventures, or offering emotional support during life's ups and downs, she was always there for me. Her kindness, patience, and emotional support made this journey so much more manageable and meaningful. Thank you, Cécile, for being such a wonderful part of my work and my life. I would also like to thank **Helena**, who has been great at guiding me whenever needed, often with great flexibility and spontaneity. I have always appreciated how she took the time to sit down with me and discuss problems, revised manuscripts and showed genuine interest in the things I was doing. I would also like to thank **Max**, the newest addition to the supervisor team. Max probably only saw the stressed version of me as he joined the supervisor team in the last year, so his enthusiasm and calmness were a great support. Max is great at explaining things that I for some reason still did not understand after three years.
I also appreciate all the help of the people who did their PhD at Human Evolution at the same time as I did; **Mario**, **Gwenna**, **Luciana**, **Natalija**, **Rickard**, **James**, **Pedro**, **Nik**, and **Jules**. Thank you for listening to me, understanding the ups and downs of the PhD journey, sharing laughs, dinners, and drinks, and offering tips and support. I would like to give a special shout-out to the supervisor during my master's project, Mario. If it was not for his great supervision style using a water-spray bottle, great fashion style and brutal honesty about

the status of my plants, I would not have been able to start this PhD position. Another special thanks to Rickard, my big PhD brother, who has become a really good friend over the years. Despite the demands of his expanding family with a wife and two tiny ones, he always made time for me, listened to my complaints and he and is wife **Maya** welcomed me in his home. Thank you! I would also like to thank Jules specifically, they have been my office-mate for a long time. They are always open to help, the best way to do that was often "een koffietje doen".
I am thankful for all the past and present members of the Schlebusch lab; **Concetta**, **Cesar**, **Nina**, **Afifa**, **Ezekiel**, **Johanna**, **Grace**, **Modeste**, **Andrea** and **Ivany**. Specifically, Afifa has been a great support in the lab, helping me with the vast amount of labwork that is part of this PhD thesis. I want to acknowledge all the members of Human Evolution for their great support throughout this journey; **Mattias**, **Magdalena**, **Magnus**, **Torsten**, **Per**, **Hanna**, **Rita**, **Tiina**, **Miguel**, **TJ**, **André**, **Freyja Ayca**, **Marianne**, **Tina**, **Mariam**, **Gérman**, **Mika**, **Alex**, and **Anna**. I am very thankful for all the great input I have gotten on projects and presentations, and was very grateful for the nice fika's everyone made, and the fun we had at defenses, outings, and departmental parties. In specific, I would like to thank **Carro** for helping me with tons of bioinformatic questions, that then suddenly took up half her day. Special thanks to **E-Jean** as well, for supporting me with great coffee and knitting talks.
I would like to thank my friend and colleague **Mariam** specifically. For the short period that we have known each other, you have become such a good friend and a great support at the end of my thesis, both emotionally as well as scientifically. I appreciate you picking me up when I am down, listening to me complain, picking berries with me, and laughing with me about the craziness of life. Also thanks to **Mar**, because she, by proxy, has to hear about all of this as well, and opens her home to me, cooking me great Spanish food.
I want to thank all the collaborators that have made the studies presented in this thesis possible. Those that were part of the sampling teams, have done analyses, and have given me feedback on my manuscripts. Specifically, **Nikolay** and **Claudio**, who have done a lot of work on the protein manuscript. Thank you for being patient with me always, and for teaching me so many new things!
I want to express my thanks to **ChatGPT**, who has been able to answer 823 questions that I have had during my PhD. Those were probably not the right answers, but I am happy for the insight they have provided me.
Teaching is a large part of the PhD and I met so many wonderful people while teaching. Many thanks to **Shreya**, **Lars**, **Fabio**, **Carlos**, and **Charlotta**. And many thanks to all my other friends at EBC; **Eleonora**, **Sifra**, **Matt**, **Mathieu**, and **Jake**.
Then there is a part of my life that most of you have not heard about, I also climb sometimes. I do that with some amazing people, they motivate me, take me on climbing trips, and I am very thankful that they share my passion.

Thank you **Bart**, **Sofia**, **Jenny**, **Craig**, **Marco**, **Jack**, **Martin**, **Mario**, **Tom**, **Venu**, **Jess**, **Jana**, **Jimmy**, **Tobias**, **Ramiro**, **Elena** and **Baptiste**. Special thanks to **Jipp** and **Fons**, who are just the cutest and I will miss you now that you do not live a bike ride away.

I would like to thank my friends **Hannah**, **Nadja**, and **Alessia** for being such a big support during my PhD time. You always listened to me and gave me great advice. If I was slightly angry or feeling low, you helped me greatly.

Many thanks to my friends back home; **Annick**, **Lotte**, **Frédérique**, and **Laura**. I am also grateful for all my friends from the knitting group: **Anna**, **Nina**, **Freja**, **Tuuli**, **Karin**, and **Jes**. They supported me at least once a week. Specifically, I would like to thank **Frida**, who made the cover of this thesis.

