The origins and spread of the Neolithic in the Old World using Ancient Genomes

UNIVERSITY OF CAMBRIDGE

Marcos Gallego Llorente

Selwyn College

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2017

Author's Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma, or other qualification at the University of Cambridge or any other University or similar institution. The main text does not exceed 60,000 words.

Summary

One of the biggest innovations in human prehistory was the advent of food production, consisting of the ability to grow crops and domesticate animals for consumption. This wide-scale transition from hunting and gathering to food production led to more permanent settlements, and set in motion major societal changes. In western Eurasia, this revolution spread from the Near East into Europe, Africa and diverse regions of Asia.

Agriculture was brought into Europe by the descendants of early Anatolian farmers starting approximately 8,000 years ago. But little was known of the people who developed agriculture in the Fertile Crescent: where they all closely related to the early Anatolian farmers, or were there multiple ethnic groups who developed agriculture in parallel? In the first data chapter, I use the first genome from a Neolithic woman from Ganj Dareh, in the Zagros Mountains (Iran), a site with evidence of early goat domestication 10,000 years ago. I showed that Western Iran wan inhabited by populations mostly similar to Hunter-gatherer populations from the Caucasus, but remarkably, very distinct from the Anatolian farmers who spread the Neolithic package into Europe. While a degree of cultural diffusion between Anatolia, Mesopotamia and the Zagros highlands likely happened, genetic dissimilarity supports a model in which Neolithic societies of that area were distinct.

The second chapter deals with how Africa was affected by population movements, originating in the Near East, during the Neolithic times. Characterising genetic diversity in Africa is a crucial step for analyses reconstructing human evolution. Using Mota, an ancient genome from a male from the Ethiopian highlands, I showed a backflow into Africa by populations closely related to the Anatolian Neolithic farmers.

The third chapter deals with some common problems and themes in the analysis of ancient DNA, such as merging capture datasets with diverse number of ascertained SNPs, combining capture and shotgun data in the same analysis, and the effect of UDG treatment in ancient samples. I describe the most common problems and their effect in summary statistics, and propose a guide on how to work with ancient DNA to avoid data compatibility problems.

Acknowledgements

The past four years have been, without any doubt, the most enriching and exciting years of my life. I have developed myself in ways I never thought would have been possible, and that has been thanks to the people that have been around me, and have provided me with the support, trust, company, and happiness that I needed to complete my PhD.

First and foremost, my thanks go to Andrea Manica, by far the best supervisor I could have ever wished for. His availability, his trust, his support in the good moments and in the difficult ones, his enthusiasm, and his vast knowledge have always been there. He has undoubtedly been the main pillar of my time in the office. I have learned so much from him, academically, professionally and personally, that he has been an inspiration for my life, and definitely will continue to be so. Thank you so much for everything, Andrea.

The Evolutionary Ecology Group has been my home for the last four years, and I am grateful to have had such a great group of people in which to trust, rely upon, and generally expect the best of times while in the office and outside. Anders, Robert, Riva, Veronika, Eppie, Liisa, Pier Paolo, Marius, Tommy, Anne Sophie, Mario: it was great to work in such a great environment. Robert especially: I won't forget our late evenings in the office discussing whatever random philosophical aspect of life while doing work. Also, those sunsets in the old office: unforgettable. Anders, Veronika and Eppie also deserve special mentions for helping me out whenever I needed a bit of bioinformatical advice – thanks so much. And thanks Riva for our occasional deep talks over coffee breaks, they were often very much needed. You all have been inspirations in one way or another, and have definitely shaped my experience here in a great way.

Life in Cambridge, however, wouldn't have been the same without my closest friends from Selwyn. My most special and heartfelt thanks go to Michał, Jessica, Alex and Max. You have been my home. Our evenings together, our unforgettable MCR dinners, our trip to Marrakech, our summery and wintery days out, and, in short, the discovery of Cambridge we've made together, have constituted the memories I'll treasure the most of my time here.

To everybody else that I've had the pleasure to call a friend: Pablo, Nathalie, Max, Melissa, Harvey, Emmy, Afnan, Lucía, Ben, Luci, Swati, Farid, Isabelle, Leanne, James, Julia, Raj, Fedir, Eva: I am glad I have had the pleasure to share my Cambridge experience with you.

To the Selwyn College Boatclub: for helping me overcome every challenge I never thought would be possible, for the early winter mornings, the sunny summer evenings, the bumps, the 2016 blades, and for making me proud of being part of such a great group of people. Nick, Nigel, Sam, Harry, George, Will, DBR, JPG, Charlie & Charlie, and Felix: Been a pleasure.

I would like to thank Selwyn College for being my home away from home all these years, and for helping me attend the SMBE 2016 in Gold Coast, Australia, where I presented part of this thesis. In addition, I would like to thank the BBSRC for the funding needed for this PhD.

During my PhD and throughout all my education, I have been incredibly lucky to receive the trust and support from my parents, my numerous siblings, and my family: They have been such a great motivation that has allowed me to follow my dreams and aspirations. Muchísimas gracias por todo.

Y por último, a mis amigos de toda la vida: Manu, Dani, Ricardo, Diego, Laus, Marina, Gabriel, Christian: no olvidaré nuestras tardes de navidades y nuestras noches de verano. Ojalá nunca me dejéis olvidar de dónde vengo. Ha sido un placer.

Contents

1. Introduction

1.1. Bioanthropology before the advent of palaeogenetics

The emergence and expansion of Anatomically Modern Humans

Anatomically Modern Humans (AMHs) constitute all the humans in today's world: members of the species *Homo sapiens*, with an osteological and behavioural modernity that encompasses the entirety of modern phenotypic variation.

The first defining feature of AMHs in the fossil record is osteological modernity. The human archaeological record fully supports a slow, gradual accumulation of the osteological features that define AMHs, which happened more than 200k years ago (Bräuer, 2008; McDougall et al., 2005). This fossil record equally suggests that these modern morphological features evolved in Africa, while the only inhabitants of Eurasia were the Neanderthals and other archaic humans (Pearson, 2008; Weaver, 2012). There are fossils in Ethiopia (the Herto and Omo skulls), dated to 195k and 160k years ago, which start to resemble the form that AMHs would eventually acquire, although there is still a hint of robustness resembling more archaic features. There is a temporal gap between the origin of the fully anatomically modern form and the eventually successful expansion out-of-Africa between 100k and 60k years ago. This has been subject to an intense paleoanthropological debate, and it is a study area where future climate reconstructions and genetic studies will undoubtedly shed some light in the future.

The second defining feature of all current humans is our current behavioural characteristics, sometimes labelled as *behavioural modernity*: an accumulation of a series of changes in cognition and behaviour, which include language, abstract thinking, symbolic behaviour (art, music and rites), deep planning, and the usage of inanimate natural resources for daily tasks such as hunting megafauna and toolmaking. Some of these characteristics might however been already present in more archaic humans, such as language (D'Anastasio et al., 2013). Although this suite of developments is evident from 60k years ago, Richard Klein in 1995 argued that these changes that accumulated in order to give rise to AMHs were very gradual (Klein, 1995). This work was however criticised

by John Shea, who criticises this concept of behavioural modernity in a context of the strategic underpinnings of human behavioural variability in Palaeolithic Archaeology (Shea, 2011). Prior to the successful out-of-Africa expansions, the earliest forms of AMHs in Africa were behaviourally indistinguishable from their more archaic, Eurasian contemporaries. Strikingly, the major behavioural differences between AMHs and archaic humans them appeared only around the period between 100k and 60k years ago. These changes include growth and formal standardization of artefacts ("points", "needles", etc.), appearance of art, evidence of spatial organisation of camp surfaces, evidence for transport of large quantities of "raw" materials such as stones and bones, evidence of ritual art and elaborate graves, and evidence for fishing with tools and other advances in human ability to use nature (Klein, 1989; Mellars, 1989; Stringer and Gamble, 1993). The most economical explanation for this rapid change is the final modernisation of the brain, which has been argued that allowed for a collection of developments to be rapidly acquired. These developments, generally related to full behavioural modernity, would then open the door for the end of a strong dependence of climatic factors or food availability by AMHs, thereby allowing for successful expansions into Eurasia (Klein, 1995).

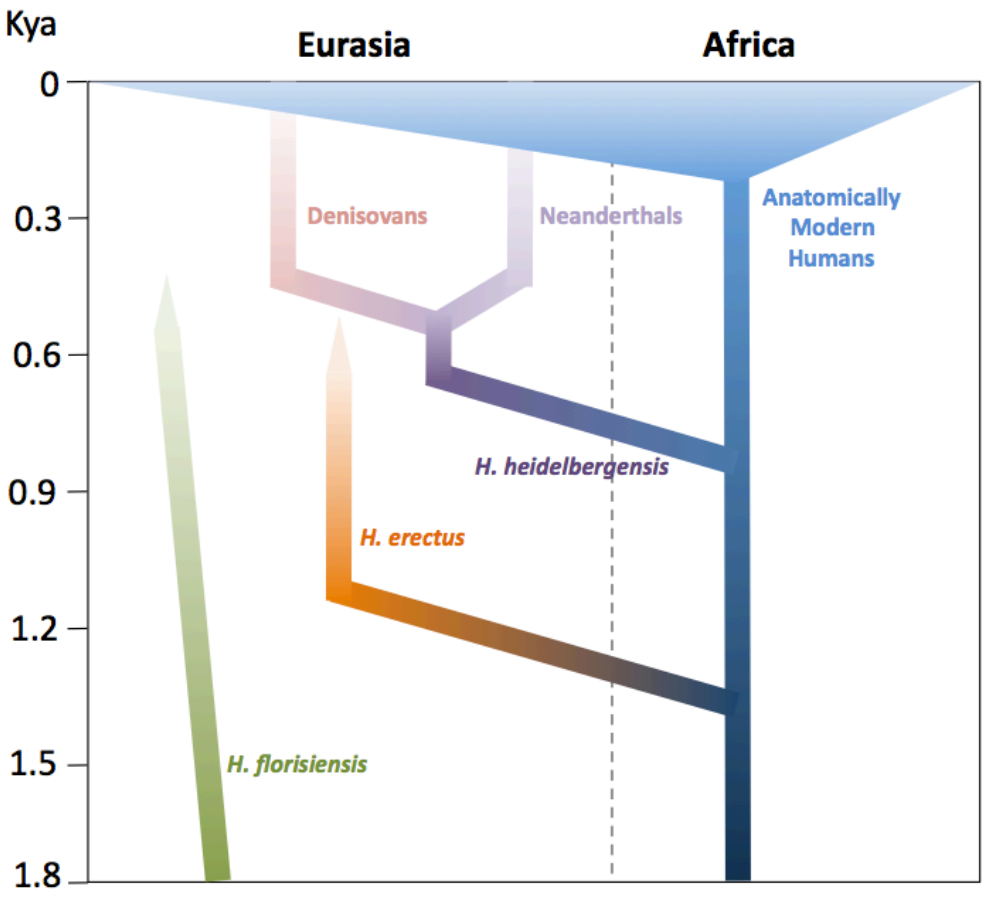


Fig. 1. While the *Homo* genus has evolved in Africa, there have been various episodes through which Homo populations have left Africa and established themselves in Eurasia. *Homo florisiensis*, a species of a yet undiscovered ancestry, was found in the Flores Island in Indonesia. Furthermore, *Homo erectus* left Africa > 2 million years ago, as did the archaic human subspecies that gave rise to Neanderthals and Denisovans. These out-of-Africa episodes, with their associated population bottlenecks, have been repeated throughout hominin history, with striking similarity.

Most authors argue that *Homo habilis* evolved into *Homo erectus* around 1.8m years ago. The partially-incomplete samples of *Homo erectus* found in Lake Turkana (Kenya) include lower jaws, a near-complete but fragmented skull, and several cranial bones (Rightmire, 1992, 1993; Wood, 1992), featuring for the first time a large brain size, a larger body size, and probably regular bipedalism, in contrast to a mix of bipedalism and tree climbing seen in earlier species (McHenry, 1992; Spoor et al., 1994). These developments allowed ancient humans to expand their geographical area of presence, which eventually created the oldest safe evidence of ancient human presence in Eurasia, the Dmanisi site in Georgia,

1.81m years ago (Gabunia et al., 2000), and various sites in the Jordan Valley, 1.4-1.3m years ago (Bar-Yosef, 1994).

By 200k years ago, the ancient hominins who left Africa had evolved in distinct fashions around the world: Europe and Western Eurasia was inhabited by the Neanderthals, a species that looked remarkably distinct from the then almost-Anatomically Modern Humans of Africa (Arsuaga et al., 1994, 1993; Stringer and Gamble, 1993).

There is paleoanthropological evidence that present human population is the outcome of a large demic expansion that began between 60k and 100k years ago out of Africa, and resulted in a rapid human occupation of almost all of the habitable areas of the planet. Before the advent of genetic studies, different forms of the Out-of-Africa theories advocated for different ways that AMHs expanded from Africa and replaced the Neanderthals (and the equally archaic Denisovans, discovered through ancient DNA and described in section 3 of the introduction). At the time, some theories supported some gene flow between expanding humans and resident archaic humans, while others did not (Smith, 1994).

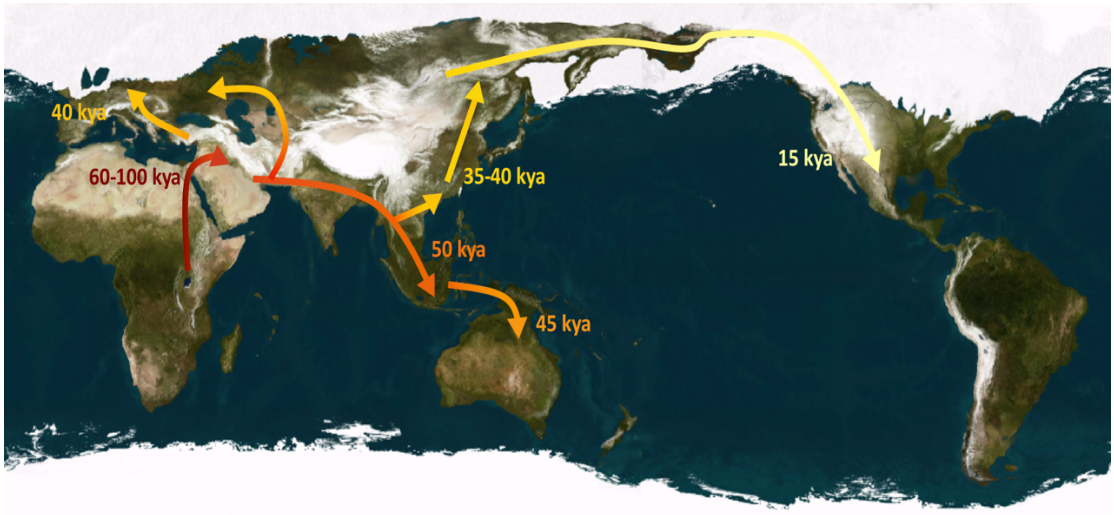


Fig. 2. The spread of Anatomically Modern Humans throughout Eurasia and the Americas. Once Anatomically Modern Humans left Africa (60-100k years ago), they quickly spread around the Indian Ocean via the Middle East and India into the Sahul, around 50k years ago, in a West-to-East axis. Subsequently, humans made their way into Northern Eurasia, reaching Europe around 40k years ago, Siberia around 25k years ago, and the Americas around 15k years ago. *In the figure, kya = thousands of years ago.*

The Advent of the Neolithic

Anatomically Modern Humans, therefore, have occupied the entirety of the inhabitable Old World for 40,000 years; and for almost all of that period, they lived as foragers, hunters and gatherers. Around 12,000 years ago, as the global temperatures began to increase at the end of the Pleistocene (the Ice Age), and the transition to the modern Holocene happened, humans developed, in a number of geographically-independent areas, a series of agricultural systems based on plant cultivation and sometimes domesticated animal farming (Barker, 2006).

The first component of this Neolithic Revolution probably happened in areas where the wild ancestors of the modern domesticates were naturally found. Wild forms of wheat and barley, for example, thrived in the eastern Mediterranean and the Mesopotamian valleys. Additionally, wild goats and sheep probably ranged across that same geographical area too. However, while authors such as Diamond and Bellwood (2003) argue for a revolutionary change that gave great demographic and cultural to those societies that first developed the Neolithic, other authors such as Dorian Fuller argue for a more protracted process, which happened through individual domestication events, coupled to different enabling technologies in each region (Fuller et al., 2015).

Secondly, the Urban Revolution that accompanied the Neolithic Revolution occurred in the same region, as the two most ancient known civilizations in History are the Sumerian in Mesopotamia, around 5,500 years ago; the Indus valley civilisation, starting around 5,300 years ago; and the Egyptians in the Nile valley, which started around 5,000 years ago (Breasted, 1916). Both the Sumerians and the Egyptians were civilizations whose food systems were based on wheat and barley, watered by the yearly flooding cycles of the Tigris, Euphrates and Nile, respectively; and complemented by sheep and goat herding.

Fig. 3. River Valley civilizations in the Old World. The Neolithic revolution was mostly developed in river valleys in regions of warm, seasonal climates, which allowed for a constant arrival of fresh water, the cultivation of seasonal crops such as wheat and rice, and the domestication of the first animals. It also therefore led to the first permanent settlements of human history.

The third component of the Neolithic Revolution is concerned with the changes in climate at the end of the Pleistocene. As the European ice sheets melted, the rain-bearing atmospheric depressions shifted northwards, away from the river depressions of the Nile and Mesopotamian river valleys. As the grasslands were probably substituted by deserts and oases, watered regularly by river flooding events, different human groups started to congregate around those areas, seeding what would then become the first permanent settlements (Childe, 1934).

Excavations from what is now northern Israel, in 1937, revealed that hunter-gatherers from the early Holocene had developed flint tools and sickles of flint blades which show signs of having been used for cutting grass of cereal stems. This therefore is the first evidence of a group of people starting to become agriculturally based (Garrod and Bate, 1937).

Spread of the Neolithic: Cultural vs demic diffusions?

In the Late Pleistocene, large areas of Europe were abandoned by people, the northern half of Britain was covered by ice, and European Upper Paleolithic hunter-gatherers lived in so-called "refugia", such as south-west France, Spain, Italy and parts of the Balkans (Soffer and Gamble, 1990; Straus et al., 1996; Street and Terberger, 1999). As soon as the ice sheets retreated, between 14,700 and 12,800 years ago, Europe was quickly repopulated by hunter-gatherers (Charles, 1996; Housley et al., 1997; Jochim, 1998).

The earliest archaeological indications of agriculture and pastoralism in the Mediterranean come from Cyprus, around 10,500-10,000 years ago (Peltenburg et al., 2000). The material artefacts found there are very similar to those in Early Neolithic in the Near East and follow the same pattern of reliance on wheat, barley, livestock and legumes. It is likely that these Levantine farmers developed over that time an expanded knowledge of seafaring (Peltenburg et al., 2000). The Neolithic seems to have arrived in Crete around 8,800 years ago, and perhaps the first humans to ever live there (Broodbank and Strasser, 1991). Settlements in Greece, contemporaneous with the peopling of Crete, share many similarities with the villages of Anatolian Agriculturalists in terms of settlement layout, construction materials and artefacts found. This suggests that the main expansions of the Neolithic culture into Europe came from Anatolian Agriculturalists (Perlès, 2001).

Fig. 4. The expansion of the Neolithic into Europe. Isochrone map of the expansion of the Neolithic into Europe. *[Figure adapted from Burger et al., 2012].*

By 8,500 years ago there was already farming in the Morava and Danube valleys through northern Serbia. These societies, like those of the Mediterranean basin at the time, lived through a combination of farming and foraging. Their material culture was however much more simple than that of the more southern Mediterranean farmers (Chapman, 2003; Kosse, 1979). The traditional archaeological interpretation of these settlements is that agricultural colonists from Greece and Anatolia moved into the Balkan river valleys, together with the Mediterranean system of farming and their artefacts. However, the incoming agricultural advances are found side by side with evidence of more traditional foraging and waterscape fishing. This, very likely, resulted from the mixing between incoming Neolithic populations and the local hunter-gatherers (Borić and Stefanović, 2004). It was only around 7,000-6,000 years ago that the commitment to agriculture rapidly developed in these societies, as well as the organisation of their settlements, the first multi-roomed houses, etc. Was it because of more waves of inhabitants from the southern fringes of Europe, were these developments the product of knowledge diffusion after the first Neolithic wave, or was there a second, more successful wave of agriculture acquisition in Europe?

In more northern Europe, the agriculture border had stayed within the area around the Danube valleys for longer than a thousand years. It was only with this "second wave of agriculture acquisition", 7,000-6,000 years ago that the borders were drastically pushed northwards. This is also coupled with the rapid appearance of archaeological sites displaying the Early Neolithic *Linearbandkeramik* pottery (LBK) (Barker, 2006). This coupling of an improvement of agricultural techniques, a modernisation of pottery and artefacts, and an expansion of agriculture and farming beyond the borders of the Danube valley into the northern European plains suggests that there might have been a definite demographic push towards the North. This question was left unanswered by archaeology, until the advent of genetic studies on ancient, archaeological samples.

The Neolithic in Europe therefore did not arrive as an inexorable, constant spread of people and technology from the Near East to Europe via Anatolia and the Balkans. The picture is a rather complex one: communities with agriculture shifted into new territories, just as, a few millennia before, foragers had done from their ice age refugia into the depths of continental Europe (Rowley-Conwy, 2004). These movements illustrate a theme that has been repeated numerous times, and which defines human prehistory as a collection of repeated expansion events with different focal centres and directions. Each expansion event has brought palpable cultural consequences that shaped early human societies (Haak et al., 2005; Richards et al., 1996).

1.2. Ancient DNA

The path to Ancient DNA

It has always been a goal of human genetics to describe human evolution, expansions across the world, and ancient history with current-day genetic variation. The usage of reduced genetic information and a small number of loci started as early as in 1964, when Cavalli-Sforza modelled a phylogenetic tree of 15 modern human populations using 5 loci and 20 alleles, mostly blood groups (Cavalli-Sforza et al., 1964). Wright's F_{ST}, gene frequencies using blood polymorphisms, and various genetic distances were for decades the only tool to describe the main patterns of population differentiation (Cavalli-Sforza et al., 1994).

Modern and ancient data availability, however, has increased in a very fast fashion. Since the early 1980s, human mitochondrial DNA variation started to be characterised (Denaro et al., 1981). Its exclusively maternal inheritance, its relatively fast mutation rate, its lack of recombination, and its high copy number per cell, has resulted in the possibility to describe genealogical relationships between whole populations, at both continental and local scales (Hutchison et al., 1974; Merriwether et al., 1991). The Mitochondrial Eve, or the MRCA (Most Recent Common Ancestor) was placed at around 200k years ago (Kivisild et al., 2006), which means that although the most distant early stages of human evolutionary history are lost in the mtDNA record, we can very easily use mtDNA variation to discern the details of the colonization process of the Old World using regional patterns of variation; mtDNA variation was a key application in the era before nuclear genome sequencing. In the mid-1990s, different mtDNA haplogroups based on certain key mutations were starting to be labelled: A-G assigned to Asian and American lineages, H-K to Europe, and L to Africa (Torroni et al., 1994, 1993). It was later shown that mtDNA differences among populations in Africa were the highest, while Native Americans had the lowest (Lippold et al., 2014), consistent with the root of mtDNA phylogeny and the most diverse branches being located in Africa.

The first human genomic sequences were obtained in 2001 (Lander et al., 2001; Venter et al., 2001). Back then, to think that within 15 years, thousands of genomes from people around the globe would have been sequenced, was almost unimaginable (The 1000

Genomes Project Consortium, 2015). It was hence unthinkable that technology would allow us to access genomic information from thousands-years-old human fossils (Allentoft et al., 2015; Haak et al., 2015; Rasmussen et al., 2010b). Not only have anatomically modern humans been sequenced, but also Neanderthals (Green et al., 2010; Prüfer et al., 2014), and even Denisovans, a sister group of Neanderthals only discovered upon DNA sequencing (Reich et al., 2010; Meyer et al., 2012a). The sequencing depth of some of these genomes is remarkable: the genotyping error rates are almost equal to those of high-coverage sequences from modern genomes.

The analysis of ancient DNA has been successful in answering a great deal of unanswered questions, while raising new questions in the process. When and from which human populations did particular populations arise? Which populations admixed and when? Are stark changes in the archaeological record the result of quick cultural innovation, or population sweeps? Which past cultures left descendants in the present world?

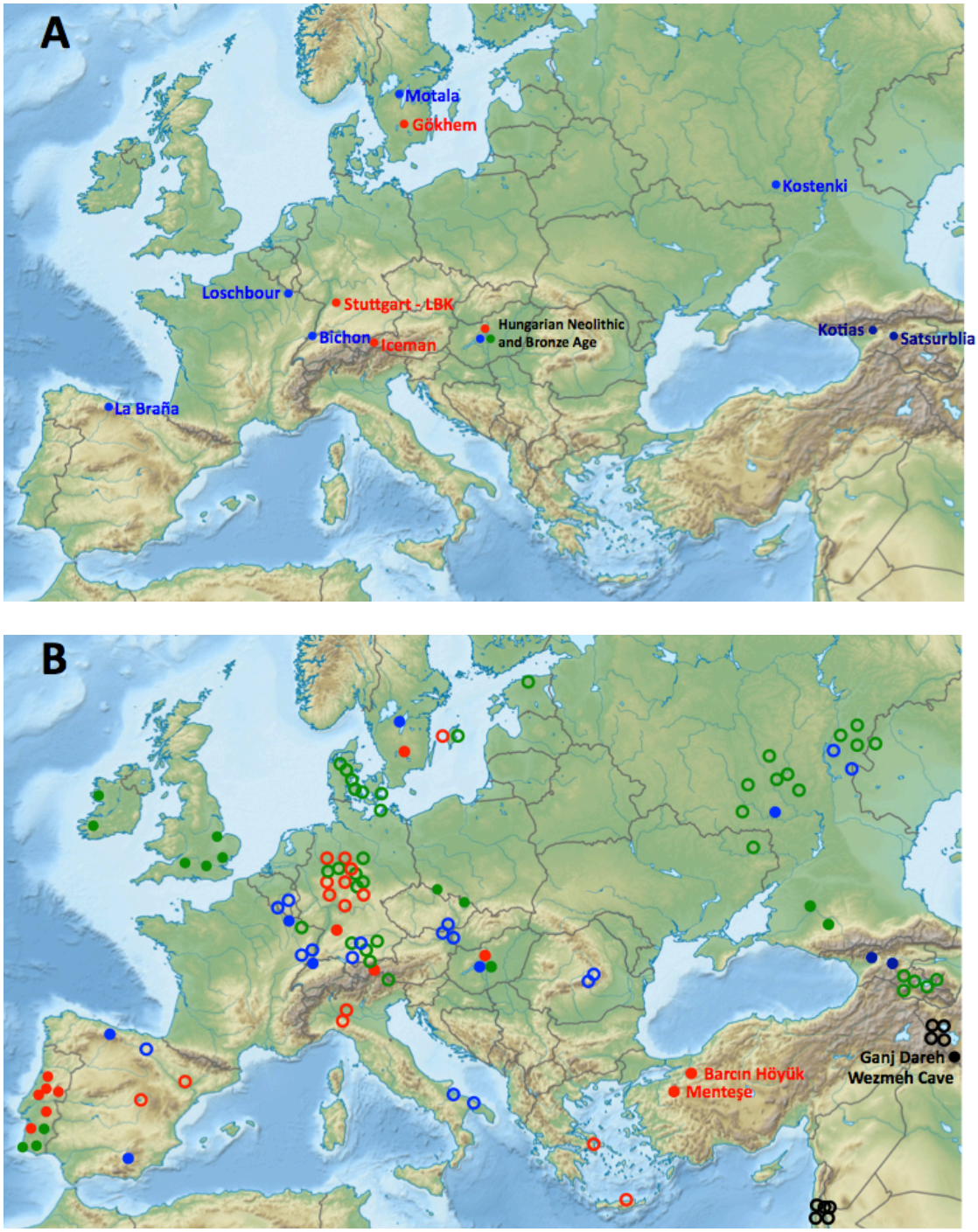


Fig. 5: Changes in the landscape of ancient genomes in Europe and the Near East, from 2014 to 2017. **A)** represents the whole genomes sequenced by 2014, with the names by which they are commonly known. **B)** represents the extent of ancient genomes obtained by SNP capture (empty circles) and whole genome sequencing (filled circles), since 2014. Colours indicate a rough idea of the ages of each sample: **blue** represents Palaeolithic samples, **red** represents Neolithic samples, **green** represents bronze age samples, and **black** represents near eastern samples of diverse epochs.

Challenges and ancient DNA work

One of the main challenges with working with aDNA has been its accessibility, due to very low endogenous DNA presence in human fossils in archaeological contests (García-Garcerà et al., 2011; Sánchez-Quinto et al., 2012b; Skoglund et al., 2012). A major breakthrough came in 2014, when Gamba et al. (2014) compared endogenous DNA from the temporal bone (specifically, its petrous portion, the densest bone in the mammalian body (Lam et al., 1999)), against fossils of other skeletal parts from six different human individuals from Hungary, at different time depths. They found that the endogenous DNA yields from the petrous portion of the temporal bone was 4-16 times higher than those from the teeth, and up to 183 times higher than those from other skeletal bones.

Another major hurdle in the ability to successfully sequence aDNA for several years was contamination. Genetic material from a recent tissue sample mostly consists of endogenous DNA (i.e. fragments of DNA from that individual). However, aDNA is very fragmented and scarce, which means that most of the genetic information found in ancient human bones tends to be exogenous DNA (i.e. from other humans who have been in contact with the fossil, or from bacteria and fungi from the environment) (Green et al., 2009). Contamination from modern human DNA is particularly problematic, as the sequences look similar to endogenous aDNA and can very easily introduce unwelcome biases (Green et al., 2006; Wall and Kim, 2007).

Strict aseptic room conditions are now standard practice in aDNA extraction: bleaching surfaces, UV radiation, and filtered refrigeration (Green et al., 2009). At the time of DNA extraction, DNA molecules are tagged so that further contamination is detected and eliminated (Briggs et al., 2007).

After DNA sequencing, several bioinformatic tools are used to remove exogenous reads and to estimate the proportion of endogenous DNA in the library. This is often done by estimating the rate of exogenous DNA in the mitochondrial DNA, which is much more abundant than nuclear DNA, and therefore sequenced to a much higher coverage (Renaud et al., 2015). A second advantage of mitochondrial DNA is the fact that it's a haploid marker, as heterozygote positions can easily be interpreted as contamination (Rasmussen et al., 2011a).

1.3. Understanding human prehistory with ancient DNA

Ancient DNA and the out-of-Africa expansions

Around the period between 100,000 and 60,000 years ago, a very rapid expansion of anatomically modern humans occurred outside of Africa, and spread quickly in all directions across the Eurasian landmass, eventually reaching nearly all inhabitable areas of the world. There are still many open questions surrounding this Great Expansion: The exact location of the population or populations that left Africa, the timing of the coastal migration around the Indian Ocean, the timing of the expansion into Northern Eurasia, the number of waves across this landmass, etc. (Macaulay et al., 2005; Oppenheimer, 2012; Rasmussen et al., 2011a).

Ancient genomes have, however, deepened our understanding of many other issues regarding this Great Expansion: a rough timing of the exit, the size of the population bottleneck associated with the event, and the general mode of the subsequent expansion.

However, there were near-AMH groups in the Near East around 130-80k years ago, as Dorothy Garrod showed in the 1930s, where a series of caves in present-day Israel showed a long sequence of Lower Palaeolithic, Middle Palaeolithic, and Epipalaeolithic, occupations (Garrod, 1937). This region was at that time ecologically very similar to north-eastern Africa (Bar-Yosef, 2000; Klein, 1989). It is likely that this primary expansion of near-AMHs was unsuccessful in persisting, and was subsequently replaced by Neanderthals during the following glacial period, starting 80k years ago. Afterwards, 50k years ago, there is evidence that AMHs were present in the Near East, and only after this point did they become widespread in Eurasia (Mellars, 2006). Genomic data from contemporary humans suggest that this expansion was accompanied by a continuous loss of genetic diversity, a result of what is called the “serial founder effect” (Manica et al., 2007; Prugnolle et al., 2005), as demonstrated by recent studies in autosomal polymorphisms and blood groups (Li and Durbin, 2011; McEvoy et al., 2011).

Ancient DNA in the admixture with other archaic humans

As the sequencing of ancient nuclear DNA and mtDNA became possible, one of the most sought-after questions was to understand the emergence of Anatomically Modern Humans in the context of all ancient human groups present at the time, and whether there was any admixture between AMHs and different archaic human populations, as AMHs expanded out of Africa.

The archaeological record clearly suggested that several groups of archaic hominins overlapped in time and space with this main out-of-Africa migration of AMHs (Higham et al., 2014). Furthermore, using the recently-sequenced Neanderthal genomes, it was shown that there was a certain degree of interbreeding: non-African AMHs have between 1.5 and 4% of Neanderthal DNA, inherited from an interbreeding episode around 50,000 years ago (Green et al., 2010; Prüfer et al., 2014; Sankararaman et al., 2014; Wall et al., 2013).

However, at the same time, there was a group of ancient humans in eastern Eurasia, which were only discovered in 2010 through genetic analysis, which also shows a large degree of divergence from the Neanderthals (Meyer et al., 2012a). These ancient humans, called Denisovans (after they were first found in the Denisova cave in the Altai mountains, Siberia) morphologically shared some characteristics with Neanderthals, while they also had some archaic features linking them back to *Homo erectus*. These Denisovans, after genetic analysis, were shown to have likely diverged from the Neanderthals as their ancestral population left Africa, later shown to be between 430k and 473k years ago, which has been shown via genome analysis (Prüfer et al., 2014), and archaeological analysis (Arsuaga et al., 2014). The split between this common ancestor of Neanderthal and Denisovans with Modern Humans was determined to be around 550k and 760k years ago (Meyer et al., 2016). Australian aboriginal populations and Papua-New-Guineans additionally have around 6-7% of Denisovan DNA (Lowery et al., 2013; Meyer et al., 2012a), while it has also been shown that modern East Asians and Native Americans also have small percentages of Denisovan ancestry in their DNA (Qin and Stoneking, 2015).

Ancient DNA in the Neolithic and its spread

Western Eurasia has had a complex history, an aspect mostly due to its complex geography, climatic changes, and different ecological areas. Also, in the past few years, this region has produced a very large number of ancient DNA genomes. The first few ancient genomes to be sequenced from Europe were from the early Neolithic times. Ötzi, a mummy from 5,300 years ago, found in the Tyrolean Alps, , showed an unexpected relationship with modern-day Sardinians, and, to a certain extent, current-day Middle Eastern populations (Keller et al., 2012; Sikora et al., 2014). Furthermore, Scandinavia revealed genomes from two drastically different ancestry sources: a 5,000 year old sample belonging to a Neolithic farmer in southern Sweden also showed the same genetic similarities to Sardinians. However, contemporaneous samples found in contexts that related them to hunter-gatherer lifestyles (then termed Scandinavian Hunter-Gatherers) did not show that relationship (Sikora et al., 2014; Skoglund et al., 2012).