I would like to thank my friends **Marcel** and **Hannah**, who have been a great support the last year, providing me with emotional support and with food, and awesome time with their kids **Margot** and **Astrid**.

Thanks to my fellow thesis-club members and dear friends from Utrecht, **Hanne** and **Brigitte**, for doing a gazillion pomodoro's with me, and literally helping me with writing, coding problems, and life's struggles in general.

Als laatste wil ik mijn familie bedanken. Dankjewel **Aimy**, dat je altijd met mij wil kletsen, mij bent gevolgd naar het Noordelijkste puntje waar we ooit zijn geweest en er altijd voor me bent. Je bent het allerliefste zusje van de hele wereld. Ook wil ik **Steven** en **Roos** bedanken, die altijd een luisterend oor bieden en geïnteresseerd zijn in wat ik doe. Jullie positieve instelling en hulp is voor mij heel waardevol. Ik wil **oma Marijke** bedanken, voor de gezellige belletjes in de weekenden. Ik wil ook mijn ouders, **Martin** en **Hanneke**, bedanken. Jullie zijn een geweldig voorbeeld en ik had me geen betere ouders kunnen wensen. Ik wil jullie bedanken voor de oneindige steun die jullie mij geboden hebben tijdens de PhD. Ondanks de afstand, waren jullie er altijd voor mij. Als het nodig was, kwamen jullie naar Zweden en anders waren jullie ook altijd bereikbaar om te bellen. Dankjewel papa, dat je mij zoveel geholpen hebt met de inhoud van mijn PhD, dat je bereid was om stukken te lezen en van commentaar te voorzien, dat ik jou kon uitleggen wat ik doe en dat je me daardoor nieuwe inzichten hebt gegeven. Papa en mama, ik ben jullie eeuwig dankbaar voor alles wat jullie voor me hebben gedaan.

10. References

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207.

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1):246–246.

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., and Bunce, M. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci*, 279(1748):4724–4733.

Alves-Silva, J., da Silva Santos, M., Guimarães, P. E., Ferreira, A. C., Bandelt, H.-J., Pena, S. D., and Prado, V. F. (2000). The ancestry of Brazilian mtDNA lineages. *The American Journal of Human Genetics*, 67(2):444–461.

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., and Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 23(2):147.

Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular biology and evolution*, 25(2):468–474.

Ávila-Arcos, M. C., Raghavan, M., and Schlebusch, C. (2023). Going local with ancient dna: A review of human histories from regional perspectives. *Science*, 382(6666):53–58.

Bajić, V., Schulmann, V. H., and Nowick, K. (2023). mtDNA "Nomenclutter" and its consequences on the interpretation of genetic data. *bioRxiv*.

Bajić, V., Barbieri, C., Hübner, A., Güldemann, T., Naumann, C., Gerlach, L., Berthold, F., Nakagawa, H., Mpoloka, S. W., Roewer, L., Purps, J., Stoneking, M., and Pakendorf, B. (2018). Genetic structure and sex-biased gene flow in the history of southern African populations. *American Journal of Physical Anthropology*, 167(3):656–671.

Bandelt, H.-J., Alves-Silva, J., Guimaraes, P., Santos, M., Brehm, A., Pereira, L., Coppa, A., Larruga, J. M., Rengo, C., Scozzari, R., et al. (2001). Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Annals of Human Genetics*, 65(6):549–563.

Bandelt, H. J., Forster, P., Sykes, B. C., and Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics*, 141(2):743–753.

Bankoff, R. J. and Perry, G. H. (2016). Hunter–gatherer genomics: evolutionary insights and ethical considerations. *Current opinion in genetics & development*, 41:1–7.

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

Barbieri, C., Hübner, A., Macholdt, E., Ni, S., Lippold, S., Schröder, R., Mpoloka, S. W., Purps, J., Roewer, L., Stoneking, M., et al. (2016). Refining the Y chromosome phylogeny with southern African sequences. *Human genetics*, 135(5):541–553.

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

Barham, L., Mitchell, P., et al. (2008). *The first Africans: African archaeology from the earliest toolmakers to most recent foragers*. Cambridge University Press.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173.

Barnard, A. (1992). *Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples*. Cambridge Studies in Social and Cultural Anthropology. Cambridge University Press.

Batini, C., Coia, V., Battaggia, C., Rocha, J., Pilkington, M. M., Spedini, G., Comas, D., Destro-Bisol, G., and Calafell, F. (2007). Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. *Molecular Phylogenetics and Evolution*, 43(2):635–644.

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. (2010). Insights into the Demographic History of African Pygmies from Complete Mitochondrial Genomes. *Molecular Biology and Evolution*, 28(2):1099–1110.

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(4):675–684.

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., Bertranpetit, J., Quintana-Murci, L., Tyler-Smith, C., Wells, R. S., and Rosset, S. (2008). The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics*, 82(5):1130–1140.

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics (Oxford, England)*, 32(18):2817–2823.

Beleza, S., Gusmão, L., Amorim, A., Carracedo, A., and Salas, A. (2005). The genetic legacy of western Bantu migrations. *Human Genetics*, 117(4):366–375.

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., van der Veen, L., Hombert, J.-M., Quintana-Murci, L., and Comas, D. (2009). Genetic and Demographic Implications of the Bantu Expansion: Insights from Human Paternal Lineages. *Molecular Biology and Evolution*, 26(7):1581–1589.

Bleasdale, M., Richter, K. K., Janzen, A., Brown, S., Scott, A., Zech, J., Wilkin, S., Wang, K., Schiffels, S., Desideri, J., et al. (2021). Ancient proteins provide evidence of dairy consumption in eastern Africa. *Nature communications*,

12(1):1–11.
Bostoen, K. (2018). *The Bantu Expansion*. Oxford University Press, Oxford, UK.
Breton, G., Barham, L., Mudenda, G., Soodyall, H., Schlebusch, C. M., and Jakobsson, M. (2024a). BaTwa populations from Zambia retain ancestry of past hunter-gatherer groups. *Nature Communications*, 15(1):7307.
Breton, G., Sjödin, P., Zervakis, P. I., Laurent, R., Froment, A., Sjöstrand, A. E., Hewlett, B. S., Barreiro, L. B., Perry, G. H., Soodyall, H., Heyer, E., Schlebusch, C. M., Jakobsson, M., and Verdu, P. (2024b). Ancient tree-topologies and gene-flow processes among human lineages in Africa. *bioRxiv*.
Brown, T. A. (1999). How ancient DNA may help in understanding the origin and spread of agriculture. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1379):89–98.
Buckley, M., Collins, M., Thomas-Oates, J., and Wilson, J. C. (2009). Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 23(23):3843–3854.
Buckley, M., Fraser, S., Herman, J., Melton, N. D., Mulville, J., and Pálsdóttir, A. H. (2014). Species identification of archaeological marine mammals using collagen fingerprinting. *Journal of Archaeological Science*, 41:631–641.
Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). MtDNA and human evolution. *Nature*, 325(6099):31–36.
Cashdan, E. (1986). Hunter-gatherers of the northern kalahari. In Vossen, R. and Keuthmann, K., editors, *Contemporary Studies on Khoisan In Honour of Oswin Köhler on the Occasion of his 75th Birthday*, volume 5 of *Quellen zur Khoisan-Forschung/Research in Khoisan Studies*, pages 145–180. Helmut Buske, Hamburg.
Cerezo, M., Černý, V., Carracedo, Á., and Salas, A. (2011). New insights into the Lake Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs. *PLoS One*, 6(4):e18682.
Černý, V., Salas, A., Hájek, M., Žaloudková, M., and Brdička, R. (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(4):433–452.
Chen, F., Welker, F., Shen, C.-C., Bailey, S. E., Bergmann, I., Davis, S., Xia, H., Wang, H., Fischer, R., Freidline, S. E., et al. (2019). A late middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature*, 569(7756):409–412.
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. *The American Journal of Human Genetics*, 66(4):1362–1383.
Chen, Y.-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S., and Wallace, D. C. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American Journal of Human Genetics*, 57(1):133.
Coutu, A. N., Taurozzi, A. J., Mackie, M., Jensen, T. Z. T., Collins, M. J., and Sealy, J. (2021). Palaeoproteomics confirm earliest domesticated sheep in southern