This led the scientific community to believe that a very large demographic shift must have happened in Europe between 5,000 and 10,000 years ago, through which Europe, populated by established communities of hunter-gatherer populations, was affected by a large expansion of farmers from the south-eastern fringes of the continent, which drastically reshaped the European genetic landscape.

This was then confirmed by Lazaridis et al., (2014), who sequenced an 8,000-year-old original European hunter-gatherer (Loschbour, in Luxembourg), which showed relatedness to the previously-described hunter-gatherer individuals, and a 7,500-year-old Neolithic sample (Stuttgart), which shared genetic resemblance to Ötzi and modern-day Sardinians. These findings, pointing towards large-scale movements, and even replacements of people, settled the long-standing debate of *demic vs cultural diffusion* in the transmission of the Neolithic package into Europe.

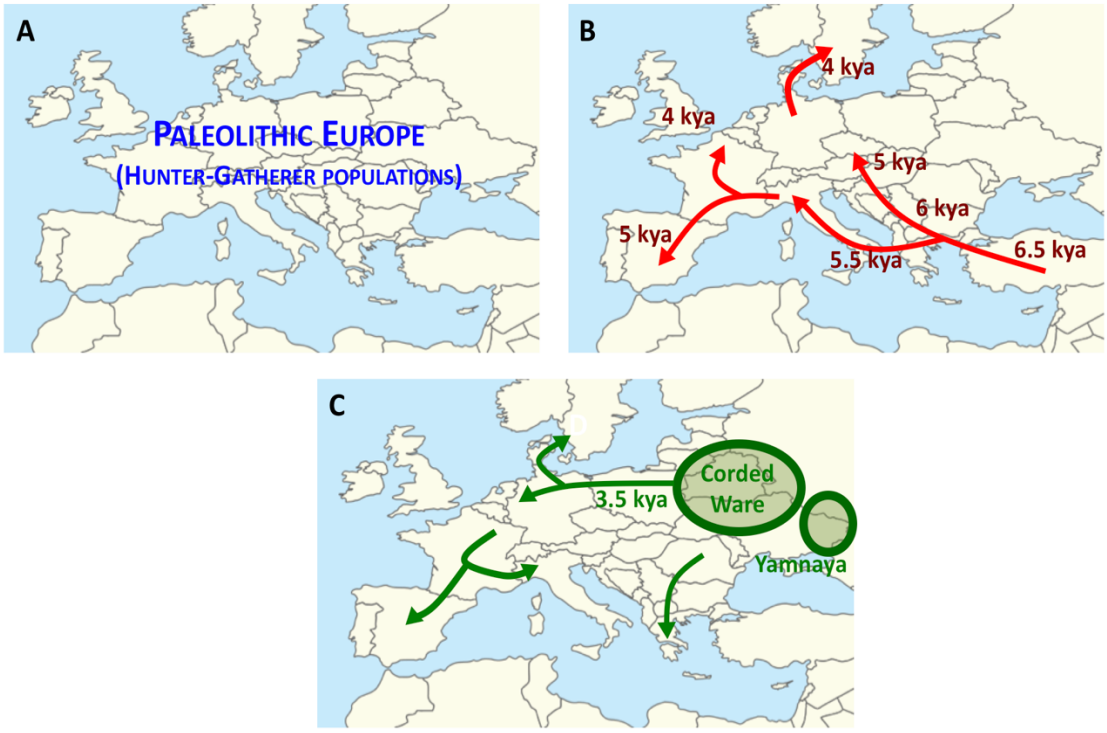
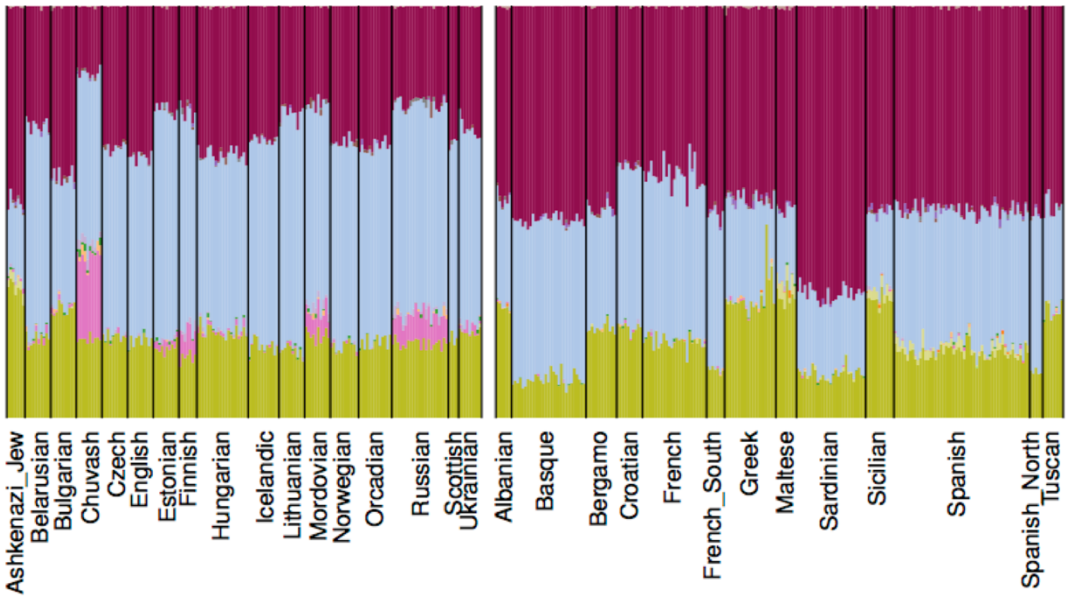


Figure 6. History of the European Genetic make-up. **A)** 10k years ago, Europe was dominated by native hunter-gatherer populations, which had just re-populated the continent after the Last Glacial Maximum. **B)** Starting 6.5k years ago, the most parts of the European continent, especially along the river valleys in Southern and Central Europe, were Neolithisised, by populations coming form the Anatolian peninsula, and a high amount of population replacement. **C)** Starting 4k years ago, the Bronze Age brought new population waves into Europe, this time bringing the wheel, bronze technology and the currently-spoken Indo-European languages.



[previous page] **Figure 7.** Admixture graph (methodical basis explained in Chapter 2) where we can observe how the genetic make-up of all European populations is a composite of the hunter-gatherer component (**blue**), the Neolithic component (**red**), and the bronze age component (**green**).

Lazaridis et al., (2014), therefore, proposed a three-ancestry model for modern Europeans: the original European hunter-gatherers (such as the Loschbour remains, in Luxembourg), the Neolithic wave of agriculturalists and farmers (the group to where Stuttgart and Ötzi belonged), and a third ancestry group which originated in the Steppes between the Black and Caspian Seas, which brought Ancient North Eurasian ancestry into the region, and were probably linked to the spread of Indo-European languages (Allentoft et al., 2015). This Steppe ancestry group was later termed the Yamnaya ancestry source, after Marja Gimbutas's Steppe Theory. In 2015, Jones et al. then linked part of this Yamnaya ancestry to the Caucasus Hunter Gatherers, a group of Hunter-gatherer populations which lived in the valleys south of the Caucasus during the Ice Age.

1.4. Objective and structure of this thesis

This thesis aims to contribute to a better understanding of the process through which the Neolithic revolution originated and expanded throughout the Old World, coupled with massive human movements and the genetic signatures that these migrations have left today.

Agriculture was brought into Europe by the descendants of early Anatolian farmers starting approximately 8,000 years ago. But little was known of the people who developed agriculture in the Fertile Crescent: were they all closely related to the early Anatolian farmers, or were there multiple ethnic groups who developed agriculture in parallel? In the second chapter, I use the first genome from a Neolithic woman from Ganj Dareh, in the Zagros Mountains in Iran, a site with evidence of early goat domestication 10,000 years ago. Were the Zagros mountains and the Iranian Plateau inhabited by populations mostly similar to from the Anatolian farmers who spread the Neolithic package into Europe , or to Hunter-gatherer populations from the Caucasus? What was the extent of the genetic and cultural exchange between Anatolia, Mesopotamia and the Zagros highlands in the early Neolithic?

The third chapter addresses how Africa was affected by population movements, originating in the Near East, during the Neolithic times. Characterising genetic diversity in Africa is a crucial step for analyses reconstructing human evolution, and little was known previous to this thesis. Using Mota, an ancient genome from a male from the Ethiopian highlands, I could answer the following questions: was there a backflow into Africa around 3,500 years ago? And if so, what was the likely source of this backflow?

The fourth chapter deals with some common problems and themes in the analysis of ancient DNA, such as merging capture datasets with diverse number of ascertained SNPs, combining capture and shotgun data in the same analysis, and the effect of UDG treatment in ancient samples. I describe the most common problems and their effect on summary statistics, and propose a guide on how to work with ancient DNA to avoid data compatibility problems.

I conclude with a final chapter in which I summarise how my thesis has shaped our understanding of some of the main migrational processes in the Old World in the last 10,000 years, and how other ancient DNA studies have advanced our understanding of human prehistory and our roots as a species, and its importance for future research. I will also discuss further paths that my research opens up, in one of the fields most interesting for understanding our origins, and our past.

2. The genetics of an early Neolithic pastoralist from the Zagros, Iran.

Abstract

The agricultural transition profoundly changed human societies. Here I present the analysis of the first genome (1.39x) of an early Neolithic woman from Ganj Dareh, in the Zagros Mountains of Iran, a site with early evidence for an economy based on goat herding, ca. 10,000 BP. This analysis shows that Western Iran was inhabited by a population genetically most similar to hunter-gatherers from the Caucasus, but distinct from the Neolithic Anatolian people who later brought food production into Europe. The inhabitants of Ganj Dareh made little direct genetic contribution to modern European populations, suggesting those of the Central Zagros were somewhat isolated from other populations of the Fertile Crescent. Runs of homozygosity are of a similar length to those from Neolithic farmers, and shorter than those of Caucasus and Western Hunter-Gatherers, suggesting that the inhabitants of Ganj Dareh did not undergo the large population bottleneck suffered by their northern neighbours. While some degree of cultural diffusion between Anatolia, Western Iran and other neighbouring regions is possible, the genetic dissimilarity between early Anatolian farmers and the inhabitants of Ganj Dareh supports a model in which Neolithic societies in these areas were distinct.

A version of this chapter has been published: Gallego-Llorente, M., Connell, S., Jones, E.R., Merrett, D.C., Jeon, Y., Eriksson, A., Siska, V., Gamba, C., Meiklejohn, C., Beyer, R., Jeon, S., Cho, Y.S., Hofreiter, M., Bhak, J., Manica, A., Pinhasi, R., 2016. The genetics of an early Neolithic pastoralist from the Zagros, Iran. Sci Rep 6.

Introduction

The agricultural transition started in a region comprising the Ancient Near East and Anatolia ~12,000 years ago with the first Pre-Pottery Neolithic villages and the first domestication of cereals and legumes (Blockley and Pinhasi, 2011; Goring-Morris and Belfer-Cohen, 2011). Archaeological evidence suggests a complex scenario of multiple domestications in a number of areas (Riehl et al., 2013), coupled with examples of trade (Aurenche and Kozlowski, 1999). Ancient DNA (aDNA) has revealed that this cultural package was later brought into Europe by dispersing farmers from Anatolia (so called 'demic' diffusion, as opposed to non-demic cultural diffusion (Mathieson et al., 2015; Omrak et al., 2016)) ~8,400 years ago. However a lack of aDNA from early Neolithic individuals from the Near East leaves a key question unanswered: was the agricultural transition developed by one major population group spanning the Near East, including Anatolia and the Central Zagros Mountains; or was the region inhabited by genetically diverse populations, as is suggested by the heterogeneous mode and timing of the appearance of early domesticates at different localities? It is unknown whether the emergence of agriculture was a geographically homogeneous event throughout the Fertile Crescent region, or whether the picture was more a mosaic of different, more localised processes of domestication. In a similar fashion, we can extend this question to the genetics: were the farming populations homogeneous, or did they have a structure that was preserved as the Neolithic package was introduced in other areas of the world?

Additionally, western Eurasians have a relative genetic homogeneity in a world context (Cavalli-Sforza et al., 1994). This could be explained by the fact that a group of people extensively migrated and admixed with local populations in a way that homogenised populations with deeper splits of ancestry.

Fig. 1. Map of the Near East at the start of the Neolithic. The Neolithic revolution was exported into Europe by populations closely related to the Anatolian Neolithic farmer communities (found in Barcin Höyük). However, it remained unknown how homogeneous the Near East was, and how the Neolithic revolution affected Central and South Asia in terms of demography.

Recent studies in ancient DNA have shown that the earliest Anatolian and Aegean farmers (from around 6,500 years ago) have a similar genetic component to that shown by the early European farmers (Hofmanová et al., 2016; Mathieson et al., 2015). However, it is yet unknown whether this component arrived from the farming populations from the Fertile Crescent proper.

To answer these questions, the genome of an early Neolithic female from Ganj Dareh, GD13a was sequenced. This individual, from the Central Zagros (Western Iran), dated to 10000-9700 cal BP (Zeder and Hesse, 2000), a region located at the eastern edge of the Near East. Ganj Dareh is well known for providing the earliest evidence of herd management of goats beginning at 9,900 BP (Zeder, 2011, 2008; Zeder and Hesse, 2000). It is a classic mound site at an altitude of ~1400m in the Gamas-Ab Valley of the High Zagros zone in Kermanshah Province, Western Iran. It was discovered in the 1960s during survey work and excavated over four seasons between 1967 and 1974. The mound, ~40

m in diameter, shows 7 to 8 m of early Neolithic cultural deposits. Five major levels were found, labelled A through E from top to bottom. Extended evidence showed a warren of rooms with evidence of under-floor inhumations within what may be burial chambers and/or disused houses (Smith, 1990). The current Minimum Number of Individuals is 116, with 56 catalogued as skeletons that had four or more bones recovered (Merrett, 2004). The individual analysed here was part of burial 13, which contained three individuals, and was recovered in level C in 1971 from the floor of a brick-walled structure. The individual sampled, 13A (referred to as GD13a throughout the text), was a 30-50 year old female; the other individuals in the burial unit were a second adult (13B) and an adolescent (13).

The site has been directly dated to 9650-9950 cal BP (Zeder and Hesse, 2000), and shows intense occupation over two to three centuries. The economy of the population was that of pastoralists with an emphasis on goat herding (Zeder and Hesse, 2000). Archaeobotanical evidence is limited (van Zeist et al., 1984) but the evidence present is for two-row barley with no evidence for wheat, rye or other domesticates. This implies that the overall economy was at a much earlier stage in the development of cereal agriculture than that found in the Levant, Anatolia and Northern Mesopotamian basin.

Results

Sequence processing, alignment, and authenticity of results

Although most fossil specimens of interest in palaeogenetics only contain less than 1% of endogenous DNA, it was shown in 2015 that the petrous section of the temporal bone reliably contains a higher proportion of endogenous DNA than any other skeletal element (Pinhasi et al., 2015). In this case, the petrous bone of GD13a was isolated, cleaned and sequenced, following the approach carried out in Gamba et al. (2014). This sample yielded sequencing libraries comprising 18.57% alignable human reads that were used to generate 1.39-fold genome coverage (Table 1). The sequence data showed read lengths and nucleotide misincorporation patterns, which are indicative of post-mortem damage, and hence are compatible with that of ancient DNA, supporting the authenticity of results and discarding modern contamination (Fig. 2). Furthermore, the estimation of the mitochrondrial contamination rate is <1%. This was based on evaluating the proportion of non-consensus bases at haplogroup-defining positions, in bases with qualiy ≥20 (Fig. 3). The mitochondrion of GD13a (91.74X) was assigned to haplogroup X, most likely to the subhaplogroup X2, which has been associated with an early expansion from the Near East (Reidla et al., 2003; Richards et al., 2000) and has been found in early Neolithic samples from Anatolia (Mathieson et al., 2015), Hungary (Gamba et al., 2014) and Germany (Haak et al., 2015).

Sample	Total reads	Aligned reads	(%)	High quality reads	(%)	Coverage (x)
GD13a	728,931,167	135,327,301	18.57	90,189,417	12.37	1.39

Table 1. Alignment statistics for GD13a, showing the total reads, the percentage of reads that were aligned, and the percentage of which were high quality. The total coverage, 1.39x, shows that every position was covered an average of 1.39 times.

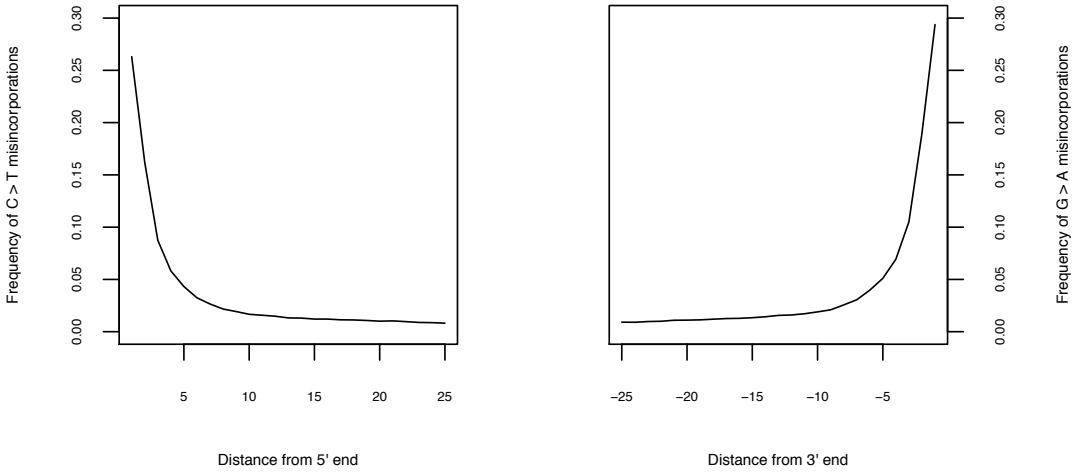


Fig. 2. Damage patterns for GD13a. Plots show mismatch frequency relative to the reference genome as a function of read position. The left hand figure shows the frequency of C to T misincorporations at the 5' ends of reads (first 25 bases) while the right hand figure shows the frequency of G to A transitions at the 3' ends of reads (last 25 bases).

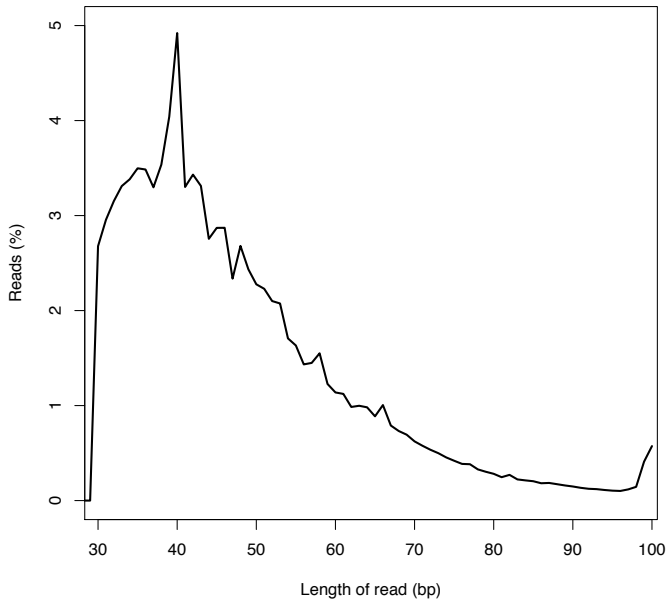


Fig. 3. Sequence length distribution for GD13a. Plot shows the proportion of reads of any given length. After the cut-off length of 30, the highest proportion of reads was between 30 and 50 base pairs, with longer reads being more infrequent.

GD13a shows affinity with Caucasus Hunter Gatherers and Central South Asian populations

I compared GD13a with a number of other ancient genomes and modern populations (Allentoft et al., 2015; Cassidy et al., 2016; Fu et al., 2015; Gamba et al., 2014; Günther et al., 2015; Haak et al., 2015; Jones et al., 2015; Keller et al., 2012; Lazaridis et al., 2014; Olalde et al., 2015, 2014; Omrak et al., 2016; Raghavan et al., 2014; Seguin-Orlando et al., 2014), using principal component analysis (PCA) (Patterson et al., 2006), ADMIXTURE (Alexander et al., 2009) and outgroup f_3 statistics (Patterson et al., 2012) (Figs 3, 4 and 5). GD13a did not cluster with any other early Neolithic individual from Eurasia in any of the analyses. PCA also revealed some affinity with modern Central South Asian populations such as Balochi, Makrani and Brahui (Fig. 4).

[figure 4 in next page]

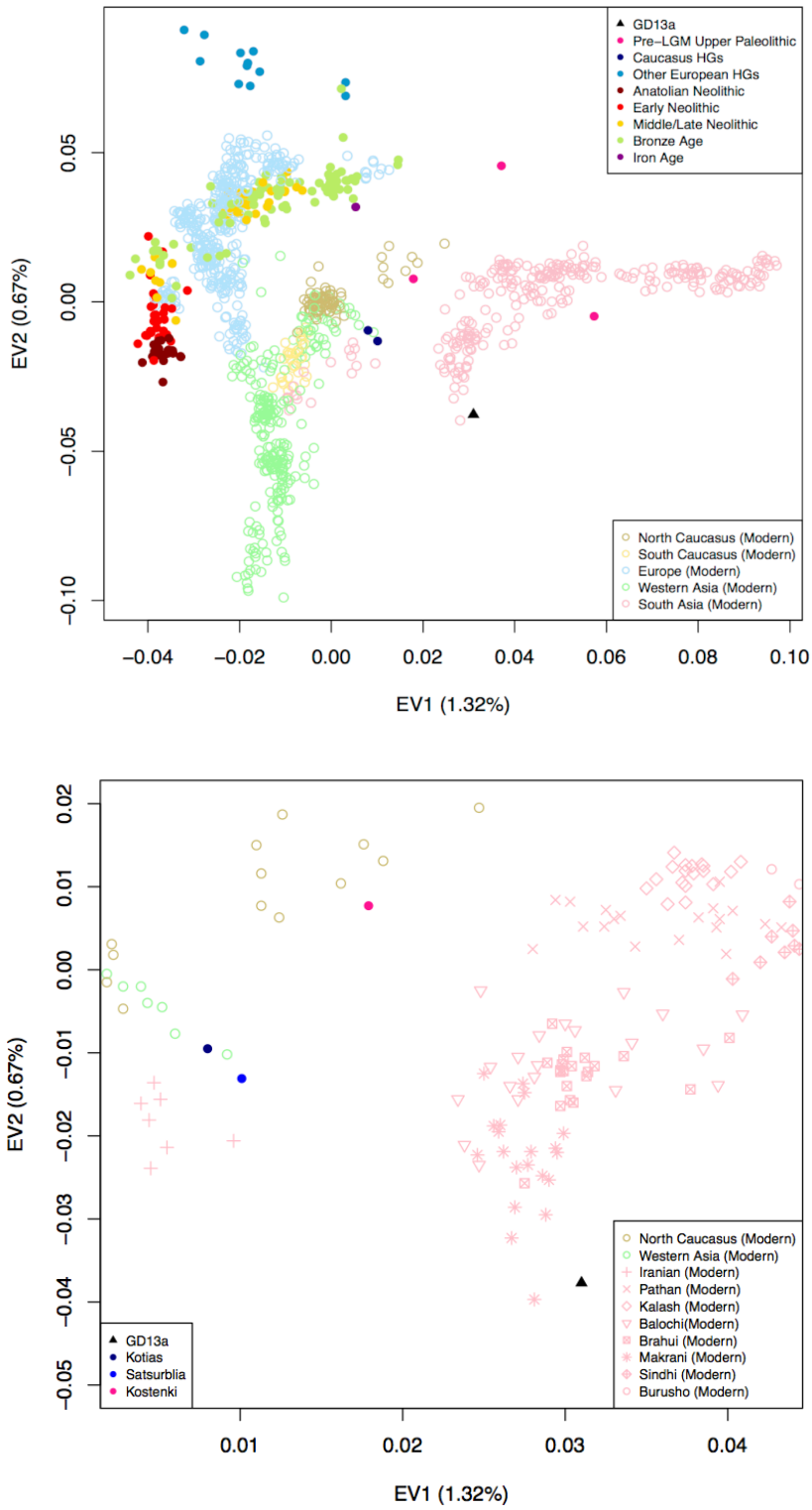


Fig. 4. A) PCA loaded on modern populations (represented by open symbols). Ancient individuals (solid symbols) are projected onto these axes. **B)** Zoom into the populations visually close to GD13a, revealing affinities to modern Balochi, Brahui and Makrani populations.

Outgroup f_3 statistics, a statistical tool analysis that estimates the shared genetic drift of two populations from an outgroup, identified Caucasus Hunter-Gatherers of Western Georgia, just north of the Zagros mountains, as the group genetically most similar to GD13a (Figs. 5 and 6, Tables 2 and 3). ADMIXTURE, a clustering maximum likelihood algorithm, showed that also genetically close to GD13a (with mostly green-coloured component) were ancient samples from Steppe populations (Yamanya & Afanasievo), which share the green component of GD13a, together with blue. These steppe populations were part of one or more of the Bronze age migrations into Europe, as well as the early Bronze age cultures in that continent (Corded Ware) (Allentoft et al., 2015; Haak et al., 2015) (Fig. 7). The UPGMA (unweighted pair group method with arithmetic mean) tree also showed GD13a to be genetically close to Caucasus Hunter Gatherers and modern Caucasus populations. These results, therefore, are in line with previous relationships observed for the Caucasus Hunter-Gatherers (Jones et al., 2015).

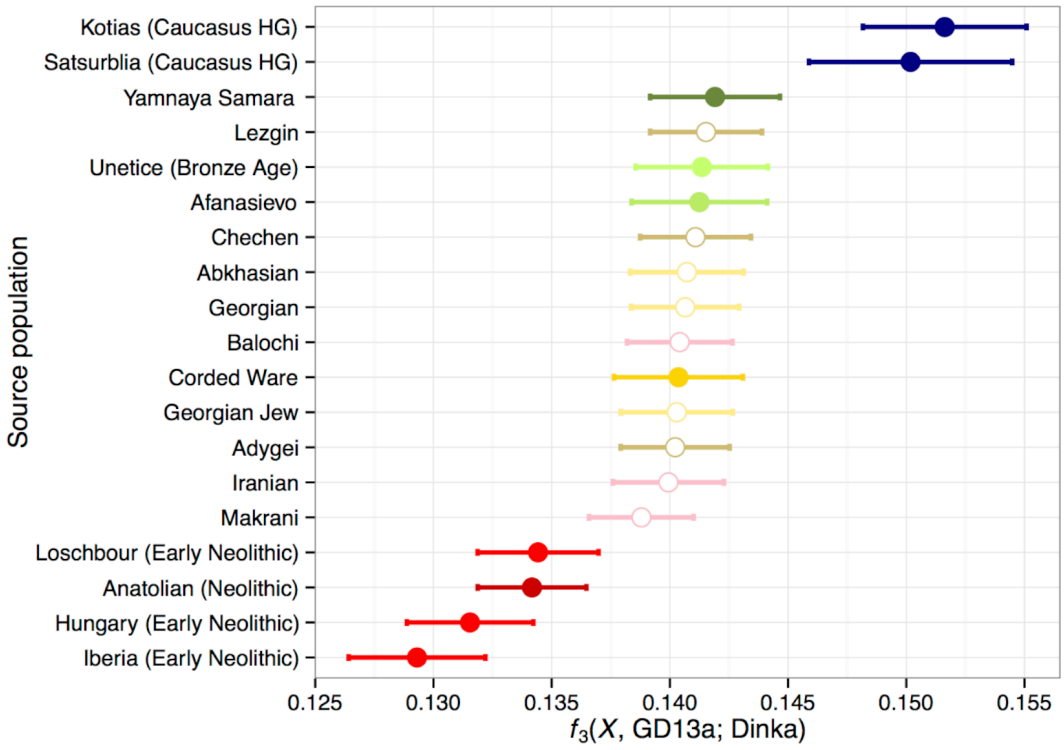


Fig. 5. Outgroup f_3(*X*, GD13a; Dinka), where Caucasus Hunter Gatherers (Kotias and Satsurblia) share the most drift with GD13a. Ancient samples have filled circles whereas modern populations are represented by empty symbols.

X	f_3	Standard Error
Kotias	**0.152**	**0.003**
Satsurblia	**0.150**	**0.004**
Russia (Early Bronze Age)	**0.142**	**0.007**
Yamnaya Samara	**0.142**	**0.003**
Lezgin	0.142	0.002
Unetice (Early Bronze Age)	**0.141**	**0.003**
Afanasievo	**0.141**	**0.003**
Chechen	0.141	0.002
Abkhasian	0.141	0.002
Georgian	0.141	0.002
Balochi	0.140	0.002
Corded Ware Germany	**0.140**	**0.003**
Georgian Jew	0.140	0.002
Adygei	0.140	0.002
Bell Beaker Germany	**0.140**	**0.006**
Yamnaya Kalmykia	**0.140**	**0.003**
Iranian	0.140	0.002
Brahui	0.140	0.002
Iranian Jew	0.140	0.002
Kalash	0.140	0.002
Armenian	0.139	0.002
Iraqi_Jew	0.139	0.002
Srubnaya	**0.139**	**0.003**
Irish (Bronze Age)	**0.139**	**0.003**
Tajik_Pomiri	0.139	0.002
Nordic (Middle Neolithic)	**0.139**	**0.006**
Makrani	0.139	0.002
Pathan	0.139	0.002
Kumyk	0.139	0.002
Balkar	0.138	0.002
Sindhi	0.138	0.002

Table 2. f_3(*X*, GD13a; Dinka) where X represents a modern or ancient individual/population. Ancient individuals/populations are shown in bold. Populations/individuals with the largest f_3 values are shown.

X	f_3	Standard Error
Kotias	**0.247**	**0.004**
Satsurblia	**0.243**	**0.005**
Spain (Middle Bronze Age)	**0.237**	**0.006**
Corded Ware Germany	**0.236**	**0.008**
Chechen	0.236	0.003
Abkhasian	0.236	0.003
Lezgin	0.236	0.003
Georgian Jew	0.236	0.003
Balochi	0.236	0.003
Yamnaya Samara	**0.236**	**0.003**
Yamnaya Kalmykia	**0.236**	**0.004**
Unetice (Early Bronze Age)	**0.235**	**0.003**
Brahui	0.235	0.003
Adygei	0.235	0.003
Georgian	0.235	0.003
Iranian Jew	0.235	0.003
Iranian	0.235	0.003
Bell Beaker Czech	**0.235**	**0.005**
Iraqi Jew	0.235	0.003
Kalash	0.235	0.003
Tajik Pomiri	0.234	0.003
Makrani	0.234	0.003
Pathan	0.234	0.003
Afanasievo	**0.234**	**0.003**
Armenian	0.234	0.003

Table 3. f_3(*X*, GD13a; Ju'Hoansi) where X represents a modern or ancient individual/population. Ancient individuals/populations are shown in bold. Populations/individuals with the largest f_3 values are shown.

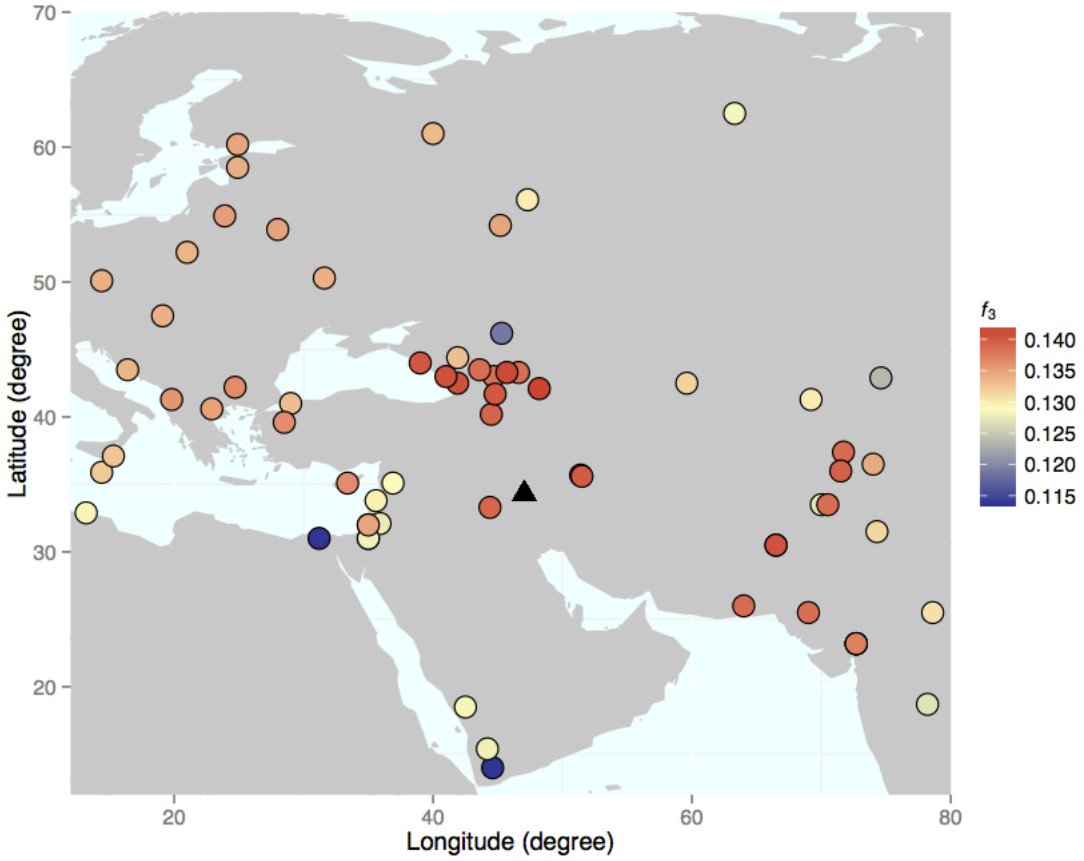


Fig. 6. GD13a shares genetic drift with modern Caucasus and South Asian populations. The statistic f_3(*X*, GD13a; Dinka) shows that the closest modern populations to GD13a are Caucasus populations and, to some extent, South Asian populations such as Balochi and Makrani. Map of populations was generated with the library "ggplot2" with R software (v3.1.2, https://cran.r-project.org/). Populations in blue with the lowest f_3 values represent Yemenite, Egyptian or Russian Jew populations, with more complex population histories. (R Development Core Team, 2001).

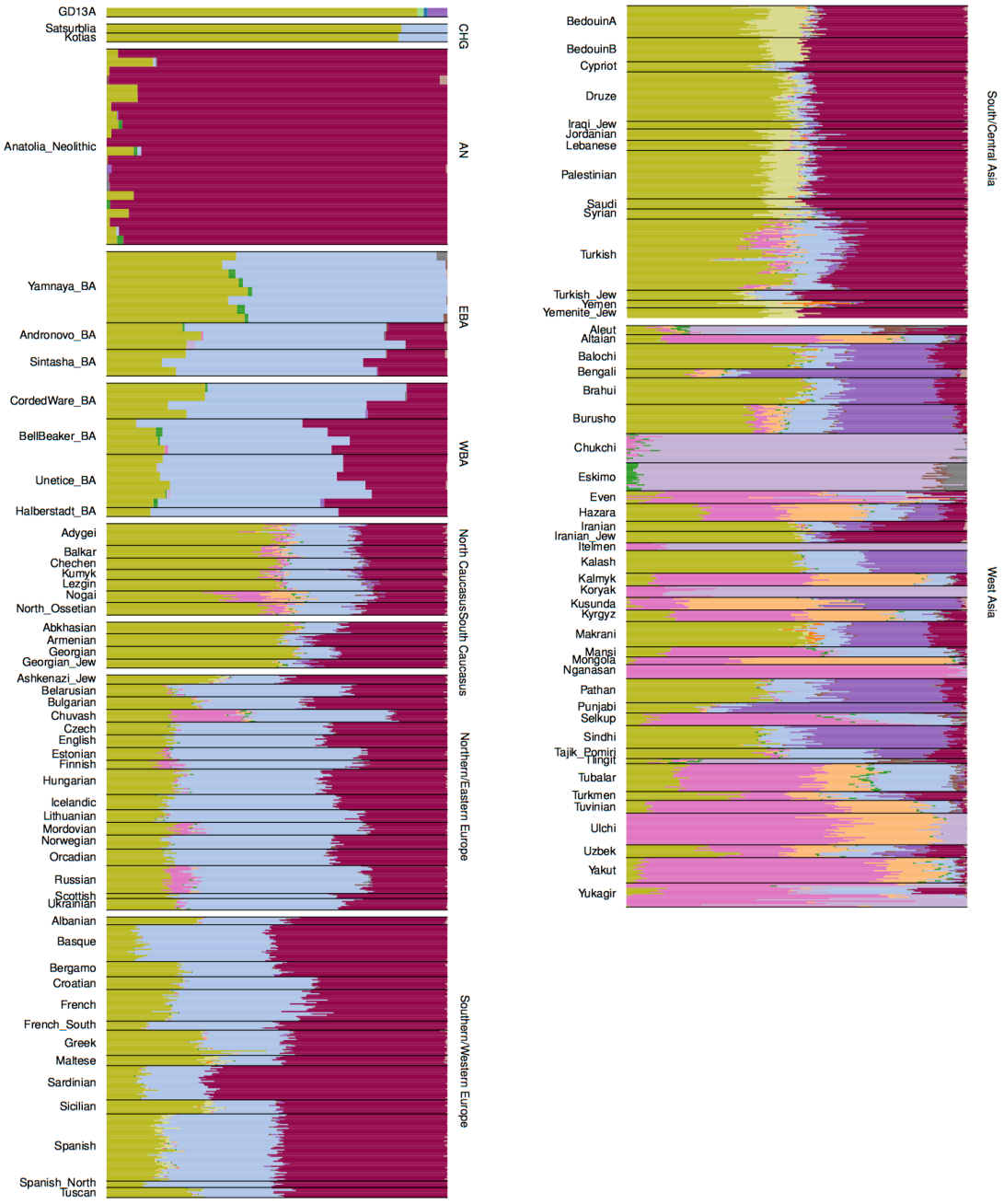


Fig. 7. **ADMIXTURE** using K=17, where GD13a appears very similar to Caucasus Hunter Gatherers, and to a lesser extent to modern south Asian populations.

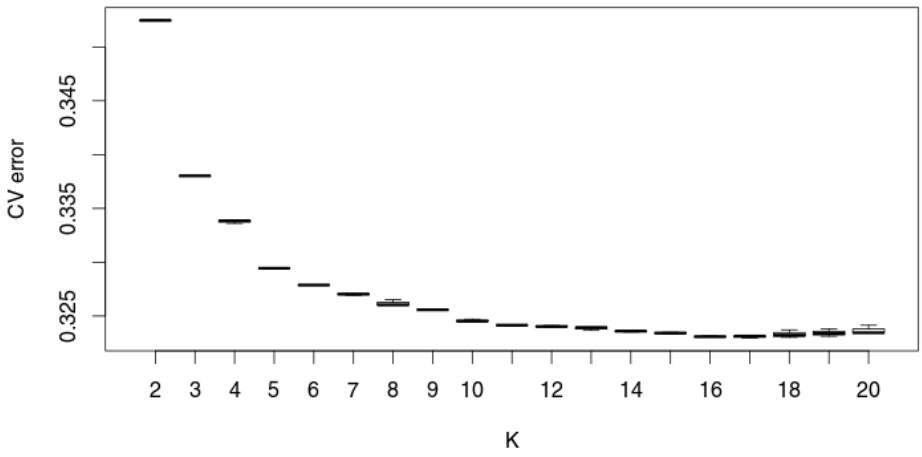


Fig. 8. ADMIXTURE analysis cross validation (CV) error as a function of the number of clusters (K). **Both the lowest minimal and mean value was attained at K=17.**

The UPGMA tree shows that GD13a clusters in the same branch as the Caucasus Hunter-Gatherers (Kotias and Satsurblia) (Fig. 9).

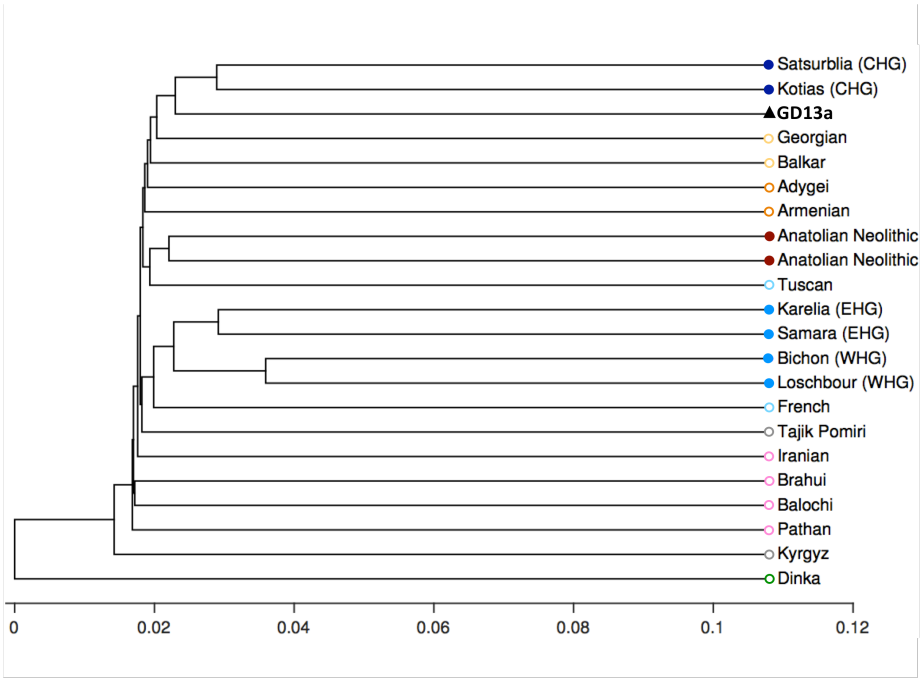


Fig. 9. UPGMA Tree (Unweighted Pair Group Method with Arithmetic Mean), showing that GD13a clusters together with Caucasus Hunter Gatherers (CHG). EHG, Eastern Hunter Gatherers; WHG, Western Hunter Gatherers.

I further investigated the relationship between GD13a and Caucasus Hunter-Gatherers using *D*-statistics (Green et al., 2010; Patterson et al., 2012) to test whether they formed a clade to the exclusion of other ancient and modern samples (Table 4). Every tested population, including a large number of Western Eurasian samples (both modern and ancient) showed significant excess genetic affinity to the Caucasus Hunter-Gatherers, whilst none did with GD13a. Overall, these results point to GD13a having little direct genetic input into later modern populations, European or Asian, compared to its northern neighbours, the Caucasus Hunter-Gatherers.

GD13a did not undergo a recent large population bottleneck

To better understand the history of the population to which GD13a belonged, I investigated the distribution of lengths of runs of homozygosity (ROH) (Fig. 10). A bias towards a high frequency of both long and short ROH is indicative of past strong bottlenecks followed by population re-expansion. GD13a has a distribution with few long ROH (>2 Mb), similar to that of the descendants of Anatolian early farmers (represented by the European farmers NE1 (Gamba et al., 2014) and Stuttgart (Lazaridis et al., 2014)). In contrast, both Western (Lazaridis et al., 2014) and Caucasus Hunter-Gatherers (Jones et al., 2015) have relatively more long as well as short ROH. Thus, GD13a is the descendant of a group that had relatively stable demography, i.e. without large shifts in population numbers, and did not suffer the bottlenecks that affected more northern or mountainous populations. Small effective population sizes and recent inbreeding would give a different pattern to the one we see in GD13a or Kotias, but more similar to the one presented by Satsurblia. Satsurblia presents a higher proportion of longer ROH, due to less recombination events due to fewer elapsed generations. This supports the fact that the similar ROH distribution shared by GD13a and Kotias is indicative of deep population bottlenecks, which since then have created a large number of very fragmented ROH.

[Next page] **Table 4**. ***D*-statistics of the form *D*(Dinka, *X*; GD13a, Kotias) where *X* represents a modern or ancient individual/population.**

Ancient individuals/populations are shown in bold. MN: Middle Neolithic, LN: Late Neolithic, LBA: Late Bronze Age, MBA: Middle Bronze Age, EBA: Early Bronze Age, HG: Hunter Gatherer, BA: Bronze Age. Populations/individuals with the largest values of *D* are shown.

X	*D*-statistic	*Z*-score
Satsurblia	**0.090**	**6.70**
Esperstedt_MN	**0.058**	**4.56**
Alberstedt_LN	**0.057**	**4.81**
Corded_Ware_Estonia	**0.055**	**3.67**
Srubnaya_Outlier	**0.053**	**4.10**
Halberstadt_LBA	**0.053**	**4.56**
Corded_Ware_Germany	**0.050**	**5.92**
BenzigerodeHeimburg_LN	**0.050**	**3.83**
Bell_Beaker_Germany	**0.049**	**2.55**
Afanasievo.	**0.049**	**5.28**
Iberia_Mesolithic	**0.048**	**4.02**
Samara_HG	**0.047**	**3.04**
Sintashta_MBA	**0.047**	**4.56**
Georgian	0.046	6.99
Unetice_EBA	**0.046**	**4.51**
Orcadian	0.046	6.82
Iberia_Chalcolithic	**0.045**	**5.01**
Poltavka	**0.045**	**4.71**
Karelia_HG	**0.044**	**3.70**
Bell_Beaker_Germany	**0.043**	**4.99**
Loschbour	**0.043**	**4.09**
Iberia_EN	**0.043**	**4.48**
MA1	**0.043**	**3.20**
Estonian	0.042	6.11
Abkhasian	0.042	6.23
Yamnaya_Kalmykia	**0.042**	**4.69**
Ukrainian	0.042	6.11
Croatian	0.041	6.10
Yamnaya_Samara	**0.041**	**5.07**
Czech	0.041	5.91
Norwegian	0.040	5.85
English	0.040	5.75
Remedello_BA	**0.040**	**3.12**
French_South	0.040	5.51
Lithuanian	0.040	5.79
Baalberge_MN	**0.039**	**2.70**
Kumyk	0.039	5.75
Hungarian	0.039	5.83
Srubnaya	**0.039**	**4.92**
Icelandic	0.039	5.60
North_Ossetian	0.039	5.73
French	0.038	5.80
Adygei	0.038	5.74
BattleAxe_Sweden	**0.038**	**2.49**
Balkar	0.038	5.60
Iberia_MN	**0.038**	**3.98**
Spanish_North	0.037	4.91
Motala_HG	**0.037**	**4.13**
Belarusian	0.037	5.38
Spanish	0.037	5.72

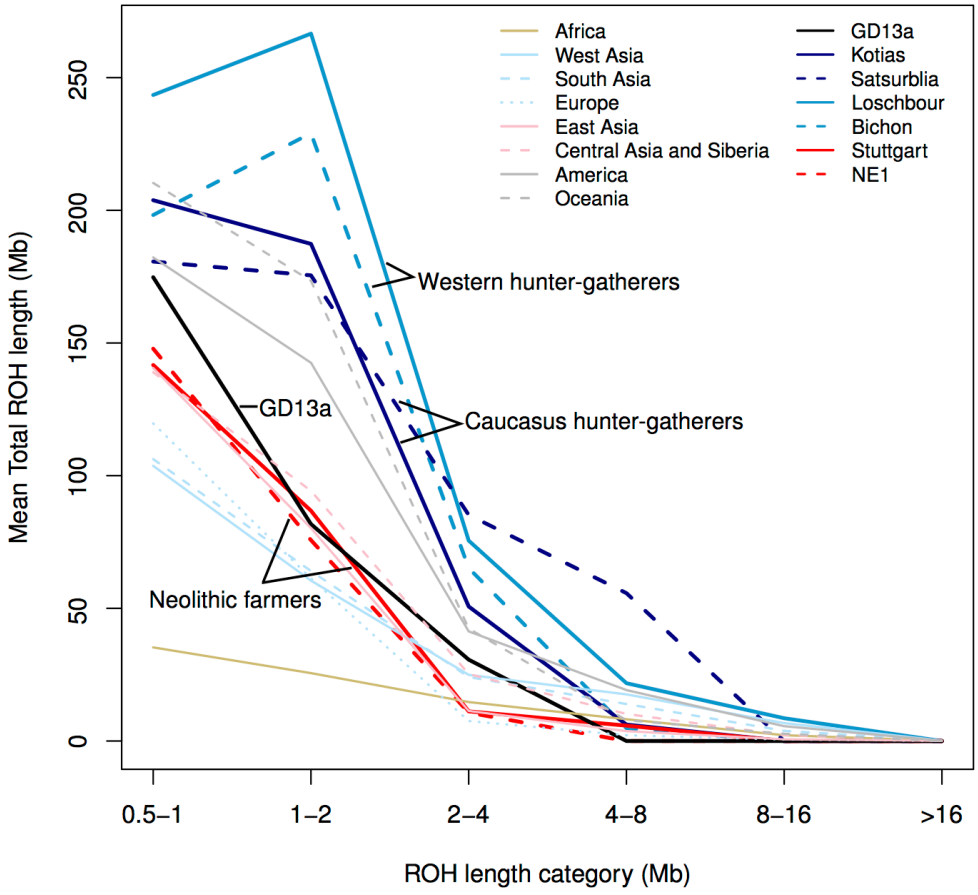


Fig. 10. GD13a has similar runs of homozygosity (ROH) lengths to Neolithic individuals, while Caucasus Hunter Gatherers (Kotias and Satsurblia), like European Hunter Gatherers (Loschbour and Bichon), underwent recent large population bottlenecks potentially associated with the LGM.

GD13a and Kotias are possible surrogates for Ancestral Northern Indians

After the fact that analysing GD13a showed a large degree of similarity between GD13a and Kotias, the next step was to check whether these two populations contributed are equally related to the steppe component of Ancient Northern Indians, or whether one of both is genetically closer to said component. It remained a possibility that farmers from the Near East contributed to the eastern diffusion of agriculture from the Near East that reached Turkmenistan (Harris et al., 2010) by the 6th millennium BP, and continued further east to the Indus Valley (Gangal et al., 2014). However, detecting such a contribution is complicated by a later influx from Steppe populations with Caucasus Hunter-Gatherer ancestry during the Bronze Age. I tested whether the Western Eurasian component found in Indian populations can be better attributed to either of these two

sources, GD13a and Kotias (a Caucasus Hunter Gatherer), using *D*-statistics to detect gene flow into an ancestral Indian component (represented by the Onge in Fig. 11 and Kharia in Fig. 12). Overall, for all tests where a difference could be detected, Kotias was a slightly closer source than GD13a. In other tests, Kotias and GD13a were equally likely sources (Fig. 11 and Table 5). These results remained in line with the ones observed in Table 4, where every tested population showed significant excess genetic affinity to the Kotias. Using the Kharia, a non-admixed population from Central India, as the representative of the ancestral Indian component yielded similar results (Fig. 12 and Table 6). Whilst the attribution of part of the Western Eurasian component seen in India to Bronze Age migrations is supported by dating of last contact based on patterns of Linkage Disequilibrium (Kirin et al., 2010), this analysis highlights the possibility that part of that component might derive from earlier contact during the eastern diffusion of agriculture, as GD13a and Kotias have similar genetic affinity to many subcontinental populations.

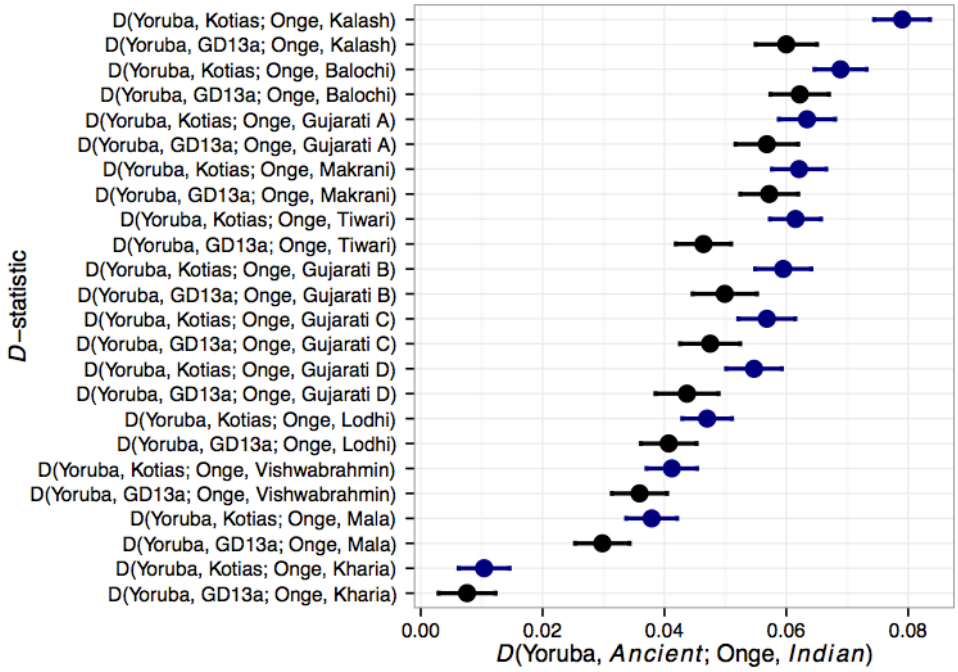


Fig. 11. *D*-statistics of the type *D*(Yoruba, *Ancient*; Onge, *South Asian*), where *Ancient* is represented by Kotias or GD13a, whereas *South Asian* is represented by Modern South Asian populations. Overall, both genomes were equally good proxies for ANI, even though Kotias was marginally better for a few comparisons (Kalash and Tiwari).

	***D*(Yoruba, GD13a; Onge, *S. Asian*)**		***D*(Yoruba, Kotias; Onge, *S. Asian*)**	
Population	***D*-statistic**	***Z***	***D*-statistic**	***Z***
Gujarati A	0.057	11.086	0.063	13.737
Gujarati B	0.050	9.462	0.060	12.943
Gujarati C	0.048	9.641	0.057	12.176
Gujarati D	0.044	8.455	0.055	11.992
Lodhi	0.041	8.914	0.047	11.531
Mala	0.030	6.673	0.038	9.092
Vishwabrahmin	0.036	7.946	0.041	9.952
Tiwari	0.046	10.236	0.062	14.701
Kharia	0.008	1.626	0.010	2.493
Kalash	0.060	11.935	0.079	17.256
Balochi	0.062	12.981	0.069	16.124
Makrani	0.057	12.060	0.062	13.959

Table 5. *D*-statistics of the form *D*(Yoruba, *Ancient*; Onge, *S. Asian*), where *Ancient* is either GD13a or Kotias, while *S. Asian* are different modern Indian and South Asian populations.

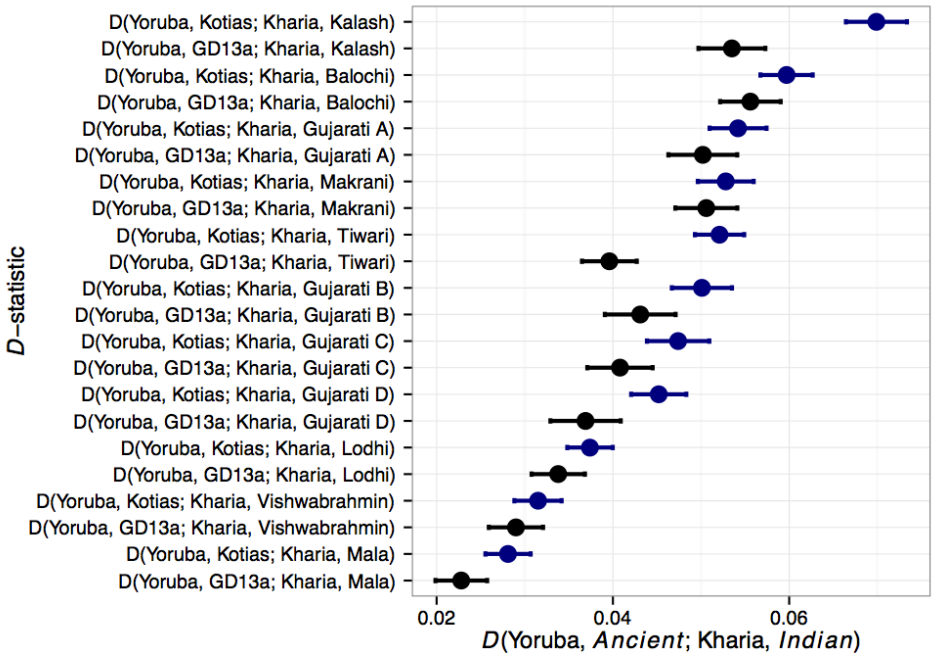


Fig. 12. *D*-statistics of the type *D*(Yoruba, *Ancient*; Kharia, *South Asian*), where *Ancient* is represented by Kotias or GD13a, whereas *South Asian* is represented by Modern South Asian populations. Overall, both genomes were equally good proxies for ANI, even though Kotias was marginally better for a few comparisons (Kalash and Tiwari).

	***D*(Yoruba, GD13a; Kharia, *S. Asian*)**		***D*(Yoruba, Kotias; Kharia, *S. Asian*)**	
Population	***D*-statistic**	***Z***	***D*-statistic**	***Z***
Gujarati A	0.050	12.856	0.054	16.79
Gujarati B	0.043	10.743	0.050	14.688
Gujarati C	0.041	11.016	0.047	13.398
Gujarati D	0.037	9.252	0.045	14.45
Lodhi	0.034	11.175	0.037	14.507
Mala	0.029	7.781	0.028	10.964
Vishwabrahmin	0.029	9.427	0.032	11.749
Tiwari	0.040	12.752	0.052	18.66
Kalash	0.054	14.101	0.070	20.194
Balochi	0.056	16.199	0.058	20.142
Makrani	0.051	14.399	0.053	16.662

Table 6. *D*-statistics of the form *D*(Yoruba, *Ancient*; Kalash, *S. Asian*), where *Ancient* is either GD13a or Kotias, while *S. Asian* are different modern Indian and South Asian populations.

Phenotypes of Interest

The phenotypic attributes of GD13a are similar to the neighbouring Anatolian early farmers and Caucasus Hunter-Gatherers. Based on diagnostic SNPs, she had dark, black hair and brown eyes (see Table 7). She lacked the derived variant (rs16891982) of the *SLC45A2* gene associated with light skin pigmentation but likely had at least one copy of the derived *SLC24A5* allele (rs1426654) associated with the same trait. The derived *SLC24A5* variant has been found in both Neolithic farmer and Caucasus Hunter-Gatherer groups (Gamba et al., 2014; Jones et al., 2015; Mathieson et al., 2015) suggesting that it was already at appreciable frequency before these populations diverged. Finally, she did not have the most common European variant of the *LCT* gene (rs4988235) associated with the ability to digest raw milk, consistent with the later emergence of this adaptation (Allentoft et al., 2015; Gamba et al., 2014; Mathieson et al., 2015).

Using the Hirisplex prediction model (Walsh et al., 2013), GD13a was predicted to have brown eyes (p-value = 0.993) and dark (p-value=0.997), black (p-value=0.899) hair. This was confirmed using imputed genotypes. The eye-colour *HERC2* variant rs12913832 was assigned almost equal likelihoods of being homozygous for the ancestral allele (A; genotype probability = 0.501) and heterozygous (AG; genotype probability = 0.499). Given this result, and that the ancestral allele was observed (2-fold coverage) in the sample it is very likely that GD13a had at least one copy of the ancestral dominant allele associated with brown eyes. Using either state (homozygous ancestral or heterozygous) in the Hirisplex model and imposing a genotype probability cut-off of 0.9 for the other imputed genotypes, GD13a was predicted with the imputation approach to have dark (p-value ≥ 0.974), black (p-value ≥ 0.703) hair and brown eyes (p-value ≥ 0.952).

We did not observe the derived *SLC45A2* variant (rs16891982) associated with light skin pigmentation in GD13a (also supported by the imputed genotype) but did observe the derived *SLC24A5* variant (rs1426654) which is also associated with the same trait in modern populations. The imputed genotype for the latter suggests that this individual was heterozygous at this position (genotype probability > 0.999). Using either observed or imputed genotypes, GD13a did not show the most common variant of the *LCT* gene (rs4988235) associated with lactase persistence in Europeans (Table 7). I visually inspected the rs12913832 (HERC), rs1110400 (MC1R), and rs1426654 (SLC24A5) sites using the Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al.,

2013) as the observed alleles found at these sites may be the result of deaminated cytosine residues. I found that the alleles called at the rs12913832 and rs1110400 variant sites are greater than 10 bp from either end of the read. For the rs1426654 site at least one allele was 10 bp from either end of the read. This, combined with the imputed genotypes, suggests that the alleles called at these positions are unlikely to be the result of postmortem DNA damage which is more prevalent at read termini.

Gene	Marker	Observed genotype	Coverage	Imputed genotype	Imputed Genotype probability
EXOC2	rs4959270	-	-	CA	> 0.999
HERC2	rs12913832	AA	2A	AA/GA	0.501/0.499
IRF4	rs12203592	-	-	CC	0.999
KITLG	rs12821256	-	-	TC	0.752
MC1R	N29insA	-	-	-	-
MC1R	rs1110400	TT	1T	TT	0.998
MC1R	rs11547464	-	-	GG	> 0.999
MC1R	rs1805005	-	-	GG	> 0.999
MC1R	rs1805006	-	-	CC	0.996
MC1R	rs1805007	CC	1C	CC	> 0.999
MC1R	rs1805008	CC	1C	CC	0.999
MC1R	rs2228479	-	-	GG	0.999
MC1R	rs885479	-	-	GG	0.913
MC1R	Y1520CH	-	-	-	-
MCIR	rs1805009	GG	1G	GG	0.999
OCA2	rs1800407	-	-	CC	0.985
PIGU/ASIP	rs2378249	-	-	AA	> 0.999
SLC24A4	rs12896399	GG	3G	GG	> 0.999
SLC24A4	rs2402130	GG	1G	GA	> 0.999
SLC45A2	rs16891982	CC	2C	CC	> 0.999
SLC45A2	rs28777	CC	1C	CC	0.999
TYR	rs1042602	CC	1C	CC	> 0.999
TYR	rs1393350	GG	3G	GG	> 0.999
TYRP1	rs683	CC	2C	CC	> 0.999
SLC24A5	rs1426654	AA	2A	AG	> 0.999
LCT	rs4988235	GG	2G	GG	> 0.999

Table 7. Observed and imputed genotypes for GD13a at variant sites associated with phenotypes of interest.

Discussion

GD13a had little direct genetic input into later European populations compared to the Caucasus Hunter-Gatherers (its northern neighbours) as demonstrated using *D*-statistics. This lack of connectivity with neighbouring regions might have arisen early on, since this report also finds that hunter-gatherers from the Caucasus show higher affinity to Western Hunter-Gatherers and early Anatolian farmers; this result suggests the possibility of gene flow between the former and these two latter groups to the exclusion of GD13a. An alternative, but not mutually exclusive, explanation for this pattern is that GD13a might have received genetic input from a source equally distant from all other European populations, and thus basal to them. This possible deep lineage of non-African ancestry, which branched off before all the other non-Africans branched off from one another, was suggested by Lazaridis et al. in 2014, and termed Basal Eurasians.

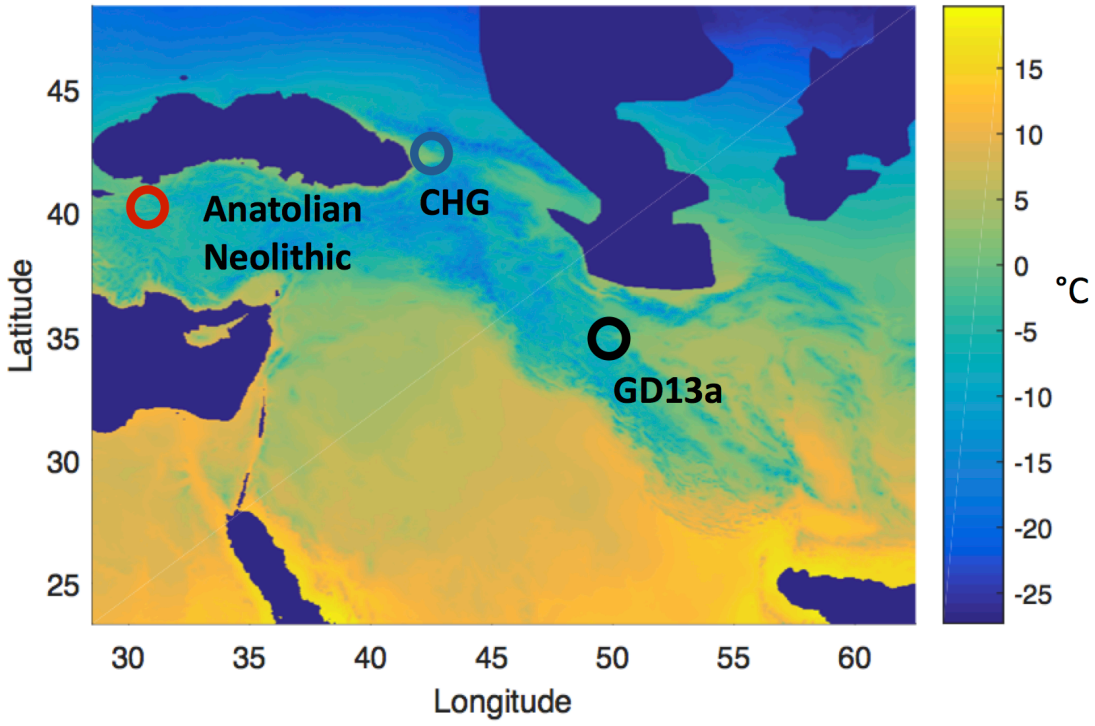


Fig. 12. Map showing geographical location of Anatolian Neolithic samples, Caucasus Hunter Gatherers (CHG) and GD13a. Background colours indicate mean temperature (°C) of coldest quarter during the LGM, with LGM sea levels. Map showing geographical location of Anatolian Neolithic samples, Caucasus Hunter Gatherers (CHG) and GD13a. Background colours indicate mean temperature (°C) of coldest quarter during the LGM (data from the worldclim database60 generated by the CCSM4 model) (Hijmans et al., 2005), with LGM sea levels. Map of temperatures was generated with MATLAB R2015b (Mathworks, http://www.mathworks.com/).

The Last Glacial Maximum (LGM) made entire regions in northern Eurasia uninhabitable, and therefore a number of hunter-gatherer populations likely moved to the south. For Europe there may be a separation between Western and Eastern populations with minimal occupation of the Central European plains (Seguin-Orlando et al., 2014). For Eastern Europe, Central Asia and the northern Near East, glaciation itself was less a factor. In these areas, our understanding of how populations weathered the LGM is still vague at best. It has previously been suggested that differences in the frequency of long and short runs of homozygosity in ancient samples may be associated with different demographic experiences during the LGM (Gamba et al., 2014; Jones et al., 2015). Neolithic farmers, with their lower frequency of short ROH, have been argued to have been relatively little affected by the LGM compared to Western and Caucasus Hunter-Gatherers (Gamba et al., 2014; Jones et al., 2015) which are characterised by more long ROH (>2Mb). GD13a has a profile similar to that of the descendants of Anatolian farmers (i.e. early European farmers), suggesting that her ancestors also faced more benign conditions compared to populations further north during the LGM. This would have then allowed for populations which didn't undergo glaciation-induced shrinkages of population numbers, hence preventing the population bottlenecks that occurred in other populations further north, such as the Caucasus Hunter-Gatherers A bigger population size of origin would also have been subject to large population bottlenecks due to decreasing resources and harsher climatic conditions, hence why it would not fully explain the patterns seen in Figure 10. Superimposing the sampling locations of these groups onto climatic reconstructions from the LGM (Fig. 2b), however, does not reveal clear climatic differences among the regions. It is possible that the ancestors of the Anatolian and Ganj Dareh farmers spent the LGM in areas further south or east, which experienced milder climate. But it is also possible that they exploited local pockets of favourable climate (refugia). Whilst high elevation sites in the Zagros were abandoned during the LGM (Matthews and Nashli, 2013), there are a number of sites in the valleys that were occupied during that period and might have experienced more favourable conditions (Tsuneki, 2013).

The archaeological record indicates an eastward Neolithic expansion from the eastern regions of the Near East into Central and South Asia (Harris et al., 2010; Weeks et al., 2006). This analysis shows that both the Caucasus Hunter Gatherer Kotias and GD13a are plausible sources for the Eurasian Ancestry found in that part of Asia. Even though part of the Western Eurasian component found in India can be linked to Bronze Age migrations by dating the last contact using Linkage Disequilibrium (thus coming from the Kotias lineage), these results highlight the possibility of an older contribution from a source genetically close to GD13a (which would be hard to disentangle from the later gene flow

from the Steppe). Eventually, ancient DNA from the Indus Valley will be needed to detect conclusively whether any genetic traces were left by the eastward Neolithic expansion from the Near East, or whether this process was mostly cultural.