Africa ca. 2000 BP. *Scientific Reports*, 11(1):6631.
De Bruyn, G. (1976). Die samestelling van die Afrikanervolk [The composition of the Afrikaner population]. *Tydskr Geesteswet*, 15:39–42.
De Filippo, C., Bostoen, K., Stoneking, M., and Pakendorf, B. (2012). Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3256–3263.
De Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., Van Helden, P. D., Seoighe, C., and Hoal, E. G. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human genetics*, 128:145–153.
Demarchi, B., Hall, S., Roncal-Herrero, T., Freeman, C. L., Woolley, J., Crisp, M. K., Wilson, J., Fotakis, A., Fischer, R., Kessler, B. M., Rakownikow Jersie-Christensen, R., Olsen, J. V., Haile, J., Thomas, J., Marean, C. W., Parkington, J., Presslee, S., Lee-Thorp, J., Ditchfield, P., Hamilton, J. F., Ward, M. W., Wang, C. M., Shaw, M. D., Harrison, T., Domínguez-Rodrigo, M., MacPhee, R. D., Kwekason, A., Ecker, M., Kolska Horwitz, L., Chazan, M., Kröger, R., Thomas-Oates, J., Harding, J. H., Cappellini, E., Penkman, K., and Collins, M. J. (2016). Protein sequences bound to mineral surfaces persist into deep time. *eLife*, 5:e17092.
Diamond, J. and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science*, 300(5619):597–603.
Dimmendaal, G. (2008). Language Ecology and Linguistic Diversity on the African Continent. *Language and Linguistics Compass*, 2:840–858.
Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.
Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.
Edman, P., Högfeldt, E., Sillén, L. G., and Kinell, P.-O. (1950). Method for determination of the amino acid sequence in peptides. *Acta chem. scand*, 4(7):283–293.
Evans, P. D., Mekel-Bobrov, N., Vallender, E. J., Hudson, R. R., and Lahn, B. T. (2006). Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. *Proc Natl Acad Sci U S A*, 103(48):18178–18183.
Fernandes, V., Triska, P., Pereira, J. B., Alshamali, F., Rito, T., Machado, A., Fajkošová, Z., Cavadas, B., Černỳ, V., Soares, P., et al. (2015). Genetic stratigraphy of key demographic events in Arabia. *PloS one*, 10(3):e0118625.
Fortes-Lima, C. A., Burgarella, C., Hammarén, R., Eriksson, A., Vicente, M., Jolly, C., Semo, A., Gunnink, H., Pacchiarotti, S., Mundeke, L., Matonda, I., Muluwa, J. K., Coutros, P., Nyambe, T. S., Cikomola, J. C., Coetzee, V., de Castro, M., Ebbesen, P., Delanghe, J., Stoneking, M., Barham, L., Lombard, M., Meyer, A., Steyn, M., Malmström, H., Rocha, J., Soodyall, H., Pakendorf, B., Bostoen, K., and Schlebusch, C. M. (2024). The genetic legacy of the expansion of Bantu-speaking peoples in Africa. *Nature*, 625(7995):540–547.
Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., et al. (2013). A revised timescale for human

evolution based on ancient mitochondrial genomes. *Current Biology*, 23(7):553–559.

Giles, R. E., Blanc, H., Cann, H. M., and Wallace, D. C. (1980). Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 77(11):6715–6719.

Goldberg, A. and Rosenberg, N. A. (2015). Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. *Genetics*, 201(1):263–279.

Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A., and Tishkoff, S. A. (2006). Whole-mtDNA Genome Sequence Analysis of Ancient African Lineages. *Molecular Biology and Evolution*, 24(3):757–768.

González, A. M., Larruga, J., Abu-Amero, K. K., Shi, Y., Pestano, J., and Cabrera, V. M. (2007). Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics*, 8:223.

Greeff, J. M. (2007). Deconstructing Jaco: Genetic Heritage of an Afrikaner. *Annals of Human Genetics*, 71(5):674–688.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722.

Greenberg, J. H. (1963). The Languages of Africa. *International Journal of American Linguistics*, 29(1).

Grollemund, R., Schoenbrun, D., and Vansina, J. (2023). Moving Histories: Bantu Language Expansions, Eclectic Economies, and Mobilities. *The Journal of African History*, 64(1):13–37.

Güldemann, T. (2014). *"Khoisan" linguistic classification today*, pages 1–41.

Güldemann, T. (2018). *The Languages and Linguistics of Africa*. De Gruyter Mouton, Berlin, Boston.

Hammarén, R., Goldstein, S. T., and Schlebusch, C. M. (2023). Eurasian back-migration into Northeast Africa was a complex and multifaceted process. *PLoS One*, 18(11):e0290423.

Heese, J. (1971). *Die herkoms van die Afrikaner, 1657-1867*. A. A. Balkema.

Heinz, T., Pala, M., Gómez-Carballa, A., Richards, M. B., and Salas, A. (2017). Updating the African human mitochondrial DNA tree: Relevance to forensic and population genetics. *Forensic Science International: Genetics*, 27:156–159.

Henn, B. M., Gignoux, C., Lin, A. A., Oefner, P. J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S. A., Mountain, J. L., and Underhill, P. A. (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A*, 105(31):10693–10698.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J., 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. *Proceedings of the National Academy of Sciences*, 108(13):5154–5162.

Henn, B. M., Steele, T. E., and Weaver, T. D. (2018). Clarifying distinct models of modern human origins in Africa. *Current Opinion in Genetics & Development*, 53:148–156.

Hofreiter, M. and Hartmann, S. (2020). Chapter Two - Reconstructing protein-coding sequences from ancient DNA. In Pelosi, P. and Knoll, W., editors, *Odorant Binding and Chemosensory Proteins*, volume 642 of *Methods in Enzymology*, pages 21–33. Academic Press.

Hollemeyer, K., Altmeyer, W., Heinzle, E., and Pitra, C. (2008). Species identification of Oetzi's clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests. *Rapid Communications in Mass Spectrometry*, 22(18):2751–2767.