The presence of two distinct lineages (Anatolian-like agriculturalists and Zagros mountain herders) in the Near East at the beginning of the Neolithic transition raises an interesting question regarding the independence of innovations arising at different locations. Even within the Central Zagros, economies vary greatly in their rate and pathway towards Neolithisation (Matthews and Nashli, 2013). Ganj Dareh, in the high Zagros, has the earliest known evidence for goat domestication (Zeder, 2011, 2008; Zeder and Hesse, 2000), and the foothills of the Zagros mountains have also been argued to have been the site of early farming (Riehl et al., 2013). In addition, early sites such as Sheikh-e Abad (11.650-9,600 cal BP) provide evidence of early stages of barley cultivation (Whitlam et al., 2013). Were these innovations independent of similar achievements that made up the Neolithic package that North West Anatolians brought into Europe? Or were they exchanged culturally? If the latter, it would imply a cultural diffusion in the absence of much genetic interchange.

Methods

DNA extraction and library preparation

Sample preparation, DNA extraction and library preparation were carried out in dedicated ancient DNA facilities at University College Dublin. The dense part of the petrous bone was isolated, cleaned and sequenced following experimental procedures outlined in Gamba et al. (2014). DNA was extracted from 310 mg of ground bone powder using a double-digestion and silica column method as described in Gamba et al. (2016). Indexed Illumina sequencing libraries were constructed with a protocol based on Meyer and Kircher (2010) with modifications including blunt end repair using NEBNext End Repair Module (New England BioLabs Inc) and heat inactivation of Bst DNA polymerase (Gamba et al., 2014).

Sequence processing and alignment

Libraries were sequenced over a flow cell on a HiSeq 2000 at the TheragenEtex (South Korea) using 100 bp single-end sequencing. Adapter sequences, either complete or partial adapter matches, were trimmed from the 3' ends of sequences using cutadapt version 1.3 (Martin, 2011), conservatively requiring an overlap of 1 base pair (bp) between the adapter and the read. Reads were aligned using BWA (Li and Durbin, 2009), with the seed region disabled, to the GRCh37 build of the human genome with the mitochondrial sequence replaced by the revised Cambridge reference sequence (NCBI accession number NC_012920.1). Data from separate lanes were merged using Picard MergeSamFiles (http://picard.sourceforge.net/) and duplicate reads from the same library amplification were filtered using SAMtools rmdup (Li et al., 2009). Sequences were further filtered to remove those with mapping quality < 30 and length < 30 bp. Indels were realigned using RealignerTargetCreator and IndelRealigner from the Genome Analysis Toolkit (McKenna et al., 2010). The first and last 2 bp of each read were soft-clipped to a base quality of 2. The average genome-wide depth of coverage was calculated using the *genomecov* function of bedtools (Quinlan and Hall, 2010). A summary of alignment statistics can be found in Table 1.

Authenticity of results

The data were assessed for the presence of typical signatures of post-mortem DNA damage (Briggs et al., 2007; Brotherton et al., 2007). The sequence length distribution of molecules was examined as outlined in Gallego-Llorente et al. (2015) (Fig. 2) while the prevalence of nucleotide misincorporation sites at the ends of reads was evaluated using mapDamage 2.0 and a random subsample of 1 million reads (Jónsson et al., 2013) (Fig. 3).

The mitochondrial contamination rate was assessed by evaluating the proportion of non-consensus bases at haplogroup defining positions in the mitochondrial genome (Gamba et al., 2014; Sánchez-Quinto et al., 2012b). Only bases with quality ≥ 20 were used. The X chromosome contamination rate could not be evaluated as the sample was determined to be female, using the script described in Skoglund et al. (2013).

Mitochondrial Haplogroup Determination

To determine to which haplogroup the mitochondrion of GD13a belonged, a consensus sequence was generated using ANGSD (Korneliussen et al., 2014). Called positions were required to have a depth of coverage ≥ 3 and only bases with quality ≥ 20 were considered. The resulting FASTA files were uploaded to HAPLOFIND (Vianello et al., 2013) for haplogroup determination. Coverage was calculated using GATK DepthOfCoverage (McKenna et al., 2010).

Dataset preparation for population genetic analyses

Genotypes were called in GD13a at sites which overlapped those in the Human Origins dataset published on Lazaridis et al. (2014), filtered as described in Jones et al. (2015) using GATK Pileup (McKenna et al., 2010). Triallelic SNPs were discarded and bases were required to have quality ≥ 30. For positions with more than one base call, one allele was randomly chosen with a probability equal to the frequency of the base at that position. This allele was duplicated to form a homozygous diploid genotype for each position called in GD13a. This method of SNP calling was also used for selected ancient samples described in Jones et al. (2015), Cassidy et al. (2016), Günther et al. (2015), Omrak et al. (2016) and Olalde et al. (2015). Genotype calls for these ancient samples were merged with calls from modern samples found in the Human Origins dataset and ancient samples provided in the Mathieson et al. (2015) dataset which also included genotype calls for previously published ancient samples (Allentoft et al., 2015; Fu et al., 2015; Gamba et al., 2014; Haak et al., 2015; Keller et al., 2012; Lazaridis et al., 2014; Olalde et al., 2015; Raghavan et al., 2014; Seguin-Orlando et al., 2014). To avoid biases caused by post-mortem DNA damage, only transversion sites were used for PCA, ADMIXTURE, f_3-statistics and *D*-statistics.

Principal component analysis

To explore GD13a and other ancient samples in the context of modern variation in Eurasia, I performed PCA with a panel of contemporary populations (196 contemporary populations, 145,004 transversion SNPs). The analysis was carried out using SmartPCA (Patterson et al., 2006); the components were loaded on the contemporary populations, and the ancient individuals were projected onto these dimensions (Fig. 4).

ADMIXTURE

A clustering analysis was performed using ADMIXTURE version 1.23 (Alexander et al., 2009), using the full panel of modern and ancient samples described above. SNPs in linkage disequilibrium were thinned using PLINK (v1.07) (Purcell et al., 2007) with parameters –indep-pairwise 200 25 0.5 (Haak et al., 2015), resulting in a set of 116,834 SNPs for analysis. Clusters (K) (2-20) were explored using 3 runs with fivefold cross-validation at each K with different random seeds. The minimal cross-validation error was found at K=17, but the error already starts plateauing from roughly K=10, implying little improvement from this point onwards (Fig. 8). The ADMIXTURE proportions are shown in Fig. 7 for all samples.

Outgroup f_3-statistics and *D*-statistics

Outgroup f_3-statistics and *D*-statistics were performed using the qp3Pop and qpDstat programs from the ADMIXTOOLS package (Patterson et al., 2012).

Neighbour-joining tree

We used a custom Matlab script to calculate pairwise pi from genome-wide genotype data using a panel of 22 individuals (from the dataset described above), including GD13a, representative ancient samples, and different modern populations from the same geographic area as GD13a, and generated an unweighted pair group method with arithmetic mean (UPGMA) tree using the seqlinkage function in Matlab's Bioinformatics Toolbox. (The MathWorks, Inc., 2016)

Runs of homozygosity

In order to examine runs of homozygosity (ROH) I used imputation to infer diploid genotypes in the sample following the method described in Gamba et al. (2014). I used GATK Unified genotyper (McKenna et al., 2010) to call genotype likelihoods at SNP sites in Phase 3 of 1,000 genomes project (The 1000 Genomes Project Consortium, 2015) (version 5a downloaded from the BEAGLE website, https://faculty.washington.edu/browning/beagle/beagle.html). Genotype likelihoods were called for alleles observed in the 1,000 Genomes Project and equal likelihoods were

set for positions with no spanning sequence data as well as positions where the observed genotype could be explained by deamination. Genotypes were imputed using Beagle 4.0 with default parameters in intervals of 1Mb (Browning and Browning, 2007). I imposed a genotype probability threshold of 0.99 (any SNP without a genotype exceeding this threshold had a missing genotype assigned) while converting to PLINK-format genotype data. These data were merged with the dataset used in Jones et al. (2015) and ROH analysis was carried out as outlined in Gamba et al. (2014) and Jones et al. (2015).

Phenotypes of interest

Genes associated with a particular phenotype in modern populations were explored in GD13a. Phenotypes were chosen from the commonly-described phenotypes in ancient individuals: such as eye, hair and skin pigmentation according to the Hirisplex prediction model, and lactase persistence phenotypes. Observed genotypes were called using GATK Unified genotyper (McKenna et al., 2010), calling alleles present in Phase 1 of 1,000 genomes dataset (Consortium, 2012) with base quality ≥ 20. As many diagnostic markers had 1-fold coverage or less, I also used imputation to infer genotypes at these positions. Imputation was performed using all 1000 genomes project populations (The 1000 Genomes Consortium, 2015), imputing at least 1Mb upstream and downstream of the SNP position (this interval was reduced for those variants within the first 1Mb of the chromosome). The Hirisplex prediction model (Walsh et al., 2013) was used to explore hair and eye colour (Table 7). For the observed data, if the sample had 1x coverage, the variant was called as homozygous for that allele. Hair and eye colour predictions were confirmed using imputed data.

3. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa

Abstract

Characterizing genetic diversity in Africa is a crucial step for most analyses reconstructing the evolutionary history of anatomically modern humans. However, historic migrations from Eurasia into Africa have affected many contemporary populations, confounding inferences. This chapter presents the complete ancient genome with 12.5x coverage of an Ethiopian male ('Mota') who lived approximately 4,500 years ago. This genome is used to demonstrate that the Eurasian backflow into Africa came from a population closely related to Anatolian farmers, also genetically close to Early Neolithic farmers, and who had colonized Europe 4,000 years earlier.

A version of this chapter has been published: Gallego-Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D.G., Bhak, J., Pinhasi, R., Manica, A., 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. Science 350, 820–822.

Introduction

The ability to sequence ancient genomes has revolutionized our understanding of human evolution. However, genetic analysis of ancient material has focused on samples from temperate and arctic regions, where ancient DNA is preserved over longer time frames (Hofreiter et al., 2015).

Africa had so far failed to yield skeletal remains with much aDNA, with the exception of a few poorly preserved specimens from which only mitochondrial DNA could be extracted (Morris et al., 2014). This is particularly unfortunate, as African genetic diversity is crucial to most analyses reconstructing the evolutionary history of anatomically modern humans, by providing the baseline against which other events are defined. In the absence of ancient DNA, geneticists predominately rely on contemporary African populations, but a number of historic events, in particular a genetic backflow from West Eurasia into Eastern Africa (Pagani et al., 2012; Pickrell et al., 2014), act as factors that can confound the genetic analysis of modern African populations using exclusively modern African genomes..

Populations of hunter-gatherer humans have been present in southern and eastern Africa for hundreds of thousands of years (Phillipson, 2005). However, during the last two to three millennia, neolithisised pastoralist and agriculturalist groups joined the hunter-gatherer populations, then present in that part of the continent. Nowadays, southern and eastern Africa are inhabited by genetically and culturally diverse populations with different origins (Fig 1). Written history has only been recently introduced in southern Africa, which means that some of the biggest trends in migrations and population history of the region have uniquely been elucidated by archaeological studies and linguistics.

In 2014, Pickrell et al. used genome-wide genetic data to describe two different events of population admixture in the Khoi San populations of southern Africa. The Khoi San populations are non-Bantu-speaking, southern African populations, who mostly tend to be either hunter-gatherer or pastoralist. The most visible admixture event, genetically speaking, was the arrival of Niger-Congo-speaking populations originally from western Africa, into eastern Africa, followed by southern Africa. This arrival of Niger-Congo populations (termed the 'Bantu expansions', after the Bantu language family), was produced between 900 and 1,800 years ago. However, a first, older event was also seen: dating to between 2,700 and 3,300 years ago, was also seen. This event brought western

Eurasian ancestry into eastern Africa, and in particular, into Kenyan, Ethiopian and Tanzanian populations.

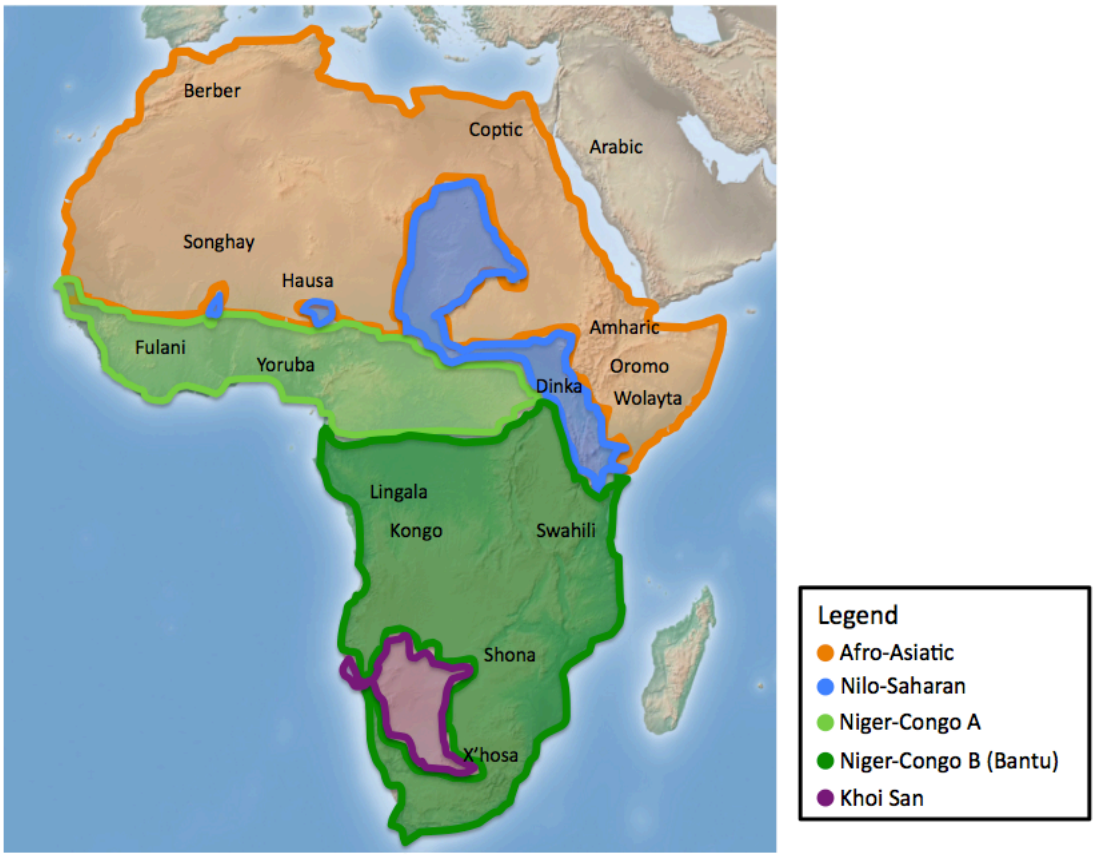


Fig. 1. Distribution of Language Families in Africa, with representative examples of each family.

This chapter presents an ancient human genome from Africa, and uses it to disentangle the effects of this recent population movement from western Eurasia into Africa. By sampling the petrous bone (Gamba et al., 2014), the genome of a male from Mota Cave was sequenced (herein referred to as 'Mota'). Mota was excavated from a cave in the southern Ethiopian highlands. The sequenced genome has a mean coverage of 12.5x. His remains were carbon dated to ~4,500 years ago (calendar date), and thus predate both the Bantu expansion (Li et al., 2014), and, more importantly, the 3ky-old West Eurasian backflow which has left strong genetic signatures in the whole of Eastern and, to a lesser extent, Southern Africa (Pagani et al., 2012; Pickrell et al., 2014).

Ethiopia, additionally, has always been on the center of research dealing with the origin and evolution of modern humans. Some of the earliest hominin species, *Australopithecus* amanensis (from around 4 million years ago), and *Australopithecus afarensis* ("Lucy", from 3-3.9 million years ago) were both unearthed in Ethiopia (Johanson and White,

1979). Furthermore, the earliest anatomically modern human remains, such as Omo 1, from 195k years ago (McDougall et al., 2005), and the *Homo sapiens idaltu*, 154-160k years ago, (White et al., 2003) were also found within modern-day Ethiopia.

Additionally, given Ethiopia's geographical position between Africa and the Eurasian landmass Ethiopia is key for our understanding of events such as the western Eurasian backflow described by Pickrell et al in 2014. In 2012, Luca Pagani published a dataset of modern Ethiopian, Sudanese and Somali populations, which became a stepping stone for the understanding of the modern genetic diversity in the region. However, the population history of the region has traditionally been very incomplete, due to the sparseness of information and sources in Ethiopia in the last 100k years. For example, while there have been local developments in both farming and agriculture (such as the domestication of the cereal teff, enset and coffee) (Phillipson, 1998), there have been external influences that have arrived during the last 5k years, such as the consumption of barley and wheat, and arrived through Egypt via early trade links (Pankhurst, 1998; Phillipson, 1993). Apart from trade and prehistorical and historical contacts between Ethiopians, Egyptians, Somalians, Arabs and populations from the rest of the African continent, another line of evidence of the complexity of Ethiopian prehistory is its linguistic diversity. There are two main linguistic families spoken in Ethiopia today: Afro-Asiatic languages and Nilo-Saharan languages. This linguistic diversity in Ethiopia is likely the result of demographic events happened during the last 10k years, such as the relatively recent arrival of Nilo-Saharan speaking populations from current-day Sudan (Ehret, 1995).

The Horn of Africa is very likely the homeland of the Afro-asiatic languages, as their present day diversity is by far greater in this region (all the Afro-asiatic languages in the Middle East are of Semitic origin, a sub-clade of Afro-asiatic languages which actually also has representatives in Ethiopia) (Ehret et al., 2004). This was studied by Diakanoff (1998), who explicitly described proto-Afroasiatic vocabulary as consistent with non-food- producing and links it to pre-Neolithic cultures in the Levant and in Africa south of Egypt, noting the latter to be older. He placed the origin of the Semitic branch in an area shared by Northern Egypt, the Sinai and Levant, but kept the remaining two branches in current-day Ethiopia, a view supported by Kitchen et al. in 2009. These arguments were at one point challenged by Diamond and Bellwood (2003), who suggested that food production and the Afroasiatic language family were brought simultaneously from the Near East by demic diffusion. However, the Afro-asiatic family has the Cushitic and Omotic

sub-clades, which not only are self-contained in the Horn of Africa, but also show the greatest linguistic divergence of the language family (Table 1 and Fig. 2).

More recent demographic changes likely resulted in the current presence of Nilotic languages, which are more widespread in Sudan and other regions of the continent (Blench, 2006).

Afro-Asiatic Languages			**Nilo-Saharan Languages**
Semitic	Cushitic	Omotic	Nilotic
Amhara Tygray	Afar Somali Oromo	Wolayta Ari	Gumuz Anuak South Sudanese

Table 1. Language families, subfamilies and individual languages present in Ethiopia and the region of the Horn of Africa. Colours assigned here will be used throughout the chapter.

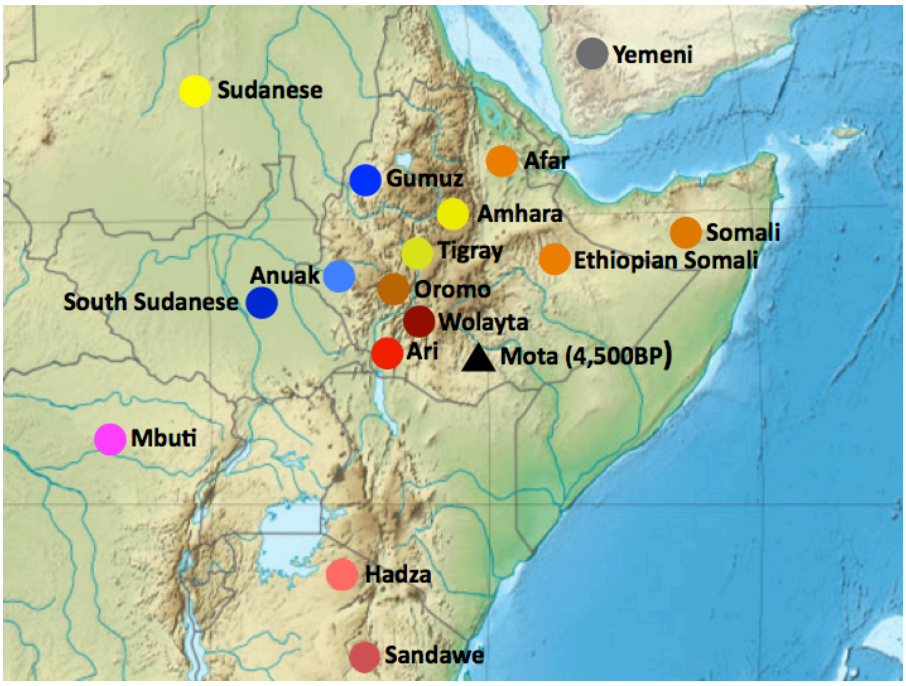


Fig. 2: Populations from Ethiopia and the surrounding area included in this study, in the context of Mota.

It was seen in Kivisild et al. (2004), that around half of the modern Ethiopian mtDNA haplotypes originate from outside Africa, and not from African mtDNA variation, hinting to a sizeable influx of non-African ancestry into Ethiopia and the Horn of Africa region. Around one fifth of Y haplotypes, as well, originate from lineages of putative non-African origin (Semino et al., 2002). These questions provide potential admixture events with implications for history, anthropology, linguistics, and genetics.

The study of the first African ancient whole genome was motivated by four main questions: Firstly, what is the position of a 4,500 year-old individual in the genetic landscape of the African continent? Secondly, what is the extent of the genetic continuity between ancient Ethiopians and modern Ethiopians? Thirdly, in the context of putative demographic events, such as the advent of Neolithic technologies into Eastern Africa, the expansions of Afro-asiatic languages, the Bantu expansions, or the making of the current landscape of Ethiopian ethnic variation, can an Ancient genome provide a picture of the genomic landscape previous to these events? And fourthly, if that is the case, can we use this genome to elucidate the origin and direction of these population movements?

I therefore analysed Mota, together with a dataset of populations from Pickrell et al. (2014), in order to understand the genetic situation of Mota in an African and global context, and answer the open questions regarding African regional population continuity, the historical aspects of Ethiopian population diversity, and the origin and extent of the latest movements back-to-Africa.

Results

The petrous bone of Mota yielded a genome with a coverage of 12.5x, which means that every position in the genome is covered by an average of 12.5 reads. Mota, additionally, has a low contamination rate of <1.3%, identified by evaluating the discordance in the rate of heterozygous calls between known polymorphic sites on the X chromosome and their adjacent sites. As the X chromosome is a haploid marker in males, any discordance may be a function of contamination. (Fig. 3). Mota was determined to be male using the script described in Skoglund et al. (2013), by considering the ratio of sequences aligning to the X and Y chromosomes.

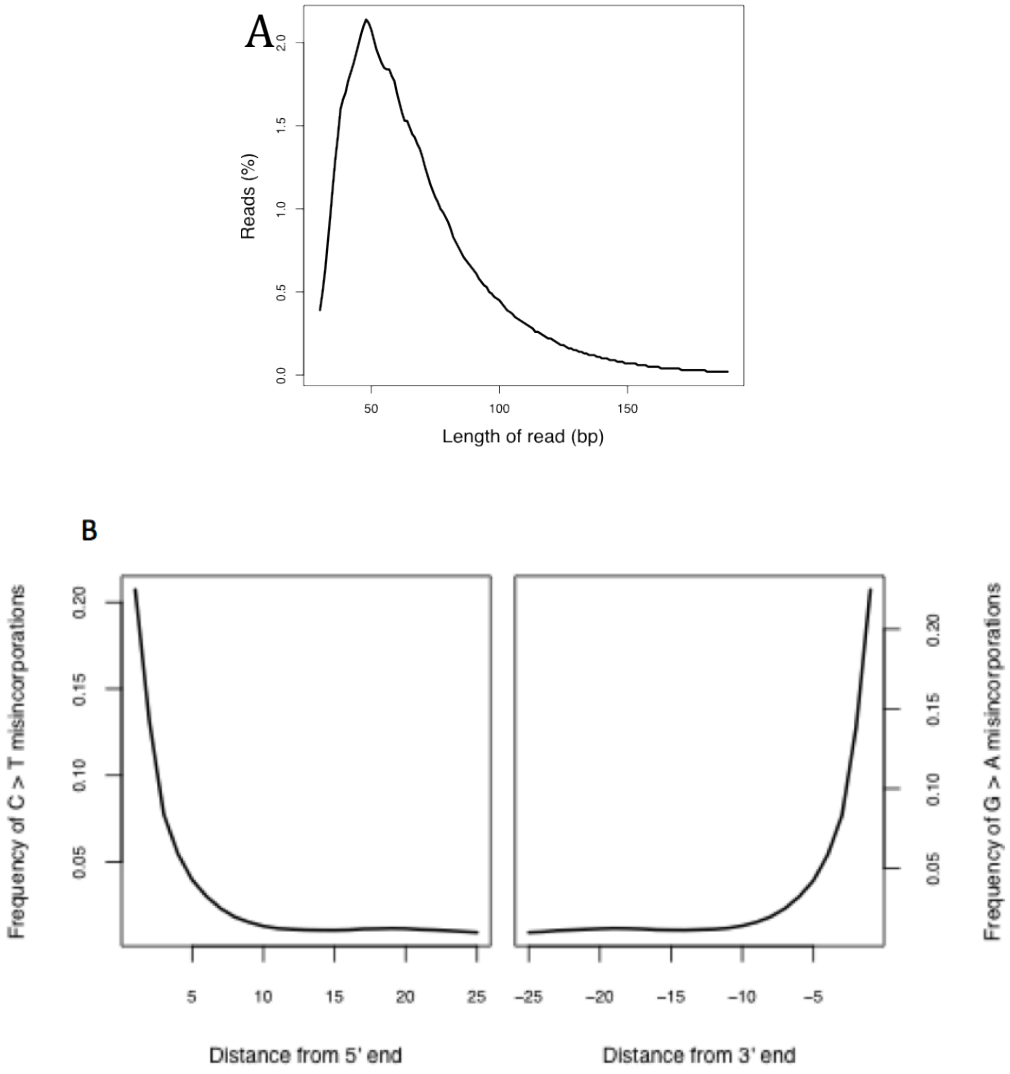


Fig 3. Sequence length distribution and patterns of molecular damage. Plots show mismatch frequency relative to the reference genome as a function of read position. The left hand figure shows the frequency of C to T misincorporations at the 5' ends of

reads (first 25 bases) while the right hand figure shows the frequency of G to A transitions at the 3' ends of reads (last 25 bases).

Mitochrondrial haplogroup

Mota was assigned to mitochondrial haplogroup L3x2a (Table 2). Haplogroup L3 arose 60-70 kya (Soares et al., 2012) in Eastern Africa where the richest present-day haplogroup diversity is found (Torroni et al., 2006). All mitochondrial haplogroups found outside Africa descend from the L3 lineage and hence this haplogroup is associated with the spread of *Homo sapiens* out of Africa to the rest of the world (Behar et al., 2008). The subhaplogroup L3x2 is restricted to the Horn of Africa and the Nile Valley in modern Ethiopian samples (Kivisild et al., 2004), suggesting a degree of maternal continuity in Ethiopia over the past 4,500 years.

Sample	Haplogroup	Haplotype
Mota	L3x2a	146T 150T 152T 195T 200G 247G **249de**l 494T 769G 825T 1018G 1442A 2758G 2885T 3435T 3483A 3594C 4104A 4312C 4769A **5899insC** 6401G 7146A 7256C 7521G 8311C 8468C 8655C **8817G** 9941G 10101C 10664C 10688G 10810T 10819G 10915T 10978G 11063T 11914G 13105A 13276A 13506C 13650C 13708A 15283C 15301A 16129G 16169T 16187C 16189T 16193T 16195C 16223C 16230A 16278C 16311T 16519T

Table 2. Mitochondrial haplogroup and haplotype for Mota.

Y-chromosome haplogroup

The Y chromosome haplogroup of Mota was assigned to haplogroup E1b1. This haplogroup was verified by looking for mutations in Mota that were described by the International Society of Genetic Genealogy (ISOGG) as defining the branches leading to haplogroup E1b1 (Table 3).

Macrohaplogroup E is the most prevalent haplogroup found in Africa with reduced frequencies in Europe and the Middle East (Semino et al., 2004; Trombetta et al., 2011). It is proposed to have originated in the East of Africa 21,000-32,000 years ago (Gebremeskel and Ibrahim, 2014; Semino et al., 2004; Trombetta et al., 2011). Mutation E-P2 (Table 3), present in Mota, represents the most widespread subclade of haplogroup E and has been found at high frequency in modern Ethiopians (Semino et al., 2002).

Haplogroup	Mutation	Alternative names	rs ID	Position (GRCh37 build)	Mutation	Mota (depth and base)
E	M96	PF1823	rs9306841	21778998	C->G	4G
E1	P147	PF1938	rs16980577	21083420	T->A	5A
E1b	P177	PF1939	rs16980473	14159846	C->T	6T
E1b1	P2	PF1940; PN2	rs9785756	21610831	G->A	3A
E1b1	P178		rs9786105	7401836	G->A	8A
E1b1	P179		rs16980621	14060308	A->C	10C
E1b1	P180	PF1941	rs9786035	18601274	G->A	6A
E1b1	P181		rs9785940	17394111	C->G	4G

Table 3. Mutations defining the E1b1 Y-haplogroup of Mota.

Ari and Sandawe are the closest contemporary populations to Mota

To explore this first Ethiopian ancient genome in the context of modern variation in Africa and the Middle East, I performed principal component analysis (PCA) with a global panel of 75 contemporary populations form Africa and the rest of the world, first published by Pickrell et al. (2014), in order to identify the populations closest to Mota in a wider context. The analysis was carried out using SmartPCA (Patterson et al., 2006); the components were loaded on the contemporary populations, and Mota was projected onto these dimensions.

Mota was placed close to the Ethiopian samples (Fig. 4), in between the clusters formed by the Ari and the Sandawe (but very close to an Ari individual that stands out from the rest of that group). The Ari can be split into two castes, Ari Cultivator and Ari Blacksmith, which share a common origin within the last 4,500 years (van Dorp et al., 2015). Since Mota was placed remarkably close to modern Ethiopian samples, and since data on a larger number of SNPs are available for 14 Ethiopian and Horn of Africa populations (Pagani et al., 2012), I repeated the PCA using this higher quality dataset, which gave 484,161 usable SNPs that could be called in Mota. Once again, Mota fell in between the Ari and the Sandawe cluster (Fig. 5).

I used outgroup f_3-statistics to estimate the amount of shared drift between Mota and contemporary populations, in order to look for the extent of population continuity in the Ethiopian region for the last 3,500 years, and the amount of external influences on the genetic composition of the region over this time frame. I computed $f_3(X, \text{Mota} ; \text{Ju|'hoansi})$, where X is a population from the Pickrell global panel, and Ju|'hoansi (Khoisan) acts as an outgoup. f_3-statistics were calculated with the 3PopTest program from the ADMIXTOOLS

package (Patterson et al., 2012).

Ari (which can be split into two castes, AriCultivator and AriBlacksmith), have by far the greatest genetic affinity to Mota (Fig. 6, 7). The Ari speak a language classified as Omotic, which is the most differentiated branch of the Afro-asiatic languages. Gumuz, a population member of the Nilo-saharan family, also shows a high level of shared drift with Mota, but significantly less than the Ari. Sandawe, which are closer to Mota in the PCAs, do not show high shared drift with Mota in the f_3, possibly because they are closer to the Khoi-San populations than the other Eastern African populations. Mota confirms the view that this divergent language family is the result of relative isolation of its speakers (Blench, 2008), and indicates population continuity over the last ~4,500 years in this region of Eastern Africa.

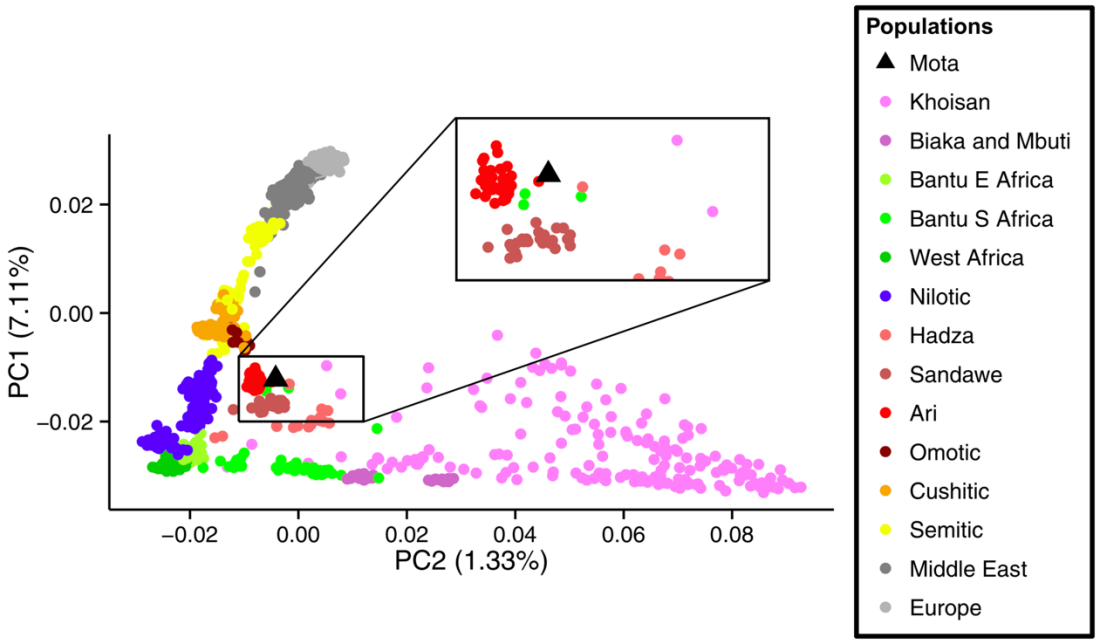


Fig. 4. Mota shows a very high degree of similarity with the highland Ethiopian Ari populations. PCA showing Mota projected onto components loaded on contemporary African and Eurasian populations. The inset magnifies the PCA space occupied by Ethiopian and Eastern African populations.

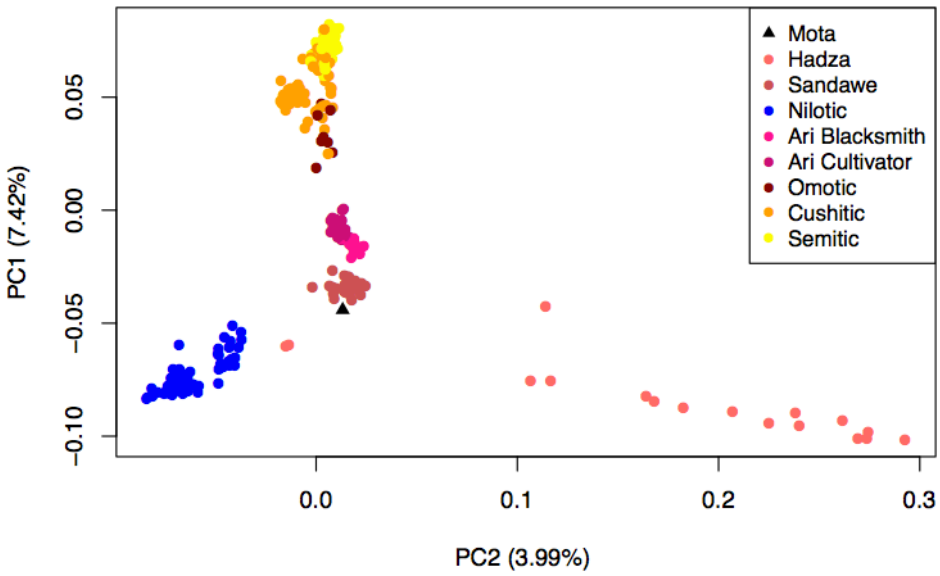


Fig. 5. PCA showing the relationship between Mota and contemporary Ethiopian populations. Components were loaded on contemporary Ethiopian populations using ~480k SNPs, with Mota projected on these dimensions.

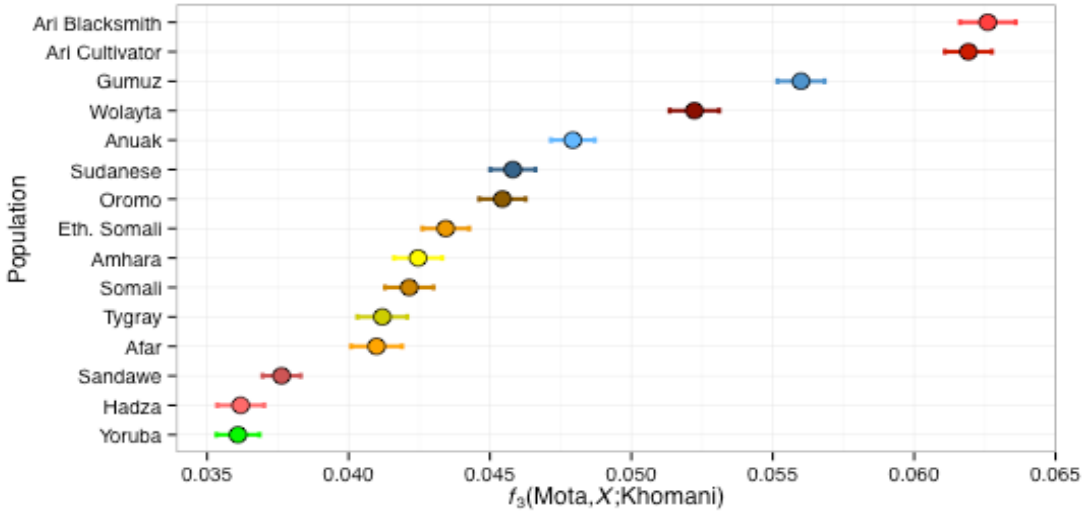


Fig. 6. Outgroup f_3 quantifying the shared drift between Mota and contemporary African populations, using Khomani (Khoisan) as an outgroup. Bars represent standard error. Populations speaking Nilo-Saharan languages are marked with blue shades, Omotic speakers with red, Cushitic with orange, Semitic with yellow, and Bantu with green. Mota is denoted by a black symbol.

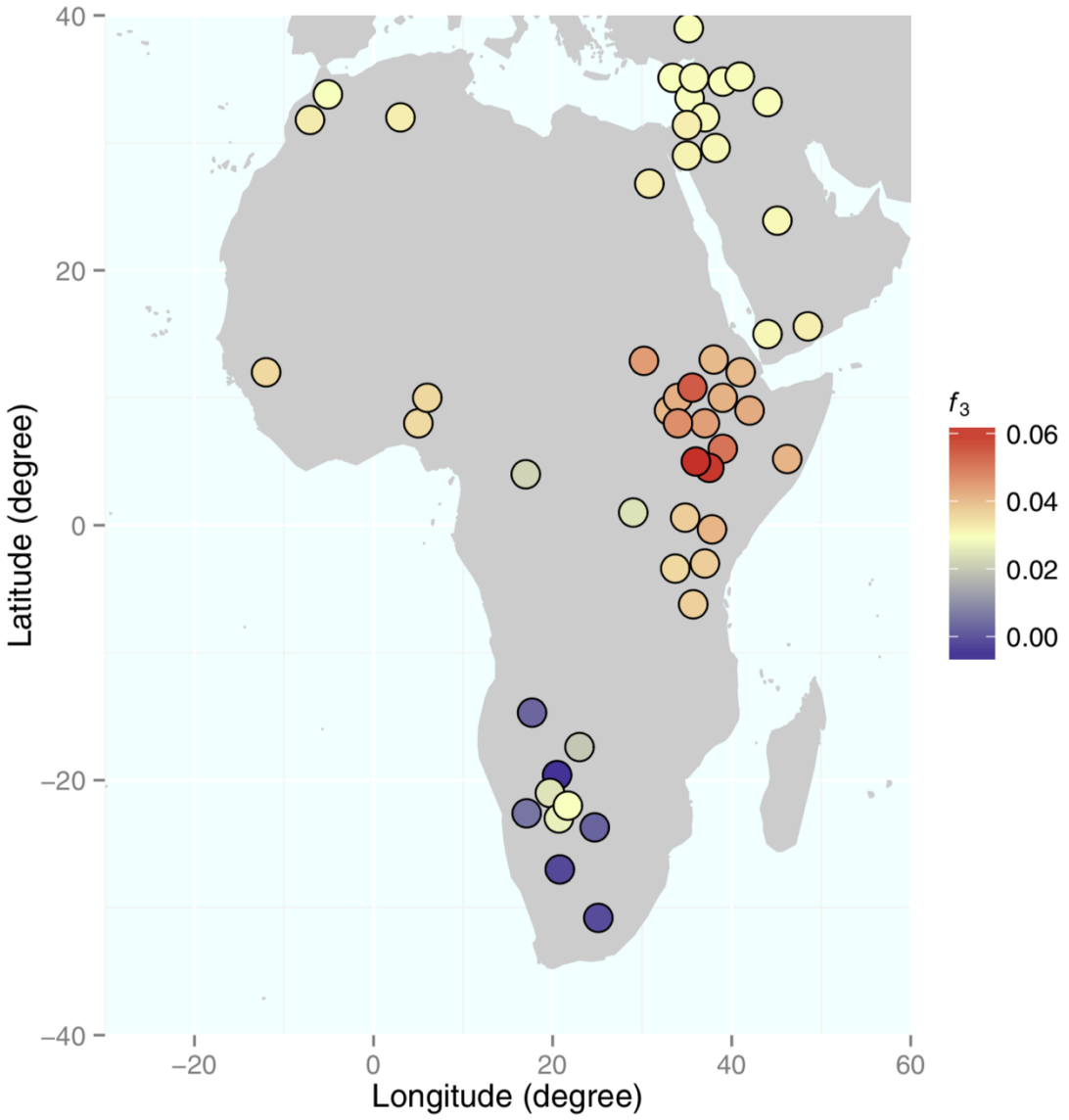


Fig. 7. Map showing the distribution of outgroup f_3 values across the African continent, f_3(Mota, X; Khomani). Populations that share the largest amount of shared drift with Mota are represented in red colour, whereas the populations with a smaller amount of shared genetic drift are represented in blue.

Figures 4-7, therefore, show that there has been a large extent of population continuity over the region of Eastern Africa in general, and Ethiopia in particular. Although it is possible that external influences have affected the genetic composition of the region, the populations currently present in the region show a large degree of similarity with Mota, who inhabited the area 3.500 years ago.

f_4 ratio analysis shows that Mota has no component of West Eurasian admixture

The age of Mota means that he should predate the putative West Eurasian backflow, which has been dated to ~3,000 years ago (Pagani et al., 2012; Pickrell et al., 2014).

I used f_4 ratio analysis (Patterson et al., 2012) to formally assess the extent of back-migration to Africa by West Eurasians, using the same logic adopted by Pickerell et al (Pickrell et al., 2014), who quantified the West Eurasian component in Africa, using Yoruba as a representative of a non-admixed African reference, and Druze as the source of non-african (i.e. western Eurasian) ancestry into eastern Africa. This was calculated with the ratio f_4(Han, Orcadian; *X*, Druze) / f_4 (Han. Orcadian; Yoruba, Druze), where *X* is a contemporary African population or Mota. However, since Druze has a small level of West African ancestry, this f_4 ratio is biased and does not show the desired fraction of West Eurasian component. To correct for this, I define $\lambda_{\text{Yoruba,Druze}}$ as the fraction of Druze-like (i.e. West Eurasian) ancestry population *X*, and *F* as the fraction of Yoruba-like (i.e. West African) ancestry in Druze (estimated in other studies to be *F*=0.05) (Moorjani et al., 2011). We can then write:

f_4 (Han, Orcadian ; *X*, Druze) / f_4 (Han, Orcadian ; Yoruba ,Druze) = (1 - $\lambda_{\text{Yoruba,Druze}}$ − *F*) / (1 − *F*)

and solve for $\lambda_{\text{Yoruba,Druze}}$ for each population *X*.

I decided to use Yoruba as a representative of a non-admixed African reference, and Druze as the source of western Eurasian ancestry, as this was needed in order to check Mota's situation relative to all modern populations analysed in the framework designed by Pickrell et al. (2014). Only this first test would enable me to identify a population to better represent this source of western Eurasian ancestry. I first checked that subsetting to the SNPs available for Mota did not affect estimates for contemporary African populations (Table 4), which are in line with those estimated by Pickrell et al. (2014) using all available positions (Pearson Correlation r=0.9998). Mota does not show any evidence of a West Eurasian component, with a $\lambda_{\text{Yoruba,Druze}}$ value that is close to zero (2.12%, ± 1.7%). This contrasts in particular with the Ari, their closest contemporary relatives, which show large West Eurasian components (17.8%±1.0% and 14.9%±1.2% for Ari Cultivator and Ari Blacksmith, respectively) (Fig. 8). I confirmed that such a difference is not due to a comparison of a single individual to population estimates by recomputing the f_4 ratio for each individual belonging to an Ethiopian population in the dataset (Fig. 9). The absence of a West Eurasian component in Mota supports the dating of the backflow into Africa,

which, at ~3.5kya, is younger than Mota (dated to 4.5 kya).

Given that Mota predates the backflow, it potentially provides a similar unadmixed African reference to contemporary Yoruba. Thus, I recomputed the extent of the West Eurasian component in contemporary African populations using Mota, $\lambda_{\text{Mota,Druze}}$, instead of Yoruba in the f_4 ratio (Table 4). Importantly, this analysis shows that the West Eurasian component can only be found also in East Africa. No West Eurasian component from this backflow is found in the Yoruba and Mbuti, which are often used a representative of an unadmixed African population

As expected, I did not find any West Eurasian component in Mota (Table 4), thus providing support for previous dating of the Eurasian backflow (Pagani et al., 2012; Pickrell et al., 2014).

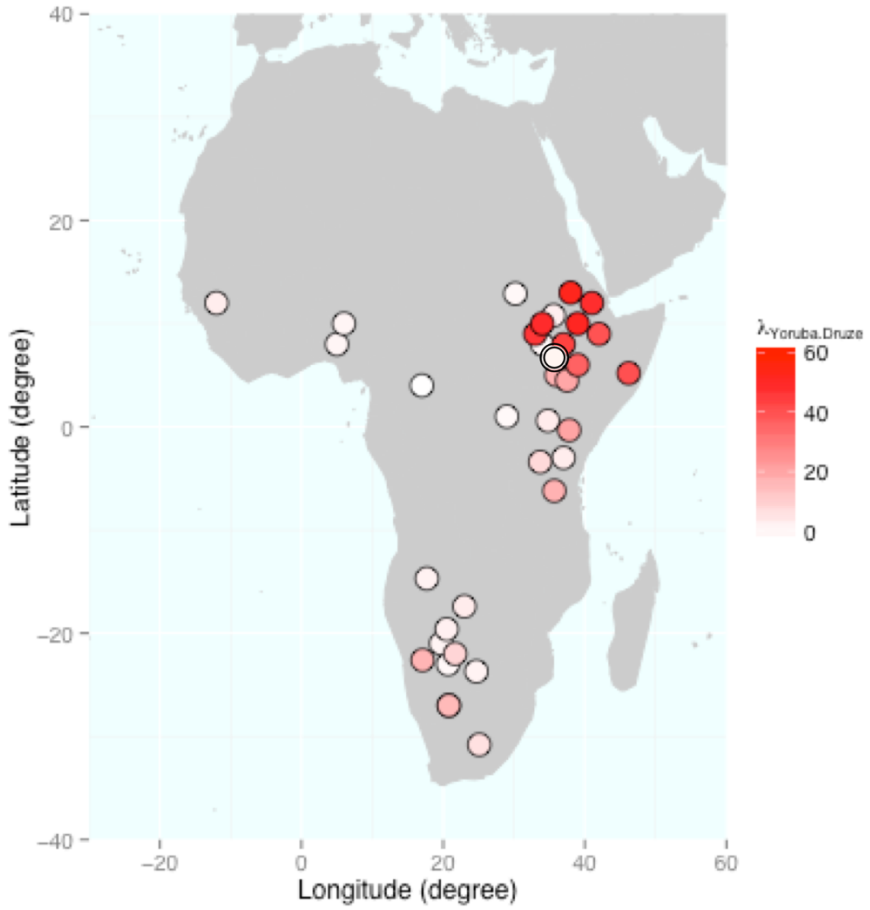


Fig. 8. The proportion of West Eurasian ancestry for all African populations in the global panel published by Pickrell et al. (2014). $\lambda_{\text{Yoruba,Druze}}$ gives estimates using Yoruba as the non-admixed reference and Druze as the source. Mota is shown with the double line.

Although no western Eurasian has been found in Mota, and although theis analysis shows a sizeable genetic input from western Eurasia into eastern Africa, it is worth mentioning that this western Eurasian backflow did not affect the entire continent, as it was initially reported in Gallego-Llorente et al. (2015). This first 2015 analysis was affected by a samtools/vcftools/PLINK compatibility issue, which skewed the results by deleting all reference homozyguous alleles in Mota, therefore skewing it away from the Eurasian range of variation. This issue showed some initial results that pointed towards a possible effect of the western Eurasian genetic backflow in distant populations such as Yoruba, Mandenka or Khoisan. Since the correction was made, Mota still does not show discernible western Eurasian component, but the results reflect that this backflow affected mostly eastern Africa.

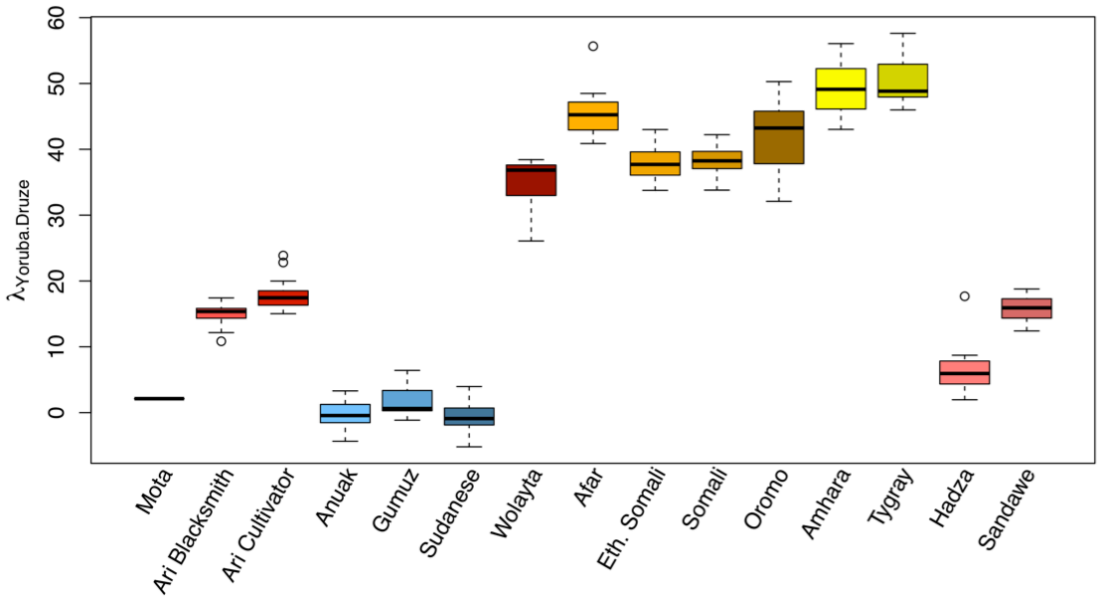


Fig. 9. The proportion of West Eurasian ancestry in modern eastern African populations. $\lambda_{\text{Yoruba,Druze}}$ (using Yoruba as the non-admixed reference and Druze as the source), estimated for individuals belonging to a number of Ethiopian populations.

Population	Language Family and Branch	$\lambda_{Yoruba,Druze}$ (%)	SE (%)	$\lambda_{Mota,Druze}$ (%)	SE (%)	$\lambda_{Mota,LBK}$ (%)	SE (%)
Mota	-	2.12	1.7	-	-		
Ari Blacksmith	Omotic (Afro-Asiatic)	15.67	1.0	14.06	1.6	11.75	1.3
Ari Cultivator	Omotic (Afro-Asiatic)	18.22	0.8	16.52	1.5	13.82	1.2
Wolayta	Omotic (Afro-Asiatic)	34.08	1.0	32.80	1.5	27.42	1.3
Afar	Cushitic (Afro-Asiatic)	45.96	0.9	44.74	1.3	37.50	1.4
Ethiopian Somali	Cushitic (Afro-Asiatic)	37.84	0.8	36.55	1.3	30.62	1.3
Somali	Cushitic (Afro-Asiatic)	38.43	0.8	37.11	1.3	31.10	1.3
Oromo	Cushitic (Afro-Asiatic)	41.59	0.7	40.41	1.2	33.88	1.2
Tygray	Semitic (Afro-Asiatic)	50.40	0.7	49.31	1.1	41.33	1.3
Amhara	Semitic (Afro-Asiatic)	49.14	0.7	48.09	1.1	40.28	1.3
Ethiopian Jews	Semitic (Afro-Asiatic)	46.57	0.9	45.52	1.2	38.15	1.4
Ethiopians	Semitic (Afro-Asiatic)	44.27	0.7	43.05	0.9	36.10	1.3
Sudanese	Nilotic (Nilo-Saharan)	-0.51	0.7	-2.65	1.8	-2.31	1.5
Gumuz	Nilotic (Nilo-Saharan)	1.68	0.8	-0.36	1.8	-0.31	1.4
Anuak	Nilotic (Nilo-Saharan)	-0.20	0.7	-2.40	1.8	-2.09	1.5
Maasai	Nilotic (Nilo-Saharan)	18.92	0.5	18.98	1.8	14.41	1.2
Hadza	Isolate	6.44	1.1	4.55	1.8	3.80	1.5
Sandawe	Isolate	15.79	0.7	14.04	1.6	11.74	1.3
Mbuti	Central Sudanic (Nilo-Saharan)	-1.71	1.0	-3.68	2.0	-3.14	1.6
Biaka	Bantu (Niger-Congo)	-0.66	0.7	-2.73	1.9	-2.36	1.5
Luhya	Bantu (Niger-Congo)	2.43	0.5	0.39	1.8	3.80	1.5
Bantu Kenya	Bantu (Niger-Congo)	2.24	0.7	0.15	1.9	11.74	1.3
South East Bantu	Bantu (Niger-Congo)	0.49	0.6	-1.57	1.9	-1.40	1.5
South West Bantu	Bantu (Niger-Congo)	7.54	0.7	5.56	1.9	4.60	1.5
Bantu South Africa	Bantu (Niger-Congo)	-0.49	0.7	-2.67	2.0	-2.32	1.6
Yoruba	Atlantic-Congo (Niger-Congo)	-	-	-2.13	1.9	-1.86	1.5
Yoruba (HapMap)	Atlantic-Congo (Niger-Congo)	0.27	0.4	-1.87	1.9	-1.60	1.5
Mandenka	Mande (Niger-Congo)	2.03	0.5	0.08	1.8	0.01	1.5
Khomani	Khoi-San	13.87	0.8	12.10	1.7	10.07	1.4
Karretjie	Khoi-San	5.02	1.0	3.25	1.9	2.68	1.5
Khwe	Khoi-San	2.49	0.7	0.50	1.9	0.37	1.5
GuiGhanaKgal	Khoi-San	0.81	0.9	-1.10	1.9	-0.96	1.5
Juhoansi	Khoi-San	1.18	1.1	-0.68	2.0	-0.70	1.6
Nama	Khoi-San	15.26	0.8	13.48	1.7	11.22	1.4
Xun	Khoi-San	0.89	1.0	-0.96	1.9	-0.85	1.6
Khomani Henn	Khoi-San	7.92	0.9	6.14	1.8	5.08	1.5

[Previous page] **Table. 4. The proportion of West Eurasian ancestry for all African populations in the global panel published by Pickrell et al. (2014).** $\lambda_{\text{Yoruba,Druze}}$ gives estimates using Yoruba as the non-admixed reference and Druze as the source, $\lambda_{\text{Yoruba,Druze}}$ using and Druze as the source, and $\lambda_{\text{Mota,LBK}}$ using Mota as the non-admixed reference and Stuttgart as a source. SE are the standard errors for these quantities.

Admixture f_3 statistics show that the West Eurasian component originated from a population similar to the early Neolithic farmers

Mota lived around 4,500 years ago. As the previous section has shown that it lacks the genetic component associated with the Eurasian backflow suggested by Pickrell et al. (2014), I searched for its most likely source by modelling the Ari, the closest contemporary population to Mota, as a mixture of Mota and another West Eurasian population. I do this by using the admixture f_3-statistics (Patterson et al., 2012) in the form $f_3(X, \text{Mota}; \text{AriCultivator})$, where X is a contemporary Eurasian population from the Pickrell global panel or one of the two then-available Eurasian ancient genomes. For the latter, I used a representative of Mesolithic hunter-gatherers (Loschbour), and one of the Early Neolithic farmers (Stuttgart, also known as LBK) (Lazaridis et al., 2014); these two genomes were chosen for their high coverage, allowing me to use most of the SNPs available for contemporary populations and Mota, given that Anatolian farmer genomes were not available at the time of publication.

In this analysis, contemporary Sardinians and the early Neolithic Stuttgart genome stand out as the most likely sources for this backflow (Fig. 10). Previous analyses have shown that Sardinians are the closest modern representatives of early Neolithic farmers (Sikora et al., 2014; Skoglund et al., 2012). Therefore, this result suggests that the backflow came from the same genetic source that fueled the Neolithic expansion into Europe from the Ancient Near East/Anatolia, before recent historic events changed the genetic makeup of populations living in that region. An analysis with haplotype sharing also identified a connection between contemporary Ethiopians and Anatolia (Kivisild et al., 2004; Pagani et al., 2012). Interestingly, archaeological evidence dates the arrival of Near Eastern domesticates (such as wheat, barley and lentils) to the same time period (circa 3,000 years ago) (Curtis, M.C., 2013; Harrower et al., 2010), suggesting that the direct descendants of the farmers that earlier brought agriculture into Europe may have also played a role in the development of new forms of food production in the Horn of Africa.

Using Mota as an unadmixed African reference and the early farmer Stuttgart as the source of the West Eurasian component, it is possible to reassess the magnitude and geographic extent of historical migrations, avoiding the complications of using contemporary populations. In Eastern Africa estimated an Eurasian backflow admixture in line with that detected by (Pickrell et al., 2014), while now the Eurasian source has been accurately located. More importantly, I quantified the impact of the backflow, using for the first time an unadmixed Eastern African (Fig. 11).

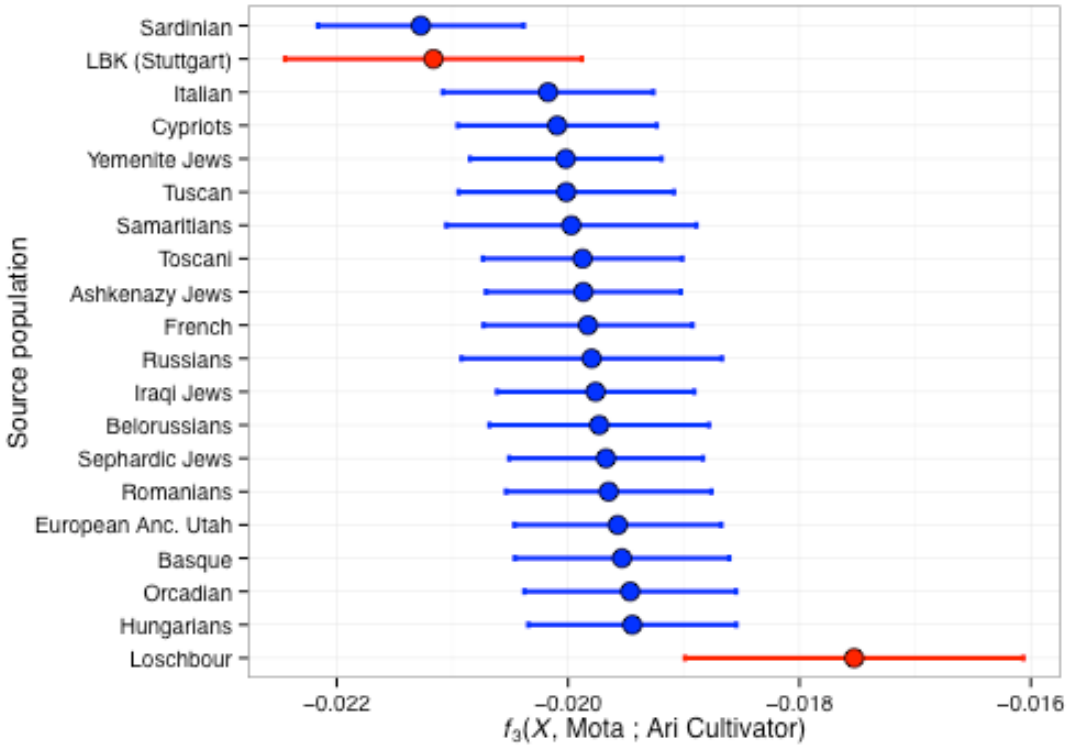


Fig. 10 Admixture f_3 identifying likely sources of the West Eurasian component (lowest f_3 values). Contemporary populations in blue, ancient genomes in red; bars represent standard error.

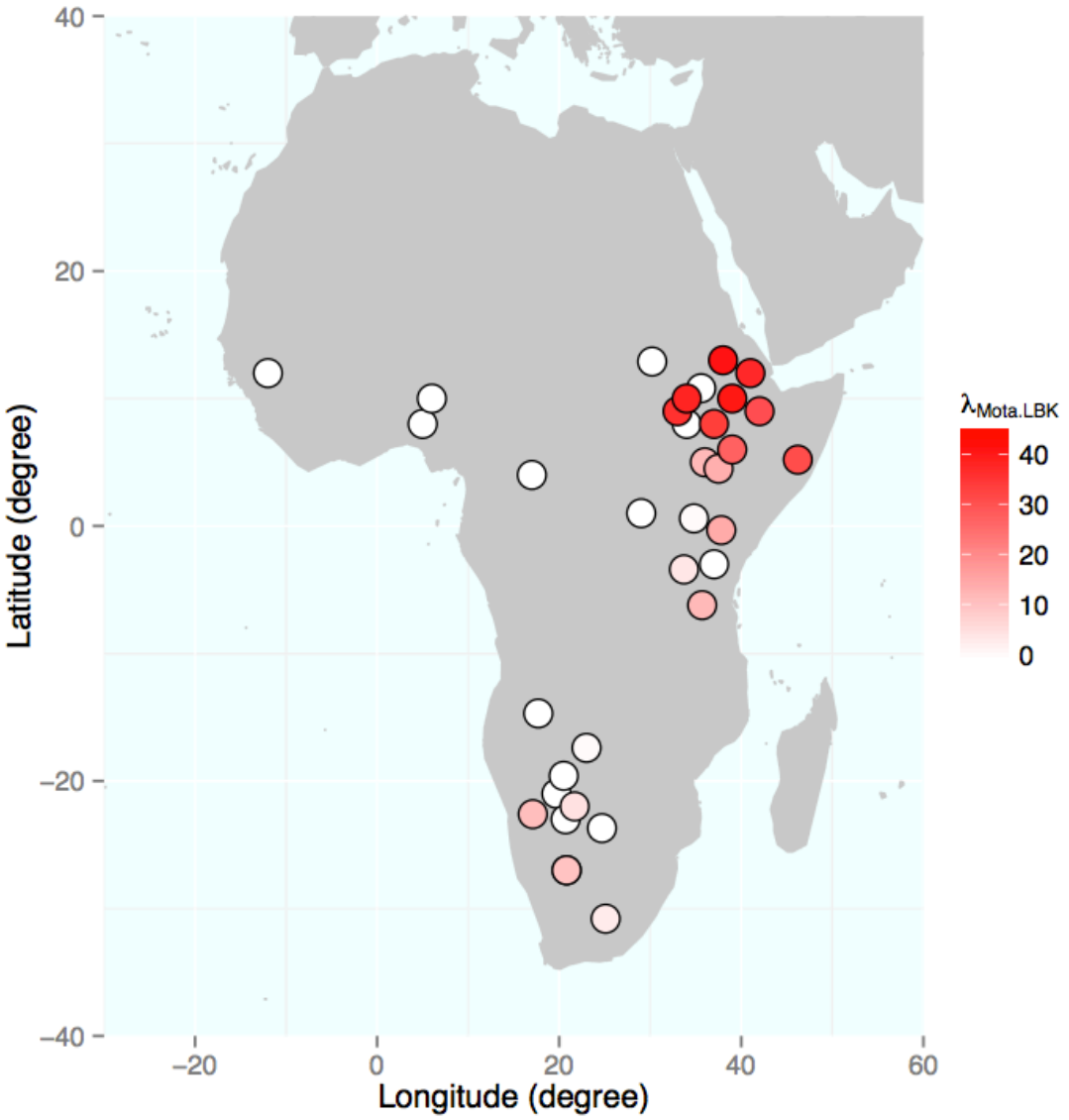


Fig. 11. Map showing the proportion of West Eurasian component, $\lambda_{Mota,LBK}$, across the African continent.

D statistics and f_4 ratios show that Mota has no discernible Neanderthal component

Since Mota predates recent demographic events, his genome can act as an ideal African reference to understand episodes during the out-of-Africa expansion. I used him as the African reference to quantify Neanderthal admixture in a number of contemporary genomes.

I performed this analysis using the complete genomes (rather than a subset of SNPs as in earlier analysis), since a large number of SNPs is needed to obtain accurate estimates.

D statistics have been used to detect gene flow between anatomically modern humans and other hominins (Green et al., 2010; Meyer et al., 2012b). Consider two contemporary populations (*A* and *B*), an ancient hominin (C), and an outgroup (D, often a chimpanzee). I can investigate differential hybridisation into *A* vs. *B* by computing:

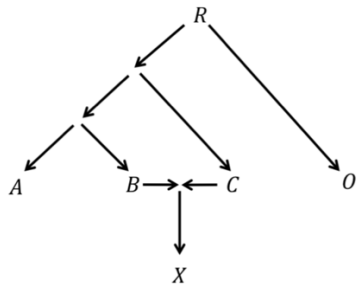
$$D(A,B;C,D) = \frac{n_{BABA} - n_{ABBA}}{n_{BABA} + n_{ABBA}}$$

where n_{BABA} represents the number of sites for which alleles are identical in *A* and *C*, and *B* and *D* respectively, but different in *AC* from *BD*. n_{ABBA}, similarly, is the number of sites for which alleles are identical in *A* and *D*, and *B* and *C* respectively, but different in *AD* from *BC*. Under the assumption that no gene flow occurred between the outgroup and any of the other populations, a positive *D(A,B;C,D)* statistic therefore indicates gene flow between *A* and *C*, whilst a negative value indicates gene flow between *B* and *C*.

The proportion of admixture into a given genome can be estimated using the f_4 ratio (Meyer et al., 2012b). This quantity is based on a ratio of f_4 statistics where f_4 is an unbiased estimate of the mean of allele frequencies in *A*, *B*, *C*, *D,* denoted respectively as a_{i}, b_{i}, c_{i}, and d_{i}, where *i* is the *i*th position out of *n*.

$$f_4(A,B;C,D) = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)(c_i - d_i)$$

Considering a phylogeny as described below:



Where population *X* is considered a mixture of *B* and *C*, and an outgroup O all descending from the same ancestor R, I can calculate the following f_4 ratio estimation:

$$\alpha = \frac{f_4(A,O;X,C)}{f_4(A,O;B,C)}$$

To compute the *D*-statistics, I first identified all positions at which the genomes of the Altai Neanderthal and the reconstructed Human-Chimpanzee Common Ancestor differ, giving ~19 million SNPs.

I computed *D*(*X*, Mota; AltaiNea, CommonAncestor), where *X* was French, Han, Yoruba, or Mbuti. CommonAncestor refers to the reconstructed alleles of the common ancestor of humans and chimpanzees. As expected, French and Han have significantly positive *D* values (Table 5), indicating that they are both more similar to Neanderthal than Mota is. The two African genomes, Yoruba and Mbuti, also have slightly positive *D* values, indicating that they are slightly more similar to Neanderthal than Mota is. This result is likely driven by the West Eurasian component found in modern Africans.

Test	***D* (Haploid, full data)**	***Z***	***D* (Haploid, transversions only)**	***Z***
Yoruba	0.017	4.56	0.002	0.52
Mbuti	0.021	5.29	0.002	0.37
French	0.039	8.50	0.033	6.34
Han	0.057	9.90	0.051	8.11

Table 5. Neanderthal component *D* statistics. *D*(AltaiNea, CAnc; Mota, *X*), where CAnc is the reconstructed human-chimpanzee common ancestor, Mota is the reference and *X* is the tested genome.

To compute the f_4 ratio, I first identified all positions at which the Altai Neanderthal (AltaiNea) genomes and Denisova differ (~3 million SNPs). All estimates were obtained both using the full data, as well as considering transversions only. (Table 6).

Test	% Neanderthal component (haploid, full data)	*Z*	% Neanderthal component (haploid, transversions)	*Z*
Yoruba	0.62 (± 0.50)	1.25	0.77 (± 0.64)	1.26
Mbuti	0.23 (± 0.45)	0.51	0.21 (± 0.72)	0.35
French	2.92 (± 0.55)	5.29	2.75 (± 0.61)	4.30
Han	2.90 (± 0.62)	4.71	2.96 (± 0.66)	4.09

Table 6. Neanderthal component based on f_4 ratio. f_4 (AltaiNea, Denisovan; *X*, Mota) / f_4 (AltaiNea, Denisovan; *X*, MezNea), where Mota is the unadmixed reference and *X* is the tested population.

I computed the statistic f_4(Denisova, AltaiNea ; *X*, Mota) / f_4(Denisova, AltaiNea ; *X*, MezNea), where *Mota* is the unadmixed individual, and *X* is the target genome. Both Yoruba and Mbuti were shown to have a small Neanderthal component (Table 6), in line with their West Eurasian ancestry. As expected, estimates for French and Han were slightly higher than than either of the two contemporary African genomes (from 0.21% in Mbuti to 2.96% in Han).