Hollfelder, N., Babiker, H., Granehäll, L., Schlebusch, C. M., and Jakobsson, M. (2020). The genetic variation of lactase persistence alleles in Northeast Africa. *BioRxiv*, pages 2020–04.

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific data*, 7(1):1–11.

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, 546(7657):289–292.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197.

Huffman, T. N. (2007). *Handbook to the Iron Age*. University of KwaZulu-Natal Press.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003.

Jobling, M., Hollox, E. J., Hurles, M. E., Kivisild, T., and Tyler-Smith, C. (2014). *Human Evolutionary Genetics, second edition*. Garland Science.

Johanson, D. C. and White, T. D. (1979). A systematic assessment of early African hominids. *Science*, 203(4378):321–330.

Lankheet, I., Hammarén, R., Caballero, L. X. A., Larena, M., Malmström, H., Jolly, C., Soodyall, H., de Jongh, M., and Schlebusch, C. (2024). Wide-scale Geographical Analysis of Genetic Ancestry in the South African Coloured Population. *bioRxiv*.

Lankheet, I., Vicente, M., Barbieri, C., and Schlebusch, C. (2021). The performance of common SNP arrays in assigning African mitochondrial haplogroups. *BMC* *genomic data*, 22:1–8.

Lavachery, P. (2001). The Holocene archaeological sequence of Shum Laka rock shelter (Grassfields, western Cameroon). *African archaeological review*, 18(4):213–247.

Lawson, D. J., van Dorp, L., and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1):3258.

Le Meillour, L., Zirah, S., Zazzo, A., Cersoy, S., Détroit, F., Imalwa, E., Lebon, M., Nankela, A., Tombret, O., Pleurdeau, D., et al. (2020). Palaeoproteomics gives new insight into early southern African pastoralism. *Scientific reports*, 10(1):1–11.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Li, S., Schlebusch, C., and Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society B: Biological Sciences*, 281(1793):20141448.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–715.

Lipson, M., Ribot, I., Mallick, S., Rohland, N., Olalde, I., Adamski, N., Broomandkhoshbacht, N., Lawson, A. M., López, S., Oppenheimer, J., Stewardson, K., Asombang, R. N., Bocherens, H., Bradman, N., Culleton, B. J., Cornelissen, E., Crevecoeur, I., de Maret, P., Fomine, F. L. M., Lavachery, P., Mindzie, C. M., Orban, R., Sawchuk, E., Semal, P., Thomas, M. G., Van Neer, W., Veeramah, K. R., Kennett, D. J., Patterson, N., Hellenthal, G., Lalueza-Fox, C., MacEachern, S., Prendergast, M. E., and Reich, D. (2020). Ancient West African foragers in the context of African population history. *Nature*, 577(7792):665–670.

Lipson, M., Sawchuk, E. A., Thompson, J. C., Oppenheimer, J., Tryon, C. A., Ranhorn, K. L., de Luna, K. M., Sirak, K. A., Olalde, I., Ambrose, S. H., Arthur, J. W., Arthur, K. J. W., Ayodo, G., Bertacchi, A., Cerezo-Román, J. I., Culleton, B. J., Curtis, M. C., Davis, J., Gidna, A. O., Hanson, A., Kaliba, P., Katongo, M., Kwekason, A., Laird, M. F., Lewis, J., Mabulla, A. Z. P., Mapemba, F., Morris, A., Mudenda, G., Mwafulirwa, R., Mwangomba, D., Ndiema, E., Ogola, C., Schilt, F., Willoughby, P. R., Wright, D. K., Zipkin, A., Pinhasi, R., Kennett, D. J., Manthi, F. K., Rohland, N., Patterson, N., Reich, D., and Prendergast, M. E. (2022). Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature*, 603(7900):290–296.

Llorente, M. G., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*, 350(6262):820–822.

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614.

Lorna Marshall, M. B. (1986). *The Past and future of !Kung ethnography : critical reflections and symbolic perspectives : essays in honour of Lorna Marshall*. Helmut Buske Verlag, Hamburg.

Luo, S., Valencia, C. A., Zhang, J., Lee, N.-C., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., et al. (2018). Biparental inheritance of mitochondrial DNA in humans. *Proceedings of the National Academy of Sciences*, 115(51):13039–13044.

Maca-Meyer, N., González, A. M., Pestano, J., Flores, C., Larruga, J., and Cabrera, V. M. (2003). Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. *BMC Genetics*, 4(1):15.

Maier, P. A., Runfeldt, G., Estes, R. J., and Vilar, M. G. (2022). African mitochondrial haplogroup L7: a 100,000-year-old maternal human lineage discovered through reassessment and new sequencing. *Scientific Reports*, 12(1):10747.

Marshall, F. and Hildebrand, E. (2002). Cattle Before Crops: The Beginnings of Food Production in Africa. *Journal of World Prehistory*, 16(2):99–143.

Marshall, J. and Ritchie, C. (1984). *Where are the Ju/wasi of Nyae Nyae?: Changes in a Bushman society, 1958-1981.* Cape Town : Centre for African Studies, University of Cape Town.

Marshall, P. A. and Marshall, P. L. (2007). *Ethical challenges in study design and informed consent for health research in resource-poor settings*, volume 5. World Health Organization.

McCarthy, R. C. and Lucas, L. (2014). A morphometric re-assessment of BOU-VP-16/1 from Herto, Ethiopia. *J. Hum. Evol*, 74(11).

Mendez, F. L., Krahn, T., Schrack, B., Krahn, A.-M., Veeramah, K. R., Woerner, A. E., Fomine, F. L. M., Bradman, N., Thomas, M. G., Karafet, T. M., 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*, 92(3):454–459.

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

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. *Proceedings of the National Academy of Sciences*, 100(1):171–176.

Mitchell, P. and Lane, P. (2013). *The Oxford Handbook of African Archaeology*. Oxford University Press.

Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., and Stein, D. J. (2018). H3Africa: current perspectives. *Pharmgenomics Pers Med*, 11:59–66.

Naidoo, T., Schlebusch, C. M., Makkan, H., Patel, P., Mahabeer, R., Erasmus, J. C., and Soodyall, H. (2010). Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investigative genetics*, 1:1–11.

Naidoo, T., Xu, J., Vicente, M., Malmström, H., Soodyall, H., Jakobsson, M., and Schlebusch, C. M. (2020). Y-Chromosome Variation in Southern African Khoe-San Populations Based on Whole-Genome Sequences. *Genome Biology and Evolution*, 12(7):1031–1039.

Nurse, G. T. and Jenkins, T. (1977). Serogenetic studies on the Kavango peoples of South West Africa. *Ann Hum Biol*, 4(5):465–478.

Ostrom, P. H., Schall, M., Gandhi, H., Shen, T.-L., Hauschka, P. V., Strahler, J. R., and Gage, D. A. (2000). New strategies for characterizing ancient proteins using matrix-assisted laser desorption ionization mass spectrometry. *Geochimica et Cosmochimica Acta*, 64(6):1043–1050.

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Romero, I. G., Ayub, Q., Mehdi, S. Q., Thomas, M. G., Luiselli, D., Bekele, E., Bradman, N., Balding, D. J., and Tyler-Smith, C. (2012). Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. *The American Journal of Human Genetics*, 91(1):83–96.

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Van der Veen, L., Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., and Quintana-Murci, L. (2009). Inferring the Demographic History of African Farmers and Pygmy Hunter–Gatherers Using a Multilocus Resequencing Data Set. *PLOS Genetics*, 5(4):1–13.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., et al. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, 356(6337):543–546.

Patterson, N., Petersen, D. C., van der Ross, R. E., Sudoyo, H., Glashoff, R. H., Marzuki, S., Reich, D., and Hayes, V. M. (2010). Genetic structure of a unique admixed population: implications for medical research. *Human Molecular Genetics*, 19(3):411–419.

Pennarun, E., Kivisild, T., Metspalu, E., Metspalu, M., Reisberg, T., Moisan, J.-P., Behar, D. M., Jones, S. C., and Villems, R. (2012). Divorcing the Late Upper Palaeolithic demographic histories of mtDNA haplogroups M1 and U6 in Africa. *BMC Evolutionary Biology*, 12(1):234.

Pereira, L., Černý, V., Cerezo, M., Silva, N. M., Hájek, M., Vašíková, A., Kujanová, M., Brdička, R., and Salas, A. (2010). Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. *European Journal of Human Genetics*, 18(8):915–923.

Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M. J., and Amorim, A. (2001). Prehistoric and historic traces in the mtDNA of Mozambique: Insights into the Bantu expansions and the slave trade. *Annals of Human Genetics*, 65(5):439–458.

Peter, B. M. (2016). Admixture, Population Structure, and F-Statistics. *Genetics*, 202(4):1485–1501.

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R. A., Hannick, L. I., Glashoff, R. H., Mukerji, M., Fernandez, P., Haacke, W., Schork, N. J., and Hayes, V. M. (2013). Complex Patterns of Genomic Admixture within Southern Africa. *PLoS Genetics*, 9(3):e1003309.

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. *Nature communications*, 3(1):1–6.

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

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). A universal

SNP and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987.

Prendergast, M. E., Lipson, M., Sawchuk, E. A., Olalde, I., Ogola, C. A., Rohland, N., Sirak, K. A., Adamski, N., Bernardos, R., Broomandkhoshbacht, N., Callan, K., Culleton, B. J., Eccles, L., Harper, T. K., Lawson, A. M., Mah, M., Oppenheimer, J., Stewardson, K., Zalzala, F., Ambrose, S. H., Ayodo, G., Gates, H. L., Gidna, A. O., Katongo, M., Kwekason, A., Mabulla, A. Z. P., Mudenda, G. S., Ndiema, E. K., Nelson, C., Robertshaw, P., Kennett, D. J., Manthi, F. K., and Reich, D. (2019). Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*, 365(6448):eaaw6275.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

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. *The American Journal of Human Genetics*, 86(4):611–620.

Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S., Balanovsky, O., Kidd, K. K., Kidd, J. R., van der Veen, L., Hombert, J.-M., Gessain, A., Verdu, P., Froment, A., Bahuchet, S., Heyer, E., Dausset, J., Salas, A., and Behar, D. M. (2008). Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences*, 105(5):1596–1601.

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

Ramsøe, A., van Heekeren, V., Ponce, P., Fischer, R., Barnes, I., Speller, C., and Collins, M. J. (2020). DeamiDATE 1.0: Site-specific deamidation as a tool to assess authenticity of members of ancient proteomes. *Journal of Archaeological Science*, 115:105080.

Rando, J. C., Pinto, F., González, A. M., Hernández, M., Larruga, J. M., Cabrera, V. M., and Bandelt, H.-J. (1998). Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. *Annals of Human Genetics*, 62(6):531–550.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J.-J., Kelso, J., Slatkin, M., and Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*,

468(7327):1053–1060.
Ribot, I., Morris, A. G., Sealy, J., and Maggs, T. (2010). Population history and economic change in the last 2000 years in KwaZulu-Natal, RSA. *Southern African Humanities*, 22(1):89–112.
Rishishwar, L., Conley, A. B., Wigington, C. H., Wang, L., Valderrama-Aguirre, A., and King Jordan, I. (2015). Ancestry, admixture and fitness in Colombian genomes. *Scientific reports*, 5(1):12376.
Rito, T., Richards, M. B., Fernandes, V., Alshamali, F., Cerny, V., Pereira, L., and Soares, P. (2013). The first modern human dispersals across Africa. *PloS one*, 8(11):e80031.
Robbins, L., Muyzer, G., and Brew, K. (1993). Macromolecules from living and fossil biominerals. In *Organic geochemistry*, pages 799–816. Springer.
Robinson, N. E. and Robinson, A. B. (2001). Prediction of protein deamidation rates from primary and three-dimensional structure. *Proc Natl Acad Sci U S A*, 98(8):4367–4372.
Rosa, A. and Brehm, A. (2011). African human mtDNA phylogeography at-a-glance. *Journal of Anthropological Sciences*, 89:25–58.
Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.
Salas, A., Richards, M., De la Fe, T., Lareu, M.-V., Sobrino, B., Sánchez-Diz, P., Macaulay, V., and Carracedo, Á. (2002). The making of the African mtDNA landscape. *The American Journal of Human Genetics*, 71(5):1082–1111.
Salter-Townshend, M. and Myers, S. (2019). Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212(3):869–889.
Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467.
Sanger, F. and Thompson, E. (1953). The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 53(3):353.
Santiago, E., Novo, I., Pardiñas, A. F., Saura, M., Wang, J., and Caballero, A. (2020). Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium. *Molecular Biology and Evolution*, 37(12):3642–3653.
Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS one*, 7(3):e34131.
Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8):919–925.
Schlebusch, C. (2010). Issues raised by use of ethnic-group names in genome study. *Nature*, 464(7288):487–487.
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. *Journal of Human Henetics*, 56(9):623–630.
Schlebusch, C. M., Lombard, M., and Soodyall, H. (2013). MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. *BMC Evolutionary Biology*, 13(1):56.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., and Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358(6363):652–655.

Schlebusch, C. M., Sjödin, P., Breton, G., Günther, T., Naidoo, T., Hollfelder, N., Sjöstrand, A. E., Xu, J., Gattepaille, L. M., Vicente, M., Scofield, D. G., Malmström, H., De Jongh, M., Lombard, M., Soodyall, H., and Jakobsson, M. (2020). Khoe-San genomes reveal unique variation and confirm the deepest population divergence in Homo sapiens. *Molecular Biology and Evolution*, 37(10):2944–2954.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H., and Jakobsson, M. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338(6105):374–379.

Sealy, J. and Yates, R. (1994). *The chronology of the introduction of pastoralism to the Cape, South Africa*, volume 68. Cambridge University Press.

Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., Prista, A., Damasceno, A., Fehn, A.-M., Schlebusch, C., and Rocha, J. (2019). Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. *Molecular Biology and Evolution*, 37(2):406–416.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing prehistoric African population structure. *Cell*, 171(1):59–71.

Smith, C. I., Chamberlain, A. T., Riley, M. S., Stringer, C., and Collins, M. J. (2003). The thermal history of human fossils and the likelihood of successful DNA amplification. *J Hum Evol*, 45(3):203–217.

Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., Costa, M. D., Musilová, E., Macaulay, V., Richards, M. B., et al. (2012). The expansion of mtDNA haplogroup L3 within and out of Africa. *Molecular biology and evolution*, 29(3):915–927.

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M. B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *The American Journal of Human Genetics*, 84(6):740–759.

Solazzo, C., Fitzhugh, W. W., Rolando, C., and Tokarski, C. (2008). Identification of protein remains in archaeological potsherds by proteomics. *Analytical Chemistry*, 80(12):4590–4597.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*, 4(1):vey016.

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. (2007a). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular biology and evolution*, 24(10):2180–2195.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–1044.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., et al. (2007b). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, 39(1):31–40.

Tishkoff, S. A. and Verrelli, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual review of genomics and human genetics*, 4(1):293–340.

Underhill, P. A. and Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.*, 41(1):539–564.

van de Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., Hajraoui, M. A. E., Amzazi, S., Hublin, J.-J., Pääbo, S., Schiffels, S., Meyer, M., Haak, W., Jeong, C., and Krause, J. (2018). Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. *Science*, 360(6388):548–552.

Van de Velde, M., Bostoen, K., Nurse, D., and Philippson, G. (2019). *The Bantu Languages*. Routledge, second edition.

Van der Auwera, G. A. and O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

van der Valk, T., Pečnerová, P., Díez-del Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, 591(7849):265–269.

van Oven, M. and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation*, 30(2):E386–E394.

Veeramah, K. R., Wegmann, D., Woerner, A., Mendez, F. L., Watkins, J. C., Destro-Bisol, G., Soodyall, H., Louie, L., and Hammer, M. F. (2012). An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol Biol Evol*, 29(2):617–630.

Vicente, M., Jakobsson, M., Ebbesen, P., and Schlebusch, C. M. (2019). Genetic affinities among southern Africa hunter-gatherers and the impact of admixing farmer and herder populations. *Molecular biology and evolution*, 36(9):1849–1861.

Vicente, M., Lankheet, I., Russell, T., Hollfelder, N., Coetzee, V., Soodyall, H., Jongh, M. D., and Schlebusch, C. M. (2021). Male-biased migration from East Africa introduced pastoralism into southern Africa. *BMC Biology*, 19:1–16.

Vicente, M. and Schlebusch, C. M. (2020). African population history: an ancient DNA perspective. *Current Opinion in Genetics & Development*, 62:8–15.

Vidal, C. M., Lane, C. S., Asrat, A., Barfod, D. N., Mark, D. F., Tomlinson, E. L., Tadesse, A. Z., Yirgu, G., Deino, A., Hutchison, W., Mounier, A., and

Oppenheimer, C. (2022). Age of the oldest known Homo sapiens from eastern Africa. *Nature*, 601(7894):579–583.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. (1991). African Populations and the Evolution of Human Mitochondrial DNA. *Science*, 253(5027):1503–1507.

Vossen, R. H. A. M. and Buermans, H. P. J. (2017). *Full-Length Mitochondrial-DNA Sequencing on the PacBio RSII*, pages 179–184. Springer New York, New York, NY.

Wall, J. D. and Hammer, M. F. (2006). Archaic admixture in the human genome. *Current Opinion in Genetics and Development*, 16(6):606–610. Genomes and evolution.

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., et al. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nature genetics*, 46(4):336–344.

Watson, E., Forster, P., Richards, M., and Bandelt, H. J. (1997). Mitochondrial footprints of human expansions in Africa. *American Journal of Human Genetics*.

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H. J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic acids research*, 44(W1):W58–W63.

Welker, F., Collins, M. J., Thomas, J. A., Wadsley, M., Brace, S., Cappellini, E., Turvey, S. T., Reguero, M., Gelfo, J. N., Kramarz, A., et al. (2015). Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature*, 522(7554):81–84.

Welker, F., Hajdinjak, M., Talamo, S., Jaouen, K., Dannemann, M., David, F., Julien, M., Meyer, M., Kelso, J., Barnes, I., et al. (2016). Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proceedings of the National Academy of Sciences*, 113(40):11162–11167.

Welker, F., Ramos-Madrigal, J., Gutenbrunner, P., Mackie, M., Tiwary, S., Rakownikow Jersie-Christensen, R., Chiva, C., Dickinson, M. R., Kuhlwilm, M., de Manuel, M., et al. (2020). The dental proteome of Homo antecessor. *Nature*, 580(7802):235–238.

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G., and Clark Howell, F. (2003). Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature*, 423(6941):742–747.

Wood, E. T., Stover, D. A., Ehret, C., Destro-Bisol, G., Spedini, G., McLeod, H., Louie, L., Bamshad, M., Strassmann, B. I., Soodyall, H., and Hammer, M. F. (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European Journal of Human Genetics*, 13(7):867–876.

Y Chromosome, C. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, 12(2):339–348.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2446

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-537831

ACTA UNIVERSITATIS
UPSALIENSIS
2024