***D* statistics to quantify Denisovan component in Mota.**

To compute the percentage of Denisovan component in Mota, I first identified all positions at which the Denisova genome and the reconstructed Human-Chimpanzee Common Ancestor differ, obtaining ~19 million SNPs.

I first computed *D*(*X*, Yoruba; Denisova, CommonAncestor), from now referred to as D_{Yoruba}, where *X* was French, Han or Mota. As expected, Mota did not show any Denisova component (Table 7). Also, recomputing *D* but using Mota instead of Yoruba as the African reference (D_{Mota}) did not lead to any noticeable changes in the estimates of Denisova components in contemporary populations out of Africa (Table 7).

D statistic	D_{Yoruba}(Denisovan, CAnc; Yoruba, *X*)				D_{Mota}(Denisovan, CAnc; Mota, *X*)			
Target	***D* (Haploid, full data)**	***Z***	***D* (Haploid, transversions only)**	***Z***	***D* (Haploid, full data)**	***Z***	***D* (Haploid, transversions only)**	***Z***
French	-0.009	-2.14	-0.002	-0.42	0.002	0.51	-0.003	-0.74
Han	0.002	0.41	0.006	1.05	0.013	2.91	0.004	0.82
Yoruba	-	-	-	-	0.010	2.68	-0.002	-0.39
Papuan	0.058	10.70	0.061	10.25	0.071	12.76	0.062	10.02
Mota	-0.010	-2.68	0.002	0.39	-	-	-	-

Table 7. Denisovan component *D* statistics. D_{Yoruba}**,** *D*(Denisovan, CAnc; Yoruba, *X*), where Yoruba is the reference and *X* is the tested genome, and D_{Mota}**,** *D*(Denisovan, CAnc; Mota, *X*), where CAnc is the reconstructed human-chimpanzee common ancestor, Mota is the reference and *X* is the tested genome

Both Yoruba and Mbuti, which are routinely used as African references for this type of analysis (Fu et al., 2015; Green et al., 2010), show a marginally closer affinity with Neanderthal than Mota based on *D* statistics, and an f_4 ratio analysis suggested a small Neanderthal component in these genomes around 0.2-0.7%, consistent with the magnitude of their Western Eurasian ancestry. Whilst the magnitude of Neanderthal ancestry in these contemporary African populations is not enough to change conclusions qualitatively (estimates of Neanderthal ancestry in French and Han only increased marginally when tested with Mota as a reference), it should be accounted for when looking for specific admixed haplotypes (Sankararaman et al., 2014) or searching for unknown ancient hominins who might have hybridized with African populations (Hammer et al., 2011).

Phenotypic information from Mota

I also investigated the Mota genome for a number of phenotypes of interest, namely phenotypes on skin, eye and hair pigmentation, lactose tolerance, and altitude adaptation, as Mota lived in a high-altitude region.

The 8-plex (Hart et al., 2013) and Hirisplex systems (Walsh et al., 2013) were used to predict skin, eye and hair colour. Genotypes, which had a minimum of 3x coverage in Mota were used and only bases with quality ≥20 were considered. Skin colour could not be determined although Mota did not have common European variants associated with light skin colour (rs16891982 and rs1426654, see Table 8). Using the Hirisplex prediction system (Table 9), Mota was determined to have had brown eyes (p-value = 0.997) and dark (p-value = 0.996), probably black (p-value = 0.843) hair.

Mota lacked any of the known alleles that give lactose tolerance (Jensen et al., 2011; Tishkoff et al., 2007) (Table 10).

However, Mota possessed all three selected alleles that have been recently shown to play a role in adaptation to altitude in contemporary highland Ethiopian populations (Udpa et al., 2014) (Table 11). LIPE, for which Mota has homozyguous alternate alleles in two key positions, is a hormone-sensitive lipase, with roles in the mobilization of glycerol and fatty acids from adipose cells, and has been identified by Udpa et al. (2014) as being selected for in high-altitude areas. UBAP2, another gene with variants selected in high-altitude areas, plays a role in uniquitination of proteins, which affects the cellular stress response and hypoxia, and immunity (Bergink and Jentsch, 2009). The presence of these mutations supports the conclusion that Mota is the descendant of highland dwellers, who have lived

in this environment long enough to accumulate adaptations to the altitude (Alkorta-Aranburu et al., 2012; Huerta-Sánchez et al., 2013).

Gene	Chrom.	Position	Marker	Genotype Mota	Coverage Mota
SLC45A2	5	33951693	rs16891982	CC	10C
IRF4	6	396321	rs12203592	CC	15C
SLC24A4	14	92773663	rs12896399	GT	13G 9T
OCA2	15	28187772	rs1545397	AA	15A
HERC2	15	28365618	rs12913832	AA	17A
SLC24A5	15	48426484	rs1426654	GG	17G
MC1R	16	89986154	rs885479	GG	3G
ASIP	20	32785212	rs6119471	CG	7C 4G

Table 8. Genotypes for SNP panel used in the 8-plex prediction system. Genotypes are reported with respect to the GRCh37 build of the human genome. Skin colour determination was inconclusive.

Gene	Chromosome	Position	Marker	Genotype Mota	Coverage Mota
SLC45A2	5	33958959	rs28777	CC	18C
SLC45A2	5	33951693	rs16891982	CC	10C
EXOC2	6	457748	rs4959270	CC	11C
IRF4	6	396321	rs12203592	CC	15C
TYRP1	9	12709305	rs683	CA	5C3A
TYR	11	88911696	rs1042602	CC	14C
TYR	11	89011046	rs1393350	GG	11G
KITLG	12	89328335	rs12821256	TT	13T
SLC24A4	14	92801203	rs2402130	AA	20A
SLC24A4	14	92773663	rs12896399	GT	13G9T
OCA2	15	28230318	rs1800407	CC	7C
HERC2	15	28365618	rs12913832	AA	17A
MC1R	16	89985753	N29insA	-	-
MC1R	16	89986091	rs11547464	GG	18G
MC1R	16	89986154	rs885479	GG	3G
MC1R	16	89986144	rs1805008	CC	3C
MC1R	16	89985844	rs1805005	GG	17G
MC1R	16	89985918	rs1805006	CC	8C
MC1R	16	89986117	rs1805007	CC	12C
MCIR	16	89986546	rs1805009	GG	19G
MC1R	16	89986122	Y1520CH	-	-
MC1R	16	89985940	rs2228479	GG	11G
MC1R	16	89986130	rs1110400	TT	9T
PIGU/ASIP	20	33218090	rs2378249	AA	11A

Table 9. Genotypes for SNP panel used in the Hirisplex prediction system. Genotypes are reported with respect to the GRCh37 build of the human genome.

Chr	Pos (hg19)	Ref/Alt	rs ID	Traditional Name	Gene	Mota
2	136608646	C-G/A	rs4988235	−13910*T	MCM6	G/G
2	136608643	G/C	rs41525747	−13907*G	MCM6	G/G
2	136608651	A-T/C	rs41380347	−13915*G	MCM6	A/A
2	136608746	G/C	rs145946881	−14010*C	MCM6	C/C

Table 10. SNPs analysed for Lactase Persistence (Jones et al., 2013). For each SNP, the chromosome and physical position, the reference and alternate (LP) alleles, dbSNP code, the gene name, and the genotype for Mota are given. Note that traditional names refer to alleles on the reverse strand.

Chr	Pos (hg19)	Ref/Alt	rs ID	Gene	Mota
9	34017106	C/T	rs1785506	UBAP2	T/T
19	42906914	G/T	rs7246232	LIPE	T/T
19	42931004	A/G	rs16975750	LIPE	G/G

Table 11. SNPs possibly related to altitude adaptation. Nonsynonymous SNPs identified by Udpa et al. (2014) as potentially involved in hypoxia tolerance (and thus altitude adaptation). For each SNP, the chromosome and physical position is given, the reference and alternative(selected) alleles, dbSNP code, the gene name, and the genotype for Mota.

Until now, it has been necessary to use contemporary African populations as the baseline against which events during the worldwide expansion of Anatomically Modern Humans are defined (Green et al., 2010; Eriksson et al., 2012; Eriksson and Manica, 2012; Pagani et al., 2015). By obtaining the first ancient whole genome from this continent, I have shown that having an unadmixed reference that predates the large number of recent historical migrations can greatly improve this inference. Mota allowed a reassessment the magnitude and spread of the West Eurasian backflow into Africa. Whilst this event had already been detected by studying contemporary genomes (Pagani et al., 2012; Pickrell et al., 2014), its true extent had been strongly underestimated, and even populations that had been previously described as unadmixed were shown to harbour a consistent proportion of Eurasian, and thus Neanderthal, ancestry. This result stresses the importance of

obtaining unadmixed baseline data to reconstruct demographic events, and the limitations of analyses that are solely based on contemporary populations.

Discussion

I aimed to study the following questions: Firstly, what is the position of a 4,500 year-old individual in the genetic landscape of the African continent? Secondly, what is the extent of the genetic continuity between ancient Ethiopians and modern Ethiopians? Thirdly, in the context of putative demographic events, such as the advent of Neolithic technologies into Eastern Africa, the expansions of Afro-asiatic languages, the Bantu expansions, or the making of the current landscape of Ethiopian ethnic variation, can an Ancient genome provide a picture of the genomic landscape previous to these events? And fourthly, if that is the case, can we use this genome to elucidate the origin and direction of these population movements?

Mota was situated very close to the rest of modern Ethiopians in a global PCA, and especially, very close to the Ari populations within Ethiopia, who speak a language classified as Omotic, which is the most differentiated branch of the Afro-asiatic languages. This showed that populations in Ethiopia have been geographically very stable, and although current-day Afro-asiatic-speaking populations have a relatively large percentage of western Eurasian admixture, Mota is the only example of a non-admixed population that forms a clade with Afro-asiatic speakers.

A proposed model would be that the Horn of Africa, home to the Afro-asiatic language family, received an input from western Eurasian farmers after Mota lived. As Ehret et al. in 2014 proposed, the Horn of Africa has been the homeland of at least the Omotic and Cushitic branches of the Afro-asiatic languages, whereas the Semitic branch probably diverged in a more northern area, surrounding the Siani peninsula and the southern Levant, only spreading southwards by demic diffusion (Diakanoff, 1998, and Kitchen et al., 2009). This is compatible with the results of this analysis, and a possible suggestion would be to link this southward expansion of western Eurasians with the current spread of Semitic languages in Ethiopia, as these populations present percentages of western Eurasian component of around 50% (such as the Tygray and the Amhara). As the Omotic and Cushitic language branches have likely been continuously spoken in Ethiopia for the entire period, and as Mota shares most similarity with the Ari population, who speak an Omotic language, it is likely that Mota spoke an early version of an Omotic language. The

input from western Eurasian peoples, therefore, permeated Ethiopia in a north-to-south gradient, mostly affecting current-day Semitic-speaking populations, and in a lesser degree Omotic and Cushitic populations. Current-day Nilotic populations are however unaffected by this admixture event, very probably because their arrival into Ethiopia from modern-day Sudan was after this western Eurasian migration (Robertshaw, 1987). In addition, the Bantu expansions originating in the area comprising current-day Nigeria and Cameroon did not affect Ethiopia to a large extent, as the modern-day populations show.

Ethiopia and the southern Nile region already had had influences of the Neolithic by the time Mota lived, as evidence of animal husbandry (goats and cows) in the Horn of Africa 5k years ago (Clark and Prince, 1978; Marshall et al., 1984). However, given the extent of the western Eurasian migration into Africa, and given the fact that this is linked to changes in the archaeological record, it is likely that this permeation of western Eurasians in Eastern Africa signified a change in agricultural practices and probably the arrival of more recent grains into the region. The first archaeological evidences of wheat, barley, lentils and flax in the Horn of Africa, are consequently, from 2,500 years ago (Bard et al., 1997; Boardman, 1999; D'Andrea et al., 2008). Hence, it is very likely that these more modern, Middle Eastern cereals were brought by communities with the same origin as those who brought agriculture into Europe and Northern Africa. This would also support a relationship between this back-migration and a further development of agriculture and farming in the Horn of Africa.

Mota, however, is an individual which is 4,500 years old, which is, according to Pickrell et al (2014) the time when the back-migration of western Eurasians into Africa was being produced. Should an even earlier genome be sequenced, it would no doubt show population movements that Mota does not show. There is no reason to think that Mota is an unadmixed African, as there could be older back-migrations of peoples with Eurasian genetic drift into Africa, which affected Mota, but not genomes previous to these hypothetical migrations. This will only become clear once we start sequencing more ancient genomes of African origin. Since Anatomical Modern Humans left Africa ~60k years ago, until Mota, there are more than 55k years of which we have no genetic information. The future will only bring exciting developments in this area.

Methods

Sample preparation and sequencing

For Mota, all molecular work and data processing work prior to analysis was carried out in the dedicated ancient DNA (aDNA) facilities at Trinity College Dublin, Ireland, by Eppie Jones, co-author of the publication Gallego-Llorente et al., (2015). The petrous portion of the right temporal bone, which has been shown to be a good source of aDNA (Gamba et al., 2014; Pinhasi et al., 2015), was sampled from Mota. A Dremel engraving cutter attached to a dental drill was used to remove the surface of the petrous and the remaining bone was exposed to ultraviolet radiation (Biometra Biolink, 5 lamps at 254 nm) for 20 minutes. The bone fragment was then ground to a fine powder using a mixer mill (MM 400, Retsch) and DNA was extracted from this powder using a silica column based protocol (Gamba et al., 2014; Yang et al., 1998), Libraries were prepared and amplified with AccuPrimeTM Pfx Supermix (Life Technology), using a modified version of Meyer and Kircher (Meyer and Kircher, 2010) as outlined in Gamba et al (Gamba et al., 2014). To evaluate the human DNA content of the samples, libraries were screened on an Illumina MiSeq platform at TrinSeq, Dublin using 70 base pair (bp) single-end sequencing. Libraries were further sequenced on a HiSeq2000 platform at the Theragen BiO Institute (South Korea) using 100 bp single-end (8 lanes) and paired-end sequencing (1 lane).

Data Processing

For all data, adapter sequences were trimmed from the ends of reads using leeHom (Renaud et al., 2014) with the --ancientdna option implemented. The program leeHom was chosen for processing the reads of the Mota genome as they were noth single-end and paired-end reads. The Ganj Dareh genome (Chapter 2) was composed of only single-end reads, and hence cutadapt was used. For paired-end data leeHom was also used to merge mate pairs which could be overlapped. For mate pairs which could not be overlapped only data from read 1 were considered for downstream analyses. Libraries for Mota were double-stranded. Reads were aligned using BWA (Li and Durbin, 2009), with the seed region disabled, to the GRCh37 build of the human genome with the mitochondrial sequence replaced by the Cambridge reference sequence (NCBI accession number NC_012920.1). Data from the same sample were merged using Picard MergeSamFiles (http://picard.sourceforge.net/) and clonal reads were removed using MarkDuplicates from the same suite of tools. Reads were filtered to allow a minimum read length of 30 bp and indels were realigned using RealignerTargetCreator and IndelRealigner from the Genome Analysis Toolkit (McKenna et al., 2010). Average depth was calculated using the

genomecov function in bedtools (Quinlan and Hall, 2010). Reads were filtered using SAMtools (Li et al., 2009) to remove sequences with a mapping quality of less than 30 and reads were rescaled using mapDamage 2.0 (Jónsson et al., 2013) to reduce the qualities of likely damaged bases, therefore lessening the effects of ancient DNA damage associated errors on analyses (Jónsson et al., 2013).

Sequence length distribution and molecular damage

The data was examined for the presence of typical signatures of ancient DNA, namely short average sequence length and a prevalence of nucleotide misincorporation sites at the ends of molecules (Briggs et al., 2007; Brotherton et al., 2007), which were assessed using all available paired-end data. These data were derived from the same libraries as those used for single-end sequencing, however, these paired-end sequences are less likely to be truncated at their 3'-termini than their single-end counterparts. Sequence length distribution was evaluated using command (1) (see below) and patterns of molecular damage were assessed using mapDamage 2.0 (Jónsson et al., 2013). Only bases with quality ≥ 30 were considered when running mapDamage.

(1) samtools view <input.bam> | awk '{print length ($10)}'| sort -n | uniq -c

Length distribution (Fig 1) of the Ethiopian data was plotted, shich showed peaks at < 50 bp. This is compatible with an ancient origin for these data as ancient DNA molecules tend to have an average sequence length of less than 100 bp (Shapiro and Hofreiter, 2014). An increase in nucleotide misincorporation sites was also observed, at the termini of molecules, namely C to T and G to A transitions at the 5' and 3' ends of reads respectively. Misincorporation frequencies were greater than 17% at the ends of reads.

X chromosome contamination

The level of X chromosome contamination in Mota was assessed, using the 'Contamination' program in the ANGSD package (Korneliussen et al., 2014). This method, based on Rasmussen et al (Rasmussen et al., 2011b), evaluates the discordance in the rate of heterozygous calls between known polymorphic sites on the X chromosome and their adjacent sites. As the X chromosome is a haploid marker in males, any discordance may be a function of contamination. I used the HapMap resources provided with the ANGSD software to define polymorphic sites and restricted the analysis to the non-recombining portion of the X chromosome (X:5,000,000-154,900,000). Only considered bases with a minimum quality of 20 were considered. Two tests were performed; "test 1" which uses all reads and "test 2" which removes the assumption of independent error rates by only

sampling a single read per site. Mota shows a low contamination rate of 1.26 ± 0.04 % (p-value < 2.2×10^{-16}) using "test 1" and 1.17 ± 0.10 % (p-value < 2.2×10^{-16}) using "test 2".

Mitochondrial contamination

To assess the level of contamination in Mota's mitochondrial genome he number of secondary (non-consensus) bases at haplogroup defining positions was examined, as a function of the total coverage for each of these sites (Gamba et al., 2014; Sánchez-Quinto et al., 2012b). GATK Pileup (McKenna et al., 2010) was used to call genotypes at haplogroup defining positions that were determined using HAPLOFIND (Vianello et al., 2013). Only bases with quality > 30 were considered in analyses. The contamination rate was estimated omitting sites which could be explained by residual deamination ("C") as well as with all available high quality bases ("C + MD") (Gamba et al., 2014; Sánchez-Quinto et al., 2012b). Contamination was estimated to be 0.64% using all sites, and 0.29% excluding sites with potentially damaged bases.

Mitochondrial DNA haplogroup assignment

The mitochondrial haplogroup was determined following the analysis described by Skoglund et al in 2014 (Skoglund et al., 2014). In brief, this involved generating a consensus mitochondrial sequence using SAMtools (Li et al., 2009) and assigning a haplogroup using HAPLOFIND (Vianello et al., 2013) (Table 2).

Y chromosome haplogroups

A maximum likelihood-based approach was used to determine the Y chromosome haplogroup of Mota. Genotypes along the Y chromosome were called with a minimum base threshold of 20 using GATK and employed YFitter (Jostins et al., 2014) to predict the most likely haplogroup. Mota was assigned to haplogroup E1b1. This haplogroup was verified by looking for mutations in Mota that were described by the International Society of Genetic Genealogy (ISOGG) as defining the branches leading to haplogroup E1b1 (Table 3).

Mutations are reported with respect to the Reconstructed Sapiens Reference Sequence (Behar et al., 2012). Mutations found in Mota, which are present in the reported haplogroup are shown here unless marked in bold or underlined. Underlined mutations are those present in Mota but not associated with the haplogroup determined. Bold mutations are those expected for the assigned haplogroup but which are absent from the sample.

SNP calling in Mota and comparison to other ancient and contemporary populations.

Genotypes in Mota were called using mpileup from samtools (Li et al., 2009), filtering for sites with base quality ≥20 and mapping quality ≥30. I then compared Mota to contemporary populations in a dataset including a large number of African and several Eurasian populations typed at 256,540 sites (Pickrell et al, 2014 (Pickrell et al., 2012), herein referred to as the "global panel"). SNPs were flipped to the positive strand, using the hg19 fasta files as reference. The Mota VCF files were converted to PLINK format using VCFtools (Danecek et al., 2011). Triallelic positionswere then filtered, and positions for which there was no call for Mota, giving 127,069 SNPs, and merged these with the global SNP panel using PLINK (Purcell et al., 2007). The merged dataset was then used for later analysis, including Principal Component Analysis and outgroup f_3 to determine the affinity of Mota to contemporary populations, and f_4 ratios and admixture f_3 to quantify the magnitude and distribution of the West Eurasian component in Africa.

Neanderthal and Denisovan Component Determination

To compute *D* statistics and f_4 ratios, I used a number of high quality whole genomes from modern populations (Meyer et al., 2012): HGDP00456-Mbuti (24.3x), HGDP00521-French (26.7x), 00778-Han (27.7x), 00927-Yoruba (32.1x), (rounded average coverage in brackets). I used the alignments to hg19 available in BAM format from: http://www.cbs.dtu.dk/suppl/malta/data/Published_genomes/bams/. The Denisovan and Neanderthal genomes were obtained from the set of vcf files from Prüfer et al. (2014), also mapped to hg19, and contained additional information such as the inferred alleles of the Human-Chimpanzee common ancestor and the Human-Orangutan common ancestor (from now on, referred as "modified vcf files"). These files are available at http://cdna.eva.mpg.de/neandertal/altai/

***D* statistics to quantify Neanderthal component**

To compute the *D*-statistics, I first identified all autosomal positions at which the genomes of the Altai Neanderthal and the reconstructed Human-Chimpanzee Common Ancestor differ, giving ~19 million SNPs. This was done using the set of modified vcf files from Prüfer et al. (2014), as detailed above. This list of discordant genomic positions was then called in the contemporary genomes as well as Mota using samtools, using MAPQ≥30 for all genomes. The resulting vcf files were then converted to PLINK format using vcftools.

PLINK-format files (ped and map) of the different genomes were subsequently merged into a single file using PLINK. Genomes were turned haploid by randomly sampling either allele at heterozygote sites. All estimates were obtained both using the full data, as well as considering transversions only. The PLINK file was in turn converted to eigenstrat format, using the Admixtools utility CONVERTF (Patterson et al., 2012). *D*-statistics were estimated using the *D*-statistics software of Admixtools (Patterson et al., 2012).

f_4 ratio to quantify Neanderthal component

To compute the f_4 ratio, I first identified all autosomal positions at which the Altai Neanderthal (AltaiNea) genomes and Denisova differ, by using the set of modified vcf files from Prüfer et al. (2014) explained above. I obtained ~3 million SNPs. This list of discordant genomic positions was then called in Yoruba, French, Han, Mbuti, Mezmaiskaya Neanderthal (MezNea) (Ovchinnikov et al., 2000), and Mota whole genomes using samtools, using MAPQ≥30 for all genomes except for the low quality Mezmaiskaya Neanderthal (MAPQ≥37). The resulting vcf files were then converted to PLINK format, using vcftools. PLINK-format files (ped and map) of the different genomes were subsequently merged into a single file using PLINK. Genomes were turned haploid by randomly sampling either allele at heterozygote sites. All estimates were obtained both using the full data, as well as considering transversions only. The PLINK files were in turn converted to eigenstrat format using the Admixtools utility CONVERTF (Patterson et al., 2012). f_4 ratio estimation was performed using the F4RatioTest software of Admixtools. Analysis was carried either on both the full data and transversions only.

***D* statistics to quantify Denisovan component in Mota.**

To compute the *D* statistics, I first identified all autosomal positions at which the Denisova genome and the reconstructed Human-Chimpanzee Common Ancestor differ. Like for the Neanderthal component analysis, I performed this analysis using the complete genomes, since a large number of SNPs is needed to obtain accurate estimates. I obtained ~19 million SNPs. This was done using the set of modified vcf files from Prüfer et al. (2014). This list of discordant genomic positions was then called in the contemporary genomes as well as Mota using samtools (Li et al., 2009). The resulting vcf files were then converted to PLINK format using vcftools (Danecek et al., 2011). PLINK-format files were subsequently merged into a single file using PLINK (Purcell et al., 2007). This was converted to eigenstrat format, using the Admixtools utility CONVERTF (Patterson et al., 2012). D statistics were estimated using the D statistics software of Admixtools.

4. Analysis of the challenges in the compatibility of ancient genetic data of different sources and their solutions

Abstract

In recent years, advances in the recovery and high-throughput sequencing of DNA have revolutionised the field of ancient DNA, which has allowed us to access an ever-increasing number of ancient individuals. However, ancient DNA continues to be very difficult to work with, due to contamination from other organisms, contamination from modern human DNA, and DNA damage due to molecular degradation over thousands of years. Hence, different methods to approach recovery of DNA and its sequencing, the processing of reads, and different bioinformatic pipelines designed to tackle the problems of contamination and damage, will result in datasets that are not easily compatible. Here, I analyse the most patent problems in terms of DNA analysis, including the use of SNP panels with different ascertainment biases obtained by DNA capture, the combination of datasets obtained by different sequencing methods, the use of uracil-DNA-glycosylase (UDG) treatment to eliminate endogenous deamination damage, how different algorithms deal with DNA damage *in silico*, and how reference bias can affect the mapping of ancient samples to a reference genome. I propose that subsetting common SNPs in capture panels will eliminate ascertainment biases of different SNP panels. Using only transversion SNPs or performing soft-clipping can reduce problems arising from differential damage patterns in aDNA. Additionally, a method to align ancient samples to a human reference genome where the reference alleles of SNPs have been substituted by a third allele (i.e. not the usual alternate allele) has been shown to fully eliminate reference bias – I name this method “third-base mapping”.

The problem with ascertainment bias in SNP datasets

Introduction

In recent years, advances in the recovery and high-throughput sequencing of DNA have revolutionised the field of ancient DNA (aDNA). As a result, aDNA studies have rapidly progressed to whole-genome sequencing and the number of sequenced ancient genomes has increased exponentially. Shotgun sequencing was used for the first sequenced ancient genomes of Anatomically Modern Humans, such as Saqqaq (Rasmussen et al., 2010a), Motala (Skoglund et al., 2012), Loschbour, Stuttgart (Lazaridis et al., 2014), Mal'ta (Raghavan et al., 2014), etc. However, in 2013, David Reich and his team pioneered the extensive use of a method to sequence ancient genomes for prehistorical demographic analysis focusing on a previously-selected subset of SNPs. To make demographic analysis of large numbers of ancient individuals economically feasible, they developed in-solution hybridization capture for ancient nuclear DNA (Haak et al., 2015), which was previously assayed by Rohland and Reich in 2012, and then used by Fu et al. in 2013, although only in chromosome 21 of the 40k year old Tianyuan individual. Haak et al. enriched sequencing libraries to a target set of 394k SNPs, 354k of which were autosomal SNPs also genotyped by the Human Origin array (Patterson et al., 2012): this reduced the amount of sequencing by a median of 262-fold.

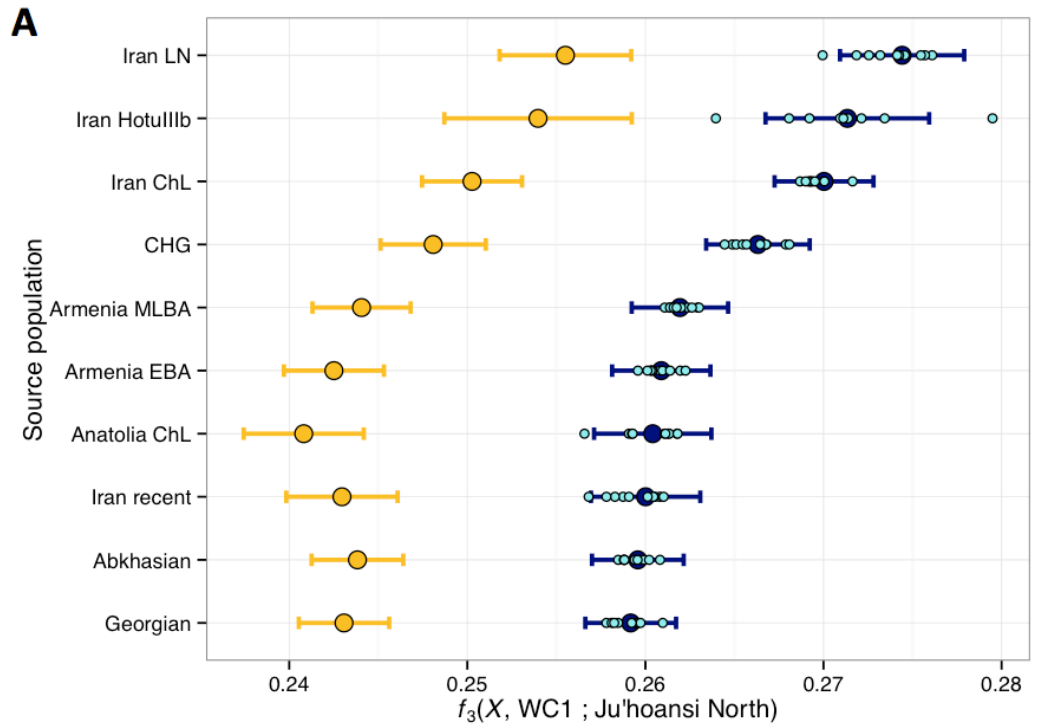
Therefore, for demographic studies, SNP capture can offer a more economical and efficient alternative (Haak et al., 2015). In 2015 a human enrichment target set with 394,577 single nucleotide polymorphisms (SNPs) ('390k capture'), most of which overlapped with the SNPs genotyped by the Affymetrix Human Origins array (HO array, or '600k capture') was published (Haak et al., 2015). This 390k capture dataset was then tripled to 1,240,000 SNPs ('1240k capture') (Mathieson et al., 2015), and subsequently expanded to 3.7 million SNPs ('3700k capture') (Fu et al., 2016).

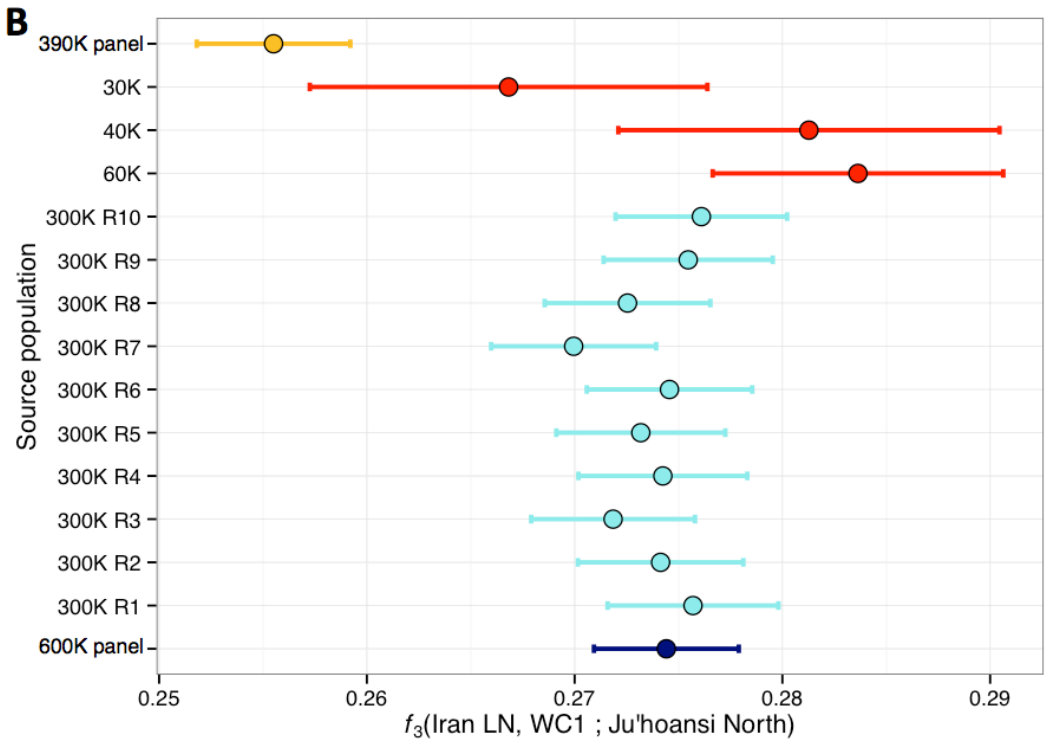
Because SNP genotypes necessarily contain a set of SNPs designed to reflect aspects of human genetic variation, any genotyping array will, by definition, be ascertained. This, in turn, means that different SNP genotyping arrays will have different ascertainment biases, and hence, different arrays will result in very different representations of human demographic history and selection (Lachance and Tishkoff, 2013). The SNP capture panels

used in aDNA pose no difference: every time a SNP is added into a panel, the ascertainment of the resulting panel subsequently changes. Hence, going from a 390k capture to a 1240k capture poses great difficulties in terms of keeping a similar ascertainment. The 390k capture panel was based on African samples: they ascertained SNPs on African samples to look at aDNA from the European Bronze Age, so that ascertainment bias would not affect theoir results. However, as these SNPs are based only on African variation, it presents potential ascertainment biases when compared with larger datasets. Additionally, the HO array (629k capture) contained more Eurasian populations, and was ascertained for a more global set of human variation. Finally, the 1240k capture was equally ascertained to reflect the full extent of human variation.

Results

I used the 1240k capture panel (Mathieson et al., 2015) merged to the original Human Origins Array (HO array, ~629k SNPs), and to a series of previously-published ancient genomes (such as Wezmeh Cave (WC1) (Broushaki et al., 2016)), to compare the effect of choosing different subsets of SNPs in different analyses. Specifically, I focussed on the outgroup f_3 statistics to compare individual and population affinities. I started by comparing f_3 values computed on the 390k and on the HO (629k) capture panels. Whilst populations are ordered in the same way in the two panels, the magnitude of the f_3 values differ greatly between the two panels (Fig. 1A). More specifically, f_3 values in any given f_3-statistics test differed greatly when done on the 390k panel and on the HO (629k) panel, by around 3-4 error bars. This is not simply due to the reduction in SNPs: subsampling a random set of 390k SNPs out of the 600k panel gave f_3 values similar to those obtained when using all the 600k SNPs (blue dots of Fig. 1A). I then went on to analyse the results of heavy random SNP subsampling, down to 30k randomly-selected SNPs, wanting to see what was the lower limit of SNPs necessary for an f_3 of this type to remain informative. Subsampling to 60k SNPs yielded f_3 results with error bars only slightly larger than the analyses done at 300K, which shows how much the number of SNPs can be reduced, while keeping the f_3 informative (Fig. 1B). I then proceeded to compare the 1240k panel against the 390k panel, obtaining the same results (Fig. 1C).


A
Iran LN
Iran HotuIIIb
Iran ChL
CHG
Armenia MLBA
Armenia EBA
Anatolia ChL
Iran recent
Abkhasian
Georgian
Source population
0.24
0.25
0.26
0.27
0.28
$f_3(X$, WC1 ; Ju'hoansi North)


B
390K panel
30K
40K
60K
300K R10
300K R9
300K R8
300K R7
300K R6
300K R5
300K R4
300K R3
300K R2
300K R1
600K panel
Source population
0.25
0.26
0.27
0.28
0.29
f_3(Iran LN, WC1 ; Ju'hoansi North)

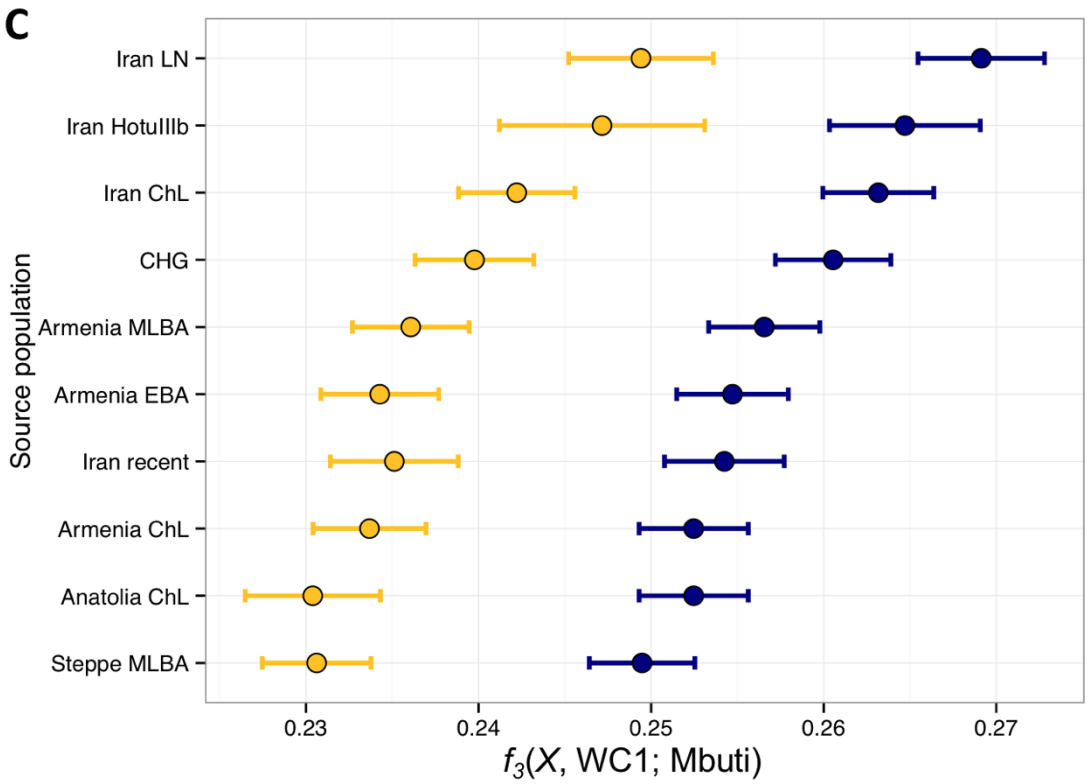


Figure 1. A. f_3(*X*, WC1; Ju'hoansi North) and **B.** f_3(Iran LN, WC1; Ju'hoansi North), where *X* is other ancient individuals and modern populations. The results in blue were performed using the 600k capture panel and the results in gold were performed on the 390k capture panel. The points in light blue show a random 300k SNP subset from the original 600k dataset. The points in red are random 60K, 40K and 30K subsets of the 600k dataset. **C.** f_3(Iran LN, WC1; Mbuti), where the results in blue are done over the 1240k panel, the results in gold are done on the 390k panel.

There have been recent publications (e.g. Fu et al., 2016) using a mix of samples from different ascertainment sets (3.7M, 2.2M, 1.2M and 390k). This poses the question of what to do in those cases; should one use the maximum number of SNPs available in each pairwise comparison, or subset to a common set of SNPs available in all samples? In **Figure 2**, I replicate an analysis done by Fu et al (2016) using all possible pairwise outgroup f_3 to describe the relationships among Ice Age European samples. I present two versions of the analysis, one using all samples subsetted to the smallest capture panel, in this case the 390k panel (**2A**), and one using the maximum number of SNPs available for each pair of samples (**2B**). Whilst using the common subset highlights similarities among samples from a similar archaeological context, as presented by Fu et al., the analysis using all available SNPs obfuscates these patterns.

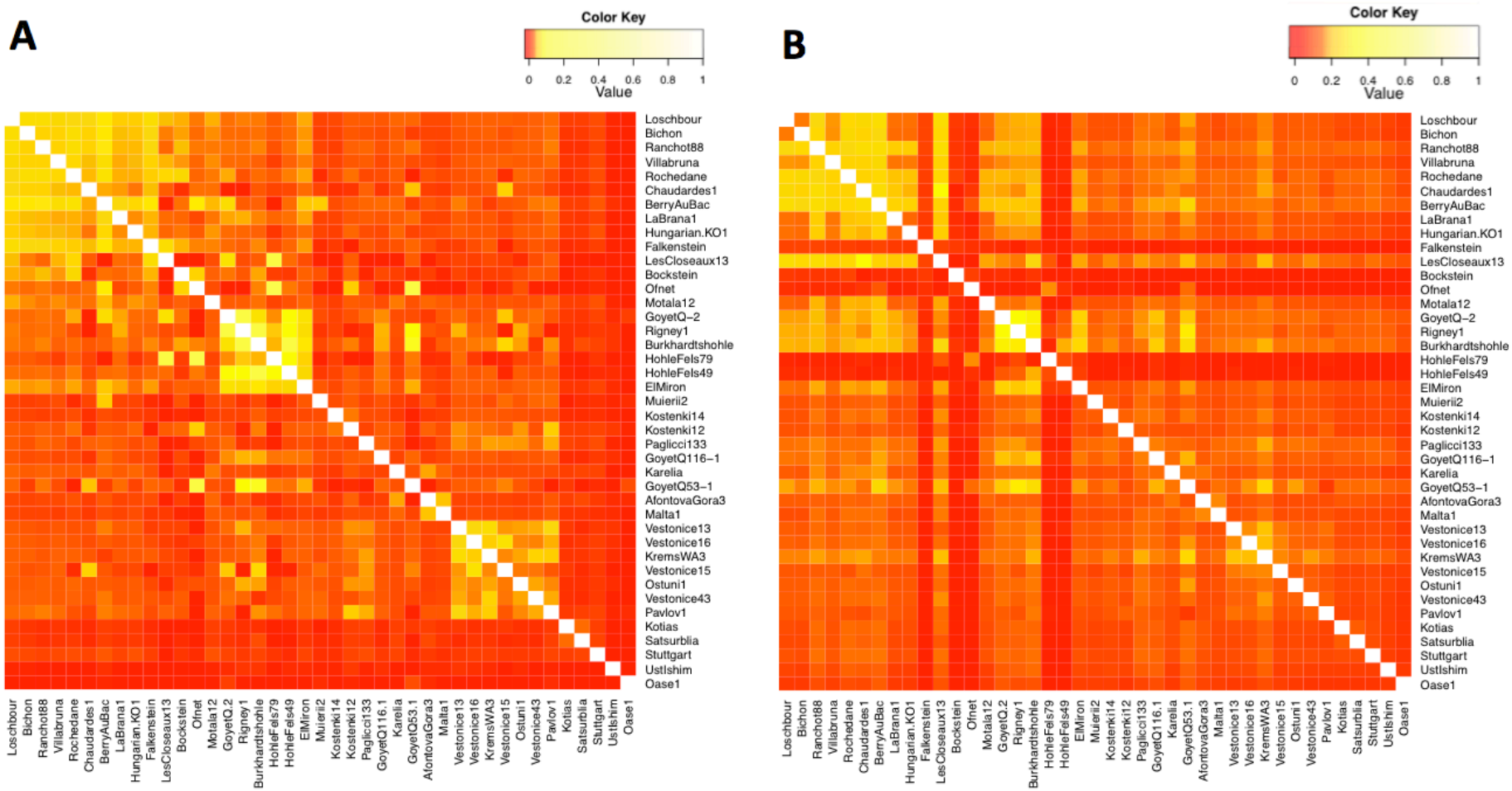


Figure 2. Subsetting to the common SNPs for all individuals in the datasaset. A. Pairwise $f_3(X, Y;$ Mbuti) performed on Fu's ancient individuals, after subsetting to only the SNPs in the 390k dataset. **B.** Pairwise $f_3(X, Y;$ Mbuti) done on Fu's ancient individuals. We can see how the capture panel of choice affects the result, for example showing lower f_3 values for those individuals for which only the 390k SNPs were captured (Bockstein, Ofnet, HohleFels79 or HohleFels 49).

Figure 2, therefore, shows that in individuals obtained by capture using different SNP platforms, we must subset the panel to SNPs common to all platforms in order to obtain comparable results. Otherwise, biases will appear, linking samples to other samples with the same ascertainment, and away from samples with a different ascertaiment. A representative example is constituted by the HohleFels individuals, captured with the 390k panel. They show a disproportionate lack of relatedness to the rest of individuals that have been captured at other resolutions 1240k or 2200k, which completely disappears when subsetting to the 390k panel

Discussion

The problem of mixing panels with different ascertainment is well known in modern samples typed with different SNP chips, but it has not been discussed explicitly in analyses of ancient genome captured for different numbers of SNPs. The analyses I presented here highlight that, unsurprisingly, ascertainment bias can have major effects on the conclusions reached on aDNA, and that care should be taken to always subset data to the smallest common set of SNPs.

In previous studies, it has been common to mix individuals from the 390k capture panel with individuals from the 600k capture panel. Additionally, the 2016 Fu et al study published European and Eurasian Ice Age individuals sequenced at different resolutions: 3.7M, 2.2M, 1.2M and 390k SNPs. This, however, causes an issue with platform compatibility. As **Figure 2 shows**, common SNPs between the panels need to be subset in order to have a meaningful analysis. Differences caused by platform bias are relatively small, but they are enough to cause differences in analyses looking at closely-related populations, as shown for the Ice Age Europe samples (Figure 2). Subsetting all samples to the 390k capture panel fully solves the problem. I recommend that the SNP set used in each of the analyses should be clearly indicated in future publications (ideally providing the commands used to subset the data), especially when individuals at different resolutions are used.

I have additionally shown that using a small random subset of SNPs is actually not a problem, and that most f3 analysis would withstand a downsampling to 300k SNPs (or even 60k SNPs) without causing a loss of resolution, which is especially important for distinguishing drift in closely-related populations (such as this analysis, done on Ice Age Europe) (Fig. 1B). Hence, we can conclude that making sure that the dataset is compatible

in terms of ascertainment should always take precedence over using the largest number of SNPs available for each individual in a single study.

The problem with platform bias

Introduction

A second problem regarding the integration of ancient data is the issue of sequencing mode. Shotgun sequencing for ancient DNA in its current form was approached and developed by studies such as Rasmussen et al., (2010); García-Garcerà et al., (2011); Sánchez-Quinto et al., (2012) and Skoglund et al., 2012). In 2014, (Gamba et al., 2014) discovered that sampling ancient DNA from the petrous section of the temporal bone, could yield 4 to 16 times more endogenous DNA than from teeth. This opened the way for extensive studies of human ancient DNA by whole genome sequencing, which have been followed in studies such as Allentoft et al., (2015); Gallego-Llorente et al., (2015); Gallego-Llorente et al., (2016); Jones et al., (2015); and Lazaridis et al., (2014). Whole genome sequencing of ancient samples paved the way for analyses that require diploid data to be carried out, such as looking at the runs of homozygosity (Gamba et al., 2015) and examining coalescence times between samples using GPhoCS (Jones et al., 2015; Kuhlwilm et al., 2016).

By far, the most common type of damage that occurs to DNA is deamination, which results in the replacement of a cytosine by a uracil. In the presence of water, deamination occurs spontaneously. Hence, samples found in humid conditions have higher amounts of this damage (Smith et al., 2003). During PCR amplification, the uracil is treated as if it was a thymine, hence rendering an original CG base-pair into an amplified TA base-pair. This type of damage has commonly been dealt with by treatment with uracil-DNA-glycosylase (UDG treatment), which excises the uracil from DNA, leading a site with no base. During PCR, these abasic sites will lead to strand breaks at these damaged sites, yielding shorter products and hence shorter reads.

While capture data is normally UDG-treated (Fu et al., 2016; Haak et al., 2015), most of the samples done by whole-genome sequencing are not UDG-treated (Broushaki et al., 2016;

Gallego-Llorente et al., 2016), as deamination damage can be dealt with in the data analysis steps, and also as UDG eliminates the damage patterns, which is one of the authenticity criteria used for ancient DNA. Some whole-genome-sequenced individuals, published by Gamba et al in 2014 (Gamba et al., 2014) were captured and published in the 2015 Mathieson paper (Mathieson et al., 2015). Additionally, the whole-genome-sequenced Iranian Farmer from Gallego-Llorente et al. (2016) was also captured and published by Lazaridis et al. (2016). Additionally, the fact that the capture kits are not publicly available, made comparison more difficult by restricting the output of all of the captured genomes to one laboratory. This creates a good dataset to compare the output of shotgun sequencing with the output of capture. In an ideal world, both datasets should be equivalent. However, this assumption has proven to be difficult to achieve for a number of reasons. Amongst them, the differences in the biochemistry of shotgun sequencing and DNA capture, the differences in the UDG treatment vs using an *in-silico* damage-removing approach, and the different sequencing pre-processing, mapping, and SNP calling steps.

Here, I analyse the reasons that account for most differences between platforms, and provide easy solutions on how to pre-process data from different platforms to make it compatible. I use the fact that we have versions of genomes that have been both shotgun sequenced (NE1 and BR2 from Gamba et al., 2014; and GD13a from Gallego-Llorente et al., 2016), and captured (NE1 and BR2 in Mathieson et al., 2016; and GD13a n Lazaridis et al., 2016), and I perform *D* statistics using two versions of the same genomes, which in ideal circumstances should be 0. An absolute *Z* value below 3 indicates non-statistically-significant differences between both samples.

Results

First, I compared the shotgun-sequenced genomes, as sequenced and processed by Gamba et al (2014), with the capture-sequenced individuals, as sequenced, processed and genotyped by Lazaridis et al (2016) and Mathieson et al (2015). I used *D* statistics of the form *D*(Mbuti, X; Y_C, Y_S) to compare the individual X, either shotgun-sequenced or captured; to individual Y, either shotgun-sequenced (Y_S) and captured (Y_C). In an ideal setting, such D should have a value of 0, as genomes Y_C and Y_S derive from the same individual. As I show below, this is not always the case.

Table 1 illustrates how samples that were shotgun sequenced showed an attraction towards the shotgun version of Y, whereas samples that were captured showed an affinity towards the captured version of Y. This was especially apparent with shotgun-sequenced samples, where the shotgun version of BR2 showed a Z-score of 7.791 towards the shotgun version of NE1, as opposed to the capture version of NE1. Other shotgun-sequenced samples showed similar results (Figure 3).

Capture samples, such as $BR2_C$, showed negative values, suggesting a similar (although non-significant) affinity towards the captured sample.

D statistic	All			Transversions only		
	D	Z	SNPs	D	Z	SNPs
D (Mbuti, $BR2_C$; $NE1_C$, $NE1_S$)	-0.0087	-1.622	537357	-0.0051	-0.419	105962
D (Mbuti, **$BR2_S$**; $NE1_C$, **$NE1_S$**)	**0.0363**	**7.791**	**702805**	**0.0349**	**3.416**	**140370**
D (Mbuti, INe_C; $NE1_C$, $NE1_S$)	-0.0084	-1.699	596816	-0.0149	-1.415	123618
D (Mbuti, **INe_S**; $NE1_C$, **$NE1_S$**)	**0.0391**	**6.418**	**406906**	0.0225	1.723	86226
D (Mbuti, IrLN; INe_C, INe_S)	-0.0139	-2.267	275825	-0.0159	-1.097	56410
D (Mbuti, $NE1_C$; INe_C, INe_S)	-0.0072	-1.317	370951	-0.0066	-0.519	75313
D (Mbuti, **$NE1_S$**; INe_C, **INe_S**)	**0.0235**	**5.044**	**544727**	0.0112	1.086	113358
D (Mbuti, $BR2_C$; INe_C, INe_S)	-0.0057	-0.989	356306	-0.0137	-1.039	72539
D (Mbuti, **$BR2_S$**; INe_C, **INe_S**)	**0.0211**	**4.437**	**544793**	0.0134	1.251	113443

Table 1. D statistics comparing attractions between capture and shotgun-sequenced genomes. Here, capture genomes have been taken directly from the dataset published in Lazaridis et al., (2016) and Mathieson et al., (2016).

These biases could be due to either platform (i.e. chemistry) or processing (i.e. bioinformatics) differences. In order to have a full picture of the biases exclusively due to platform differences, I decided to re-process the captured genomes ($NE1_C$, $BR2_C$, INe_C and $IrLN_C$) following the same guidelines published in Gamba et al. (2014), which included using only sequence data with a base quality ≥ 20 and depth ≥ 15. These re-processed captured genomes were used for the shotgun samples used in this analysis.

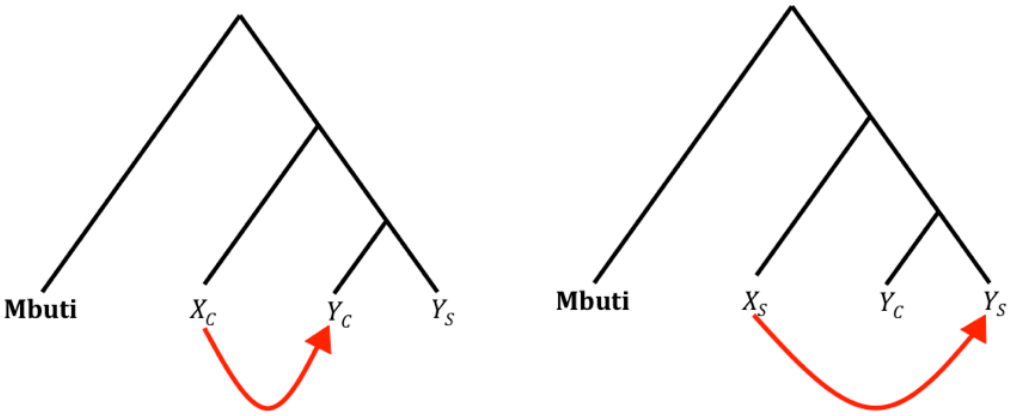


Figure 3: Shows that in ancient samples sequenced by either capture or shotgun sequencing, there seems to be an attraction towards other samples sequenced by the same method. Here, I do D statistics of the form D(Mbuti, X; Y_C, Y_S), where Y_C is a sample obtained by SNP capture, and Y_S is the same individual sequenced by shotgun sequencing. X, depending on the sequencing method, shows more affinity towards Y_C or Y_S respectively.

Table 2, however, shows that even when the biases due to processing are removed by using an identical bioninformatics pipeline for all samples, samples obtained on the same platform still show a degree of affinity. It is noticeable that while using all SNPs did not change the pattern previously observed, using only transversions resulted in the reduction of large Z-scores, such as the one from D(Mbuti, **$BR2_S$**; $NE1_C$, **$NE1_S$**) going from 3.416 to 2.229. These results conclusively show that going back to the raw data, and re-pre-processing reads and re-mapping reads using the same protocol is a key step when using datasets from different laboratories and publications, in order to avoid biases that would skew results towards an artefactual pattern created by differences in the data processing, rather than a real demographic pattern. However, this step was only effective when coupled with restricting the analysis to only transversion calls, which are not affected by deamination. This method has been used in publications such as Fu et al., (2015) and Gallego-Llorente et al., (2016).

D statistic	All (1,053,016 SNPs)			Transversions only (208,788 SNPs)		
	D	Z	SNPs	D	Z	SNPs
D (Mbuti, BR2$_C$; NE1$_C$, NE1$_S$)	-0.0033	-0.620	493259	-0.0132	-1.026	92668
D (Mbuti, **BR2$_S$**; NE1$_C$, **NE1$_S$**)	**0.0364**	**7.966**	646641	0.0238	2.229	122872
D (Mbuti, INe$_C$; NE1$_C$, NE1$_S$)	0.0064	1.248	557281	-0.0159	-1.345	105435
D (Mbuti, **INe$_S$**; NE1$_C$, **NE1$_S$**)	**0.0374**	**5.923**	390106	0.0118	0.836	75625
D (Mbuti, IrLN; INe$_C$, INe$_S$)	-0.0080	-1.094	245100	-0.0265	-1.445	47607
D (Mbuti, NE1$_C$; INe$_C$, INe$_S$)	-0.0071	-1.125	333750	-0.0116	-0.728	64476
D (Mbuti, **NE1$_S$**; INe$_C$, **INe$_S$**)	**0.0185**	**3.494**	485420	0.0061	0.491	95711
D (Mbuti, BR2$_C$; INe$_C$, INe$_S$)	-0.0019	-0.284	320561	-0.0089	-0.559	62077
D (Mbuti, **BR2$_S$**; INe$_C$, **INe$_S$**)	**0.0175**	**3.381**	485461	0.0082	0.648	95770

Table 2. D statistics comparing attractions between capture and shotgun-sequenced genomes. For this analysis, capture genomes have been processed from the raw data files using the same protocol as shotgun genomes.

Table 2 therefore shows that some of the results in the raw data are probably driven by damage. As we eliminate patterns arising from different processing and mapping options, the remaining factors that may affect the observed results in the D statistics are those driven by the biochemical basis of the sequencing, the degree of DNA endogenous damage, the use of UDG treatment, or exogenous contamination.

It was noticeable that while the absolute values of the Z scores went down in value, so did the D scores, but the relationship was not linear. I then decided to test the same D statistics using a random subset of SNPs (Table 3). This subset panel had the same number as SNPs as found in the tranversion-only dataset. Strikingly, by using a random subset of 208,788 SNPs I also removed some (though not all) of the patterns of the Z scores. When the D(Mbuti, BR2$_S$; NE1$_C$, NE1$_S$) was calculated on all SNPs, the resulting Z score was 7.966, but when it was calculated using only transversions, the resulting Z score was 2.229; when it was calculated using the same number of SNPs as the transversions, but randomly subsetted, the Z score was 4.116. In fact, when we lower the number of SNPs we are to

some extent diluting the power of the D-statistics, leading to non-significant results in comparisons that would have been significant with a higher number of SNPs. This shows that reducing the number SNPs covers up the platform bias pattern to some extent, especially when differences are small. However, this also shows that only selecting the transversion SNPs is still helpful for analyzing data without deamination biases.

D statistic	**Random subset (208,788 SNPs)**		
	D	**Z**	**SNPs**
D (Mbuti, $BR2_C$; $NE1_C$, $NE1_S$)	-0.0134	-1.129	97940
D (Mbuti, **$BR2_S$**; $NE1_C$, **$NE1_S$**)	**0.0428**	**4.116**	128111
D (Mbuti, INe_C; $NE1_C$, $NE1_S$)	-0.0083	-0.698	110430
D (Mbuti, **INe_S**; $NE1_C$, **$NE1_S$**)	0.0325	2.266	77362
D (Mbuti, IrLN; INe_C, INe_S)	-0.0271	-1.610	48649
D (Mbuti, $NE1_C$; INe_C, INe_S)	-0.0339	-2.291	66167
D (Mbuti, **$NE1_S$**; INe_C, **INe_S**)	0.0032	0.260	96242
D (Mbuti, $BR2_C$; INe_C, INe_S)	-0.0177	-1.159	63679
D (Mbuti, **$BR2_S$**; INe_C, **INe_S**)	-0.0025	-0.210	96239

Table 3 (previous page). D statistics showing the same analysis as table 2, but using a random subset of SNPs of the same number as the subset including only transversions.

As reprocessing the data did not completely correct all platform biases when using all SNPs (Table 2), I decided to perform soft-clipping on the shotgun-sequenced samples as non-UDG–treated data has, overall, more damage. It is a known pattern that the mismatch frequency exponentially increases at the ends of reads: C to T misincorporations are very common near the 5' ends of reads, and G to A misincorporations are reciprocally common near the 3' ends of reads (Brotheron et al., 2007, Briggs et al., 2007). As the shotgun genomes were not UDG-treated, I decided to subject the mapped reads to a 4-base, 7-base or 10-base trimming at both ends (termed soft-clipping), to assess if the platform comparison would benefit from completely eliminating these bases at the ends of reads from the analysis.

Table 4 shows how soft-clipping deals reduces the discrepancy between shotgun and capture data. Clipping 7 and 10 base pairs at each end almost fully removed the bias in the D statistics, which is shown in the tests D(Mbuti, $BR2_S$; $NE1_C$, $NE1_S$) and D(Mbuti, INe_S; $NE1_C$, $NE1_S$). Here we can clearly see that Z values of 7.966 reduce to 4.966 when 7bp soft-clipping on shotgun samples is performed, and 3.762 when reads are soft-clipped by 10bp. These effects are not due to a reduction in the number of SNPs, as soft-clipping will not reduce much the number of bases called in samples with coverage above 1X.

These results confirm that a big part of the platform bias is due to how different pipielines deal with damage. In non-UDG-treated shotgun-sequenced samples, damage is concentrated at the ends of reads, hence requiring a large degree of soft-clipping to deal with this damage. Whilst soft-clipping helped somewhat, it never fully removed the biases. For sake of completeness, I have added the values from the transversions-only analysis, but as we saw in Tables 2 and 3, the reduction in Z-score values is mostly due to a lower number of SNPs, rather than actually solving the artefact created by different platforms.

D statistic	All (1,053,016 SNPs)			Transversions only (208,788 SNPs)		
	D	Z	SNPs	D	Z	SNPs
4bp soft-clipping						
D (Mbuti, BR2$_C$; NE1$_C$, NE1$_S$)	-0.0038	-0.703	493166	-0.0202	-1.584	92602
D (Mbuti, BR2$_S$; NE1$_C$, NE1$_S$)	0.0297	**6.314**	646437	0.0042	0.397	122731
D (Mbuti, INe$_C$; NE1$_C$, NE1$_S$)	0.0044	0.873	557187	-0.0193	-1.678	105366
D (Mbuti, INe$_S$; NE1$_C$, NE1$_S$)	0.0268	**4.181**	390035	-0.0080	-0.564	75572
D (Mbuti, IrLN; INe$_C$, INe$_S$)	-0.0143	-1.924	245100	-0.0331	-1.817	47607
D (Mbuti, NE1$_C$; INe$_C$, INe$_S$)	-0.0074	-1.170	333750	-0.0145	-0.906	64476
D (Mbuti, NE1$_S$; INe$_C$, INe$_S$)	0.0136	2.541	485331	-0.0057	-0.452	95650
D (Mbuti, BR2$_C$; INe$_C$, INe$_S$)	-0.0016	-0.232	320561	-0.0093	-0.579	62077
D (Mbuti, BR2$_S$; INe$_C$, INe$_S$)	0.0156	**3.013**	485399	0.0001	0.009	95733
7bp soft-clipping						
D (Mbuti, BR2$_C$; NE1$_C$, NE1$_S$)	-0.0047	-0.891	493095	-0.0237	-1.857	92555
D (Mbuti, BR2$_S$; NE1$_C$, NE1$_S$)	0.0223	**4.966**	646232	-0.0013	-0.116	122596
D (Mbuti, INe$_C$; NE1$_C$, NE1$_S$)	0.0040	0.807	557104	-0.0252	-2.200	105311
D (Mbuti, INe$_S$; NE1$_C$, NE1$_S$)	0.0298	**4.426**	358042	0.0006	0.040	69438
D (Mbuti, IrLN; INe$_C$, INe$_S$)	-0.0100	-1.270	225642	-0.0354	-1.864	43843
D (Mbuti, NE1$_C$; INe$_C$, INe$_S$)	-0.0079	-1.183	306299	-0.0151	-0.907	59242
D (Mbuti, NE1$_S$; INe$_C$, INe$_S$)	0.0146	2.521	448173	0.0064	0.498	88470
D (Mbuti, BR2$_C$; INe$_C$, INe$_S$)	0.0022	0.322	294093	-0.0004	-0.024	57029
D (Mbuti, BR2$_S$; INe$_C$, INe$_S$)	0.0119	2.178	448227	-0.0043	-0.310	88542
10bp soft-clipping						
D (Mbuti, BR2$_C$; NE1$_C$, NE1$_S$)	-0.0073	-1.349	493004	-0.0261	-2.134	92498
D (Mbuti, BR2$_S$; NE1$_C$, NE1$_S$)	0.0170	**3.762**	645969	-0.0096	-0.931	122429
D (Mbuti, INe$_C$; NE1$_C$, NE1$_S$)	0.0014	0.270	556987	-0.0232	-2.059	105234
D (Mbuti, INe$_S$; NE1$_C$, NE1$_S$)	0.0255	**3.615**	320515	-0.0059	-0.376	62066
D (Mbuti, IrLN; INe$_C$, INe$_S$)	-0.0095	-1.154	202563	-0.0262	-1.249	39297
D (Mbuti, NE1$_C$; INe$_C$, INe$_S$)	-0.0081	-1.138	273991	-0.0009	-0.049	52945
D (Mbuti, NE1$_S$; INe$_C$, INe$_S$)	0.0126	2.115	404051	0.0037	0.263	79716
D (Mbuti, BR2$_C$; INe$_C$, INe$_S$)	-0.0039	-0.531	263165	-0.0062	-0.350	51018
D (Mbuti, BR2$_S$; INe$_C$, INe$_S$)	0.0115	2.010	404108	0.0006	0.042	79784

Table 4. Soft-clipping capture genomes. D statistics showing the same analysis as in tables 2 and 3, but introducing a post-mapping trimming of all reads of 4, 7 or 10 base pairs in every genome.

Whilst the approaches discussed so far greatly reduce platform bias, none of them fully removes platform bias (with the exception of focussing on transversions, but that is in

large part a result of losing power due to the lower number of SNPs). There is one last mechanism I have not considered yet: could the platform bias also arise from differential reference bias when mapping reads to the reference genome? I used a version of hg19 where the Lazaridis SNPs were changed to a third allele (not the reference nor the alternate allele), to allow for an even mapping of reads, irrespective of whether they carry the reference or alternate allele.

Table 5 shows that, in fact, when removing the reference bias at the mapping stage of the analysis, I also remove the attraction of shotgun-sequenced genomes with other shotgun-sequenced genomes, as well as the attraction between captured samples. It is worth noting that even without soft-clipping, this approach eliminated platform bias completely. This could be explained by a situation in which most of the bias that causes the attraction between platforms is actually introduced in the alignment step, and that, in fact, the way that we normally deal with damage by discarding damaged reads, taking reads longer than 30bp, trimming the ends pre-processing, and the mapping and base quality filtering actually are quite successful. A caveat would be that the SNP numbers are quite low (around 35-40% of SNPs are lost), due to a combination of the third-base mapping (which removes a percentage of the reads) and a slightly lower coverage of the samples (NE1 and BR2 were subsetted to 1.86x and 0.82x respectively for computational speed). A further check will be needed, in order to run the same analysis at the full coverage and check whether the effect of losing the platform attraction bias is real, and not an artefact of the lower number of SNPs. ANGSD calling (Korneliussen et al., 2014) without reference was not attempted, as here the analysis focused on removing the reference bias at the mapping stages of the analysis. This will be included in further analysis as part of the paper about this study that is currently in preparation, however it is worth mentioning that for low call rates, SAMtools outperforms ANGSD, as reported by Korneliussen et al. (2014).

	No soft-clipping			4bp soft-clipping			7bp soft-clipping			10bp soft-clipping		
	D	**Z**	**SNPs**	**D**	**Z**	**SNPs**	**D**	**Z**	**SNPs**	**D**	**Z**	**SNPs**
D (Mbuti, $BR2_C$; $NE1_C$, $NE1_S$)	-0.0069	-1.041	303584	-0.0091	-1.221	267796	-0.0119	-1.584	239993	-0.0158	-2.055	226726
D (Mbuti, $BR2_S$; $NE1_C$, $NE1_S$)	0.0130	1.714	232270	0.0138	1.660	201415	0.0183	2.114	177498	0.0047	0.521	160475
D (Mbuti, INe_C; $NE1_C$, $NE1_S$)	-0.0142	-2.257	345723	-0.0164	-2.339	299666	-0.0080	-1.073	263378	-0.0110	-1.448	249140
D (Mbuti, INe_S; $NE1_C$, $NE1_S$)	-0.0075	-0.928	220572	-0.0082	-0.919	177344	-0.0101	-1.013	145330	-0.0113	-1.018	117928
D (Mbuti, IrLN; INe_C, INe_S)	-0.0041	-0.491	210616	-0.0015	-0.164	175942	-0.0046	-0.469	149308	-0.0106	-1.035	141333
D (Mbuti, $NE1_C$; INe_C, INe_S)	-0.0063	-0.751	221335	-0.0175	-1.967	177565	-0.0147	-1.470	145384	-0.0171	-1.566	123664
D (Mbuti, $NE1_S$; INe_C, INe_S)	0.0069	0.938	283529	-0.0030	-0.390	226781	0.0038	0.411	184215	0.0021	0.214	151107
D (Mbuti, $BR2_C$; INe_C, INe_S)	-0.0120	-1.431	211578	-0.0093	-1.019	169307	0.0032	0.302	138343	0.0010	0.084	117780
D (Mbuti, $BR2_S$; INe_C, INe_S)	0.0144	1.718	211578	0.0091	0.961	155805	0.0274	2.545	126166	0.0221	1.883	103745

Table 5. Third-base mapping all ancient genomes. D statistics showing the same analysis as in tables 2, 3 and 4, but after having re-processed, realigned, and re-mapped all genomes to a version of the human reference genome where the Lazaridies SNPs have been substituted by a 3rd base, to remove the reference bias. I have also introduced a post-mapping trimming of all reads of 4, 7 or 10 base pairs in every genome. Here I looked at all SNPs, both transitions and transversions.

Discussion

Here I have attempted to tackle a very common problem with ancient DNA analysis, the integration of ancient data obtained by different people in different laboratories.

While in the past few years, publications such as Lazaridis et al (2014), Lazaridis et al (2016), Mathieson et al (2016), and Fu et al (2016) have made SNP datasets of ancient and modern populations available in eigenstratgeno format (admixtools) after being processed, mapped and called, it was incorrectly assumed that they could be merged with other ancient genomes, irrespective of how those other ancient genomes had been sequenced, pre-processed, mapped and called. Here, I have shown that this approach is deeply flawed, and can lead to data incompatibility problems, including platform and processing biases.

By using softclipping, or removing transitions, reference bias is not removed. However, reference bias is constant amongst coherent datasets with similar amounts of damage. Hence, by equalising the amount of damage by softclipping the ends and removing the platform and processing bias as much as possible, if reference bias was constant irrespective of the sequencing platform, then reference bias would not skew analysis.

In this way, the main goal of this chapter has been to find a way to continue using the most common tools and pipelines for analysing ancient data, but finding a way in which different sources of data, or data from different laboratories do not affect demographic results and fine demographic patterns.

Data comes with damage and biases. However, it has been a pattern in most genetic analysis using ancient genomes to keep as much as data as possible, under the assumption that more data is automatically better resolution. However, here I show that removing biases is always a priority even if a percentage of data is lost. As an example, using only transversions loses 4/5ths of the data, but this has proven as the best way to remove biases caused by deamination patterns in reads. However, here I show that care is needed when using transversions only, because heavily reducing the number of SNPs will mean

that some of the formal statistical tools might lose their power and result in non-significant results, when a higher number of SNPs would have made them significant.

Additionally, the experiments shown here suggests that softclipping of 10 bases at the end of each read might be worth to decrease biases caused by damage. In a sample containing reads that average 60bp (which is typical of ancient DNA), by soft-clipping 10 bases at each end (20 bases in total), we would keep 2/3rds of the data: a much higher percentage than using transversions only. Additionally, this does not result in eliminating SNP positions as selecting only transversions does, as in any 2-3X sample, most SNPs will have enough copies to be called. So, in fact, we will be using close to 90-95% of SNPs, just with a smaller number of copies each – but with a higher certainty that the variant is the correct one. It is worth noting, however, that soft-clipping will not be readily achievable with heavily degraded samples, as the majority of reads will be short and we would lose most of the data. However, with an extremely degraded sample, looking at transversions only would also be challenging.

We could therefore conclude that "*bad data is not data*". In my analysis of Ganj Dareh (Gallego-Llorente et al., 2016) 60% of data was eliminated, as the analysis was done using only transversions. Doing the analysis on all SNPs could have created spurious artefacts and attractions could have been introduced by the transitions data, heavily affected by deamination patterns.

Finally, it is worth noting that highly damaged samples of low coverage are inherently difficult to work with, even without UDG treatment: the low coverage will mean that softclipping would in fact potentially remove large numbers of SNPs, while a high number of short reads would be automatically discarded, due to damage patterns at the ends of reads. If to that, we add reference bias, we end up with an extremely challenging sample. UDG treatment would improve some aspects in terms of damage removal, but would fragment a high number of the scarce long reads. Hence, for samples with a high amount of predicted damage, it might be wise to wait until sequencing technology improves before attempting to use any valuable part of a bone in an experiment that might not yield the desired results.

Methods

Sources of data for f_3 statistics (Fig. 1):

The datasets from Lazaridis et al. (2016) were obtained from the following link:

http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets_files/NearEastPublic.tar.gz

It contains the AncientLazaridis2016 dataset (294 ancient individuals from Europe and the Near East), and the HumanOriginsPublic2068 dataset, with 2068 modern individuals from around the world.

I used the Iranian Farmer individual which was published in Gallego-Llorente et al. (2016), and can be downloaded from:

ftp://ftp.sra.ebi.ac.uk/vol1/ERA669/ERA669868/bam/Iranian_farmer1_sort_merge_rmdup_q30_l30_IR_2bpsoftclip.bam.

I used the Wezmeh Cave individual from Broushaki et al. (2016), which was downloaded from:

ftp://ftp.sra.ebi.ac.uk/vol1/ERA637/ERA637051/bam/WC1.all_SG_join.Mkdup.len.realg.bam.

SNP calling and processing for f_3 statistics (Fig. 1)

SNPs from bam files from the Iranian Farmer and Wezmeh were called using the following SAMtools (Li et al., 2009) command:

samtools mpileup -C50 -t DP4 -q 30 -Q 20 -uIf FASTA -l BED BAM | bcftools call -cA -Ov -> OUT

Where FASTA refers to the hg19 fasta file, BED to a file covering the 1,233,553 positions in the AncientLazaridis2016 dataset, and BAM to the previously-downloaded bam files.

I used an in-house-developed script to obtain the majority call from the vcf file, and in case of equal number of reads, one at random.

I used PLINK (Purcell et al., 2007) to merge the ancient genome calls with the human reference genome hg19, after which we converted it to eigenstratgeno format (using convertf from admixtools) (Patterson et al., 2012).

I then used the polarise option of convertf to make sure both the ancient dataset and the new modern genome are similarly polarised to hg19. I used the mergeit tool (admixtools) (Patterson et al., 2012), to merge both datasets. Finally, I used plink to merge the WC1-IranianFarmer-AncientLazaridis2016 with the HumanOriginsPublic2068.

In order to do the comparison between the 1,233,553 positions in the 1240k capture dataset, and the 354,212 positions in the Haak ancient and modern dataset(Haak et al., 2015), I used admixtools to filter out those SNPs present in Haak but not in the 1240k capture dataset. I then merged both datasets using plink (Purcell et al., 2007). I additionally created random subsets of 300k, 60k, 40k and 30k SNPs from the bigger dataset, and also merged with plink. F3 statistics were done using qp3pop from admixtools (Patterson et al., 2012).

Sources of data for pairwise f_3 statistics (Fig. 2)

The dataset from Fu et al. (2016) was downloaded from http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets_files/FuQ.zip.

I also downloaded the whole genome Yoruba (HGDP00927) from http://cdna.eva.mpg.de/denisova/BAM/human/HGDP00927.bam.

The 2.2M SNPs in the Fu2.2M dataset were called in the HGDP Yoruba genome using the following command:

samtools mpileup -C50 -t DP4 -q 30 -Q 20 -uIf FASTA -l BED BAM | bcftools call -cA -Ov -> OUT

Where FASTA refers to the hg19 fasta file, BED to a file covering the 2.2M positions in the Fu2.2M dataset, and BAM to the previously-downloaded HGDP Yoruba file.

The HGDP00927 Yoruba was merged to the hg19 reference genome using plink, as well as the Fu2.2M dataset. Both datasets were then converted to admixtools format, polarised to hg19, and merged using mergeit. F3 statistics were done using qp3pop from admixtools (Patterson et al., 2012).

I downloaded the BR2 and NE1 genomes from Gamba et al. (2014) using ENA accession code PRJEB20905.

Sources of data for *D* statistics

The BR2 and NE1 genomes were individually merged to the hg19 reference genome using plink. Both datasets were then converted to admixtools format, polarised to hg19, and merged using mergeit. Finally, the BR2-NE1-hg19 dataset was merged to the WC1-IranianFarmer-AncientLazaridis2016-HumanOriginsPublic2068 as explained before. D-statistics were performed using qpDstat from admixtools (Patterson et al., 2012).

I downloaded the FASTQ files from BR2, NE1, Ganj Dareh and Iranian Late Neolithic from ENA. Their individual codes were ERR1136467, ERR1136469, ERR1463779, and ERR1463796 respectively.

Read pre-processing and mapping

Reads were processed as indicated in Gamba et al., (2014). Given the rate of misincorporation sites at the 3' of reads (G to A), and at 5' end of reads (C to T), I trimmed two bases at the ends, before mapping (MacHugh et al., 2000). This was done with trimfq (https://github.com/lh3/seqtk), using trimfq options -b 2 -e 2. Afterwards, I mapped reads to the GRCh37 build of the human nuclear genome (hg19) and the revised Cambridge reference sequence for the mitochrondrial genome (NCBI NC_012920.1). I used bwa, the Burrows Wheeler Aligner, version 0.7.12-r1039. (Li and Durbin, 2009). I ran the program with default parameters, except with the seed option disabled(-l 1000). Duplicate reads were then removed using samtools, version 0.1.19-44428cd (Li et al., 2009).

Data was then merged as previously described, using plink, and convertf and mergeit from admixtools.

Softclipping was done prior to SNP calling, using a home script.

FASTQ files from the shotgun whole genome sequencing of GD13a were obtained through EBI ENA, accession number PRJEB13189. The Adapter sequence was trimmed from reads, and the length of selected reads was set to a minimum of 34bp, using cutadapt, and the flags -m 34 -O 1.

Third base mapping was done by using a GRCh37 build of the human nuclear genome where the bases of the SNPs identified in the phase 3 of the 1,000 genomes project have been substituted to a base not being the alternate base identified by the 1,000 genomes project. A caveat was added, that if the reference base was a C, then it would not be replaced by a T. If the base was a G, it would not be replaced by an A. This was done to guard against possible damage patterns. The modified fasta file was created using the mutfa option in seqtk.

SNP calling and processing for *D* statistics

SNP calling, was, again, done using the following SAMtools command:

samtools mpileup -C50 -t DP4 -q 30 -Q 20 -uIf FASTA -l BED BAM | bcftools call -cA -Ov -> OUT

I used plink (Purcell et al., 2007) to merge the ancient genome calls with the human reference genome hg19, after which I converted it to eigenstratgeno format (using convertf from admixtools) (Patterson et al., 2012).

I then used the polarise option of convertf to make sure both the ancient dataset and the new modern genome are similarly polarised to hg19. I used the mergeit tool (admixtools) (Patterson et al., 2012), to merge all datasets.

I only used the demographic SNPs from the HOIll.snp file (Lazaridis et al., 2016).

D statistics were performed using qpDstat from admixtools (Patterson et al., 2012).

5. General Discussion

This PhD thesis shows that past human movements and migrations can be studied using genetic analysis involving ancient and modern genomes, which can lead to close reconstructions of ancient demographic events. When this PhD was started back in 2014, there were very few ancient genomes, and most of them were sourced from Europe, Northern Eurasia and North America (Keller et al., 2012; Rasmussen et al., 2010a; Skoglund et al., 2012), whose analysis only resulted in a very general picture of some of the largest demographic shifts in the last 10,000 years in the Northern Hemisphere. However, there were no samples from Africa, no samples from the Middle East, and very little was understood about human movements outside Europe, Siberia or North America.

This thesis, which contains material from two publications (chapters 2 and 3), alongside other publications from various other authors, has meant a giant leap for our understanding of the demographic aspects of human prehistory: I have presented the first ancient genome from Africa (Mota in Chapter 3 of this thesis), which then was joined by other ancient African genomes from Schlebusch et al. (2017) and Skoglund et al., (2017). Hundreds of genomes from Europe have been made available, from every epoch in the last 15k years (Fu et al., 2016; Martiniano et al., 2017). And, perhaps most importantly, I have presented here the first ancient genome from the Middle East (Ganj Dareh in Chapter 2 of this thesis), which was simultaneously published alongside Broushaki et al. and Lazaridis et al., in 2016). East Asia (Siska et al., 2017) and the Americas (Raghavan et al., 2015) have in addition received quite a lot of attention lately. Therefore, in the last 4 years, this field has changed immensely, in a way that would have been unthinkable a few years ago. The material in this thesis forms an important part of this change.

From the start of my PhD, I have benefited from the publicly available Human Origin dataset of modern populations, as well as other datasets with modern populations specific to other areas of the world, such as Pagani et al., (2012), and Pickrell et al. (2014). My approach to analysing new samples followed the descriptive approaches used by Gamba et al. (2014) and Jones et al. (2015), in their whole genome sequencing approach to understanding European prehistory and the Caucasus Hunter-Gatherers, respectively. My analysis of Ganj Dareh and Mota followed this approach, which was useful for

understanding the genetic landscape of those areas of the world and the demographic processes that have happened in East Africa, the Near East and Central in the last few millennia.

How do human cultures expand?

In this thesis, questions about the Neolithic cultural expansion and its associated human movements get asked constantly, while other related questions are also readily mentioned: how did the Neolithic revolution affect the genetic landscape of the Near East? In which ways did the demographic landscape of Europe change, before, during and after the expansion of the Neolithic? Was this expansion a unique event in terms of cultural shift, speed, permeability and reach, or were there other similar migratory events in prehistory? And hence, were the population movements associated with the **development** of the Neolithic package, similar to the ones associated with its **expansion**? How did the development of the Neolithic affect Africa and other regions of Asia? And finally, have these migrations left any signature in terms of the genetic and linguistic landscape of today's human diversity?

My results show that the eastern fringes of the Fertile Crescent (modern-day western Iran) was inhabited by populations mostly similar to Hunter-gatherer populations from the Caucasus, but remarkably, very distinct from the Anatolian farmers who spread the Neolithic package into Europe. While a degree of cultural diffusion between Anatolia, Mesopotamia and the Zagros highlands likely happened, the notable genetic dissimilarity between individuals of both areas supports a model in which the Neolithic societies of the Near East originated from very distinct lineages. However, it was then reported by Broushaki et al. (2016), that while Early Neolithic samples from the eastern side and western side of the Fertile Crescent show high levels of genetic difference, more recent samples (such as Iron Age Iranians or Chalcolithic Anatolians) indicate that a subsequent process of post-Neolithic homogenization likely happened throughout the region. This subsequent process therefore very likely involved tri-directional gene flow between the Mesopotamian lowlands, the Anatolian highlands and the Iranian plateau – in addition to even posterior influences from the Steppes north of the Caucasus, likely brought by the first Indo-European speakers into this region.

The Neolithic transition in South and South-Western Asia included a wide range of different crops and domestic animals in different parts of the wider region, but with no archaeologically obvious unique geographical focus (Fuller et al., 2012). This archaeological view is now fully consistent with what we see in the genetics: a multi-polar development of agriculture and farming, only homogenised later by subsequent, probably unrelated processes.

However, it remains very surprising that Ganj Dareh and other of the Early Neolithic genomes from the Zagros are diverged 40-50k years ago from those from North-western Anatolia and the typically European Neolithic component. This is primarily surprising due to the fact that there is archaeological and genetic evidence of east-to-west movements of domestic cattle across Anatolia during the Neolithic (Scheu et al., 2015). There is the chance that not only prior to the Neolithic revolution, but also at the start of it, there was limited demic movement across stark boundaries between biogeographically differentiated areas, which would explain the little gene-flow between the Mesopotamian river valleys, the Anatolian highlands and the Iranian plateau. This would have allowed for cultural expansion events without an associated demic homogenisation, at least during the first few millennia of the Neolithic revolution. This view would also explain why the Early Neolithic Iranians such as Ganj Dareh cluster with the Ancient Northern Indian component of South Asian populations, as the Iranian plateau progressively merges with the Balochistan ranges to end in the Indus River Valley – incidentally the first centre of Agriculture outside the Near East.

I also explored the effect that the development of agriculture in the Near East had in movements of people in other areas of the Old World, where this development was later spread. It was already known that the Neolithic arrived into Europe via Anatolia, and it was reported that the strong Neolithic genetic component in European populations came from the Near East (Peltenburg et al., 2000). Here, I have showed how this component also reached East Africa in a really strong manner, by being the principal component of the backflow of communities of Western Eurasian origin into Africa around 4,000 years ago. These populations, closely related to Anatolian Farmers, were identified by using Mota, an ancient genome from a male from the Ethiopian highlands, in comparison with modern populations from East Africa, which show the genetic signatures of this backflow.

The importance of standardising methods and approaches in aDNA

Finally, I have also studied the origin of some of the common problems in the analysis and usage of ancient DNA, such as merging capture datasets with diverse number of ascertained SNPs, combining capture and shotgun data in the same analysis, and the effect of UDG treatment in ancient samples. I have shown that using a reduced dataset of common SNPs with the same ascertainment is a key step in order to analyse individuals from different sources and sequenced at different resolutions.

Additionally, in this thesis I have re-emphasised the importance of a common pipeline for pre-processing and mapping reads, and calling SNPs. Only by using a common pipeline will we be able to overcome problems of data compatibility from different platforms. In addition, given the structures of damage in ancient DNA, and given the drawbacks of UDG treatment in samples to be shotgun-sequenced, I have shown that there are a series of steps that one should take in order to make shotgun data fully compatible with capture data, such as the soft-clipping of reads, and the alignment to a reference genome where the bases corresponding to the SNPs being analysed have been substituted by an uncommon third-base. Third-base mapping also solves the common problem of reference bias. With this, I propose a how-to guide on how to work with ancient DNA to avoid data compatibility problems, which will undoubtedly be incredibly useful in the future.

This, however, leads us to the question of what is better: whole genome sequencing by shotgun, or SNP capture by in-solution hybridization. SNP capture provides us with a cheap and easy way to obtain ready-to-use data to study demographic processes such as migrations, admixture events, and recent cultural shifts associated with population movements. In addition, it allows the study of many, relatively undamaged modern samples of which we have many fossils of the same area, epoch and culture.

SNP capture, however, prevents us from using the entire genome for extra, very informative analysis, such as Runs of Homozygosity to look at past population bottlenecks, dating split times (e.g. using GPhoCS), calculating percentages of archaic human ancestry using the entire genomes, or looking at the dynamics of closely-related ancient populations. Hence, it can be argued that for scarce, valuable samples, a shotgun-sequencing approach should be always preferred.

It is a fact that obtaining ancient DNA entails the destruction of portions of very valuable and irreplaceable sample. We could also argue that as whole genome sequencing technology keeps improving, we will be able to fully sequence at higher quality samples that today would be challenging to study. Hence, it seems reasonable that there are samples for which we should wait before attempting any uncertain aDNA extraction, at least before the physical skulls are fully documented, morphologically studied, and digitalised. An example of these would be the Neanderthal fossils from Atapuerca, in Spain. Care must be taken in order to decide which samples to sequence in the present and in which fashion, and which samples for which we should instead delay DNA extraction.

Open questions and further research avenues

The Indo-European question

One of the biggest questions in the prehistory of the Eurasian Bronze Age has been to locate the homeland of all surviving branches of the Indo-European language family. The majority of Indo-European specialists, starting with Marija Gimbutas in the 1950s, supported the Kurgan hypothesis, which linked the PIE homeland to the Pontic-Caspian steppes, between the Caspian and the Black Seas, around 6,000 years ago (Gimbutas, 2001; Gimbutas and Dexter, 1997). There was a major alternative, the Anatolian hypothesis, which defended an earlier homeland in the Anatolian peninsula, around 10,000 years ago (Renfrew, 1987).

The Kurgan hypothesis was defined as a way to group various cultures that appeared 6,000 years ago in the Pontic-Caspian Steppes, which included the Yamna, the Samara and the Seroglazovo cultures of the region. These peoples were pastoralist and nomadic, and at some point 5,000 years ago, expanded into Eastern Europe. Haak et al, in 2015, using ancient genomes obtained from throughout Europe and sequenced by capture, linked the Kurgan model and the Bronze age expansions into Europe to the spread of the Indo-European languages and hence supported the Kurgan theory regarding the late Indo-European homeland (Haak et al., 2015). The recent publication of the Iberian genomes from the Bronze Age further supported this theory (Martiniano et al., 2017).

However, we are still waiting for ancient genomes from South Asia at the time of the Indo-European expansions, which should be incredibly informative. Their analysis will set the tone for what will be the next question that ancient genetics can answer: who were the inhabitants of the Indus Valley civilization, and how did agriculture appear in the region? How did the expansion of the Indo-Aryan languages and peoples into the subcontinent affect culture and population structure? Although it is favoured that the Indus Valley Civilization spoke a language likely related to Dravidian, it is not known how this civilization acquired agriculture: While some theories point to heavy trade and knowledge diffusion between the Near East and South Asia, other theories point to independent developments (Lockard, 2010; McPherson, 2009). Genomes from this area will particularly help elucidating these open questions. A mostly Dravidian component would point towards either an independent development of agriculture, or knowledge diffusion without any demic component. However, if genomes belonging Indus Valley Civilization present sizable components from either the Near East of Central Asia, this could point to earlier events of population movements towards South Asia and subsequent admixture.

Another pending ancient genome related to the Indo-European expansions would be the sequencing of Hittite or other Indo-European genomes from Bronze Age Anatolia. This will be a stepping-stone for the pinpointing of the Proto-Indo-European homeland, as opposed to the late Indo-European homeland. If genomes from Hittite societies have a sizeable Caucasus hunter-gatherer or the Iron Age steppe component, then this would greatly support the Kurgan theory proposed by Maria Gimbutas, through which proto-Indo-European languages developed in the Pontic steppes between the Black and the Caspian seas, to posteriorly expand westwards to Europe, southwards to Anatolia, and south-eastwards to Central Asia, the Iranian Plateau and the Indian subcontinent. However, if the Hittites were genetically similar to Neolithic Anatolians, this would support the Anatolian hypothesis developed by Colin Renfrew in 1987.

Open questions in Africa

As it happens when exploring prehistory, new sources answer old questions, while the same new sources open new questions. This is extremely patent in studies concerning Africa: modern genomes were initially used to show the genetic signature of the Bantu expansions, a movement of agriculturalist peoples, originating in western Africa (today's Nigeria) and expanded to central and south-eastern Africa, all the way to current day South Africa. This involved a replacement and displacement of the native population, most likely related to today's Khoi-San populations in Namibia. The genetic signatures left by this event are still very patent using contemporaneous genetic data (de Filippo et al., 2012; Li et al., 2014; Patin et al., 2017) and linguistic evidence (Vansina, 1995).

Ancient genomes, however, not only confirmed these old findings and added a bit more detail, but also opened a whole set of new questions: A 2017 pre-print (Schlebusch et al., 2017) and a 2017 conference abstract (Skoglund et al., 2017) featuring African ancient genome showed direct ancient genomic evidence supporting the question of the Bantu expansions and the identity of those who were replaced. It showed, for example, that populations close to current-day Khoi-San inhabited areas as north as Malawi and Tanzania before the Bantu expansions. However, these papers also showed hints that current-day western Africans might harbour ancestry from an ancient lineage that separated from the more modern human lineages earlier than any other currently known one, including the Khoi-San. More ancient genomes will surely be used to confirm these findings, while opening new windows to the even deeper past. It will be very interesting to further this research in the future, not only to better understand some details of African population prehistory, but also to understand more about African ancient population diversity, which in turn will be key to understanding the genetic aspect of Africa at the time of the first Out-of-Africa episodes, 50-70k years ago.

Human prehistory, however, has many blank spots in different parts of the world. So far, we only have an almost-full picture of the events that shaped Europe and western Eurasia in the last 10,000 years. However, we only broadly understand glimpses of the last 10,000 years in Africa, Asia and the Americas. What happened between 50,000 and 10,000 years ago remains a mystery. However, it makes sense to presume that, if during the last 10,000 years, the European demographic composition has been reshaped at least twice by different migration events, the previous 40,000 years have probably been no different.

Africa is another such example: It had been assumed that Africa remained more or less unchanged since humans left Africa, around 80,000-70,000 years ago. However, recent developments have shown that this could not be further from the truth: the Bantu examples and the possible western African old population split show how the continent might have been subject to constant internal movements and migrations that have reshaped the genetic composition of the continent. It would be tempting to link these migrations to climatic events or cultural developments.

Continuous development of techniques

Another aspect of working with ancient DNA is the physical samples and the mode in which DNA is extracted and sequenced. In 2015, David Reich and his group pioneered the large-scale use of in-solution hybridisation capture (or capture, for short) (Haak et al., 2015), to enrich next generation sequencing libraries for a target set of predetermined SNPs. This lowers the cost of sequencing ancient samples for ancient demographic analysis, which most of the time allows for the sequencing of many more individuals for the same study. However, at the same time, it does not produce a whole genome sequence that can be used for other analyses which need whole genomes or large tracts of DNA (such as GPhoCS or Runs of Homozygosity). At the same time as David Reich's group sequenced large numbers of individuals using capture (Fu et al., 2016; Lazaridis et al., 2016; Mathieson et al., 2015), other studies published whole genomes, produced by shotgun sequencing (Broushaki et al., 2016; Gamba et al., 2014; Siska et al., 2017).

The amount of usable bone for DNA recovery has traditionally been very low (García-Garcerà et al., 2011; Sánchez-Quinto et al., 2012a; Skoglund et al., 2012), and so, a large amount of material was needed. In 2014, Gamba et al (Gamba et al., 2014) pioneered the extraction of ancient genetic material from the petrous section of the temporal bone, which yielded 4 to 16 times more DNA than teeth, and up to 183 times more DNA than other skeletal bones. However, ancient remains are very scarce, and every round of aDNA isolation destroys a portion of the sample. Hence, it is possible that in the process of acquiring samples for cheaper sequencing techniques (such as capture), we might be destroying invaluable data that could be used in the future for a much more informative whole genome sequencing, once this technique becomes cheaper and possibly better at yielding high coverage results.

Late Bronze Age and more recent samples are very common in Europe, and so losing a portion of such samples by using them for capture is relatively acceptable. However, for older samples, such as Palaeolithic samples, or samples from other regions of the world where aDNA conservation is more difficult, capture would be a technique that would favour a sneak peek today, rather than an eventual full picture tomorrow.

Future of the field and final words

On the whole, the field of ancient DNA has a lot of potential, as so far the number of samples outside Europe remains extremely poor. So far, samples have been used mostly in a descriptive fashion, through which we have been able to understand individual events of expansions, admixture, introgression, migration, and replacement. However, a larger number of samples will lead us to more integrated approaches for our understanding of demic and cultural migrational patterns throughout human history and their relationship with climate and ecology. In addition, studies linking demographics to adaptation are now starting to appear, which will shed light into how different populations have developed traits in order to increase chances of survival in different environments.

One of the biggest reasons for the study of aDNA has been to explore the relationship between the cultural-linguistic shifts in human prehistory and their associated migrations. This builds from an inherent willingness from people to understand their own ancestral origins, as well as to understand the origins of their socio-cultural group. Ancient DNA has benefited from these pretensions to the extent that a lot of the research that has been done so far has tried to elucidate the origins of certain cultural aspects in regions of the world, understand geographical patterns of language distribution, or pinpoint arrival dates of certain human groups into continents.

I would like to conclude this thesis, however, by emphasising a known fact, which is that humans have been moving all the time. A quirky example of this is how the remoteness of Polynesia was colonised, during a time when Europeans, with their supposedly more advanced technology, never dared to venture beyond the sight of land. This shows how

humans, using whichever technology is available, have always strived to find new places in search of resources, and, more generally, places in which to thrive.

As our history as a species advances, and as we step into the future, we must never forget that we have always moved in search of better (or just different) lives and that we have always wanted to explore in search of hope. And that, in one way or another, we will continue doing so in the future.

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109

Alkorta-Aranburu, G., Beall, C.M., Witonsky, D.B., Gebremedhin, A., Pritchard, J.K., Di Rienzo, A., 2012. The Genetic Architecture of Adaptations to High Altitude in Ethiopia. PLoS Genet. 8, e1003110. https://doi.org/10.1371/journal.pgen.1003110

Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P.D., Dąbrowski, P., Duffy, P.R., Ebel, A.V., Epimakhov, A., Frei, K., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolář, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Pálfi, G., Pokutta, D., Pospieszny, Ł., Price, T.D., Saag, L., Sablin, M., Shishlina, N., Smrčka, V., Soenov, V.I., Szeverényi, V., Tóth, G., Trifanova, S.V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando, L., Sicheritz-Pontén, T., Brunak, S., Nielsen, R., Kristiansen, K., Willerslev, E., 2015. Population genomics of Bronze Age Eurasia. Nature 522, 167–172. https://doi.org/10.1038/nature14507

Arsuaga, J.L., Martínez, I., Arnold, L.J., Aranburu, A., Gracia-Téllez, A., Sharp, W.D., Quam, R.M., Falguères, C., Pantoja-Pérez, A., Bischoff, J., Poza-Rey, E., Parés, J.M., Carretero, J.M., Demuro, M., Lorenzo, C., Sala, N., Martinón-Torres, M., García, N., Velasco, A.A. de, Cuenca-Bescós, G., Gómez-Olivencia, A., Moreno, D., Pablos, A., Shen, C.-C., Rodríguez, L., Ortega, A.I., García, R., Bonmatí, A., Castro, J.M.B. de, Carbonell, E., 2014. Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. Science 344, 1358–1363. https://doi.org/10.1126/science.1253958

Arsuaga, J.-L., Martínez, I., Gracia, A., Carretero, J.-M., Carbonell, E., 1993. Three new human skulls from the Sima de los Huesos Middle Pleistocene site in Sierra de Atapuerca, Spain. Nature 362, 534–537. https://doi.org/10.1038/362534a0

Arsuaga, J.L., Roura, E.C. i, Risueño, J.M.B. de C., 1994. La sierra de Atapuerca: los homínidos y sus actividades: los homínidos y sus actividades. Revista de arqueología 12–25.

Aurenche, O., Kozlowski, S.K., 1999. La naissance du néolithique au proche Orient ou Le paradis perdu. Errance, Paris.

Bard, K., Fattovich, R., Manzo, A., Perlingieri, C., 1997. Archaeological Investigations at Bieta Giyorgis (Aksum), Ethiopia: 1993-1995 Field Seasons. Journal of Field Archaeology 24, 387–403.

Barker, G., 2006. The Agricultural Revolution in Prehistory: Why did Foragers become Farmers? Oxford University Press, Oxford.

Bar-Yosef, O., 2000. The Middle and early Upper Paleolithic in Southwest Asia and neighboring regions, in: The Geography of Neandertals and Modern Humans in Europe and the Greater Mediterranean. pp. 107–156.

Bar-Yosef, O., 1994. The Lower Paleolithic of the Near East. Journal of World Prehistory 8, 211–265.

Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N.M., Kivisild, T., Torroni, A., Villems, R., 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am. J. Hum. Genet. 90, 675–684. https://doi.org/10.1016/j.ajhg.2012.03.002

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., Rosset, S., Genographic Consortium, 2008. The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82, 1130–1140. https://doi.org/10.1016/j.ajhg.2008.04.002

Bergink, S., Jentsch, S., 2009. Principles of ubiquitin and SUMO modifications in DNA repair. Nature 458, 461–467. https://doi.org/10.1038/nature07963

Blench, R., 2008. Omotic livestock terminology and its implications for the history of Afroasiatic., in: SemitoHamitic Festschrift for A.B. Dolgopolsky and H. Jungraithmayr. Dietich Reimer Verlag, Berlin, pp. 63–78.

Blench, R.M., 2006. Archaeology, Language, and the African Past. AltaMira PRess, Lanham, MD.

Blockley, S.P.E., Pinhasi, R., 2011. A revised chronology for the adoption of agriculture in the Southern Levant and the role of Lateglacial climatic change. Quaternary Science Reviews 30, 98–108. https://doi.org/10.1016/j.quascirev.2010.09.021

Boardman, S., 1999. The Agricultural Foundation of the Aksumite Empire, Ethiopia: An Interim Report., in: The Explotiation of Plant Resources in Ancient Africa. Kluwer/Plenum, New York, pp. 137–148.

Borić, D., Stefanović, S., 2004. Birth and death: infant burials from Vlasac and Lepenski Vir. Antiquity 78, 526–546. https://doi.org/10.1017/S0003598X00113201

Bräuer, G., 2008. The origin of modern anatomy: By speciation or intraspecific evolution? Evol. Anthropol. 17, 22–37. https://doi.org/10.1002/evan.20157

Breasted, J.H., 1916. Ancient times, a history of the early world: an introduction to the study of ancient history and the career of early man. Ginn, Boston.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Pääbo, S., 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. U.S.A. 104, 14616–14621. https://doi.org/10.1073/pnas.0704665104

Broodbank, C., Strasser, T.F., 1991. Migrant farmers and the Neolithic colonization of Crete. Antiquity 65, 233–245. https://doi.org/10.1017/S0003598X00079680

Brotherton, P., Endicott, P., Sanchez, J.J., Beaumont, M., Barnett, R., Austin, J., Cooper, A., 2007. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res. 35, 5717–5728. https://doi.org/10.1093/nar/gkm588

Broushaki, F., Thomas, M.G., Link, V., López, S., Dorp, L. van, Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-del-Molino, D., Kousathanas, A., Sell, C., Robson, H.K., Martiniano, R., Blöcher, J., Scheu, A., Kreutzer, S., Bollongino, R., Bobo, D., Davudi, H., Munoz, O., Currat, M., Abdi, K., Biglari, F., Craig, O.E., Bradley, D.G., Shennan, S., Veeramah, K.R., Mashkour, M., Wegmann, D., Hellenthal, G., Burger, J., 2016. Early Neolithic genomes from the eastern Fertile Crescent. Science aaf7943. https://doi.org/10.1126/science.aaf7943

Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097. https://doi.org/10.1086/521987

Cassidy, L.M., Martiniano, R., Murphy, E.M., Teasdale, M.D., Mallory, J., Hartwell, B., Bradley, D.G., 2016. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. PNAS 113, 368–373. https://doi.org/10.1073/pnas.1518445113

Cavalli-Sforza, L.L., Barrai, I., Edwards, A.W., 1964. ANALYSIS OF HUMAN EVOLUTION UNDER RANDOM GENETIC DRIFT. Cold Spring Harb. Symp. Quant. Biol. 29, 9–20.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Chapman, J.C., 2003. From Franchthi to the TiszazugL Two Early Neolithic worlds', in: Morgenrot Der Kulturen: Frühe Etappen Der Menschheits-Geschichte in Mittel-Nd Südeuropa. Archaeolingua, Budapest.

Charles, R., 1996. Back into the North: the Radiocarbon Evidence for the Human Recolonisation of the North-Western Ardennes after the Last Glacial Maximum. Proceedings of the Prehistoric Society.

Childe, G., 1934. New Light on the Most Ancient East. The Oriental Prelude to European Prehistory. Routledge and Kegan Paul, London.

Clark, J.D., Prince, G.R., 1978. Use-Wear on Later Stone Age Microliths from Laga Oda, Haraghi, Ethiopia and Possible Functional Interpretations. Azania 13, 101–110.

Consortium, T. 1000 G.P., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65. https://doi.org/10.1038/nature11632

Curtis, M.C., 2013. Archaeological Evidence for the Emergence of Food Production in the Horn of Africa., in: The Oxford Handbook of African Archaeology. Oxford University Press, pp. 571–584.

D'Anastasio, R., Wroe, S., Tuniz, C., Mancini, L., Cesana, D.T., Dreossi, D., Ravichandiran, M., Attard, M., Parr, W.C.H., Agur, A., Capasso, L., 2013. Micro-Biomechanics of the Kebara 2 Hyoid and Its Implications for Speech in Neanderthals. PLOS ONE 8, e82261. https://doi.org/10.1371/journal.pone.0082261

D'Andrea, A., Schmidt, P., Curtis, M.C., 2008. Paleobotanical Analysis and Agricultural Economy in Early First Millennium BCE Sites around Asmara., in: The Archaeoloy of Ancient Eritrea. Red Sea Press, Trenton, NJ, pp. 207–216.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Filippo, C., Bostoen, K., Stoneking, M., Pakendorf, B., 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc Biol Sci 279, 3256–3263. https://doi.org/10.1098/rspb.2012.0318

Denaro, M., Blanc, H., Johnson, M.J., Chen, K.H., Wilmsen, E., Cavalli-Sforza, L.L., Wallace, D.C., 1981. Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proc. Natl. Acad. Sci. U.S.A. 78, 5768–5772.

Diakanoff, I., 1998. The earliest Semitic society. J. Semit. Stud. 43, 209–17.

Diamond, J., Bellwood, P., 2003. Farmers and Their Languages: The First Expansions. Science 300, 597–603. https://doi.org/10.1126/science.1078208

Ehret, C., 1995. Reconstructing Proto-Afroasiatic (Proto-Afrasian): Vowels, Tone, Consonants, and Vocabulary. University of Califormia Press, Berkeley, CA.

Ehret, C., Keita, S.O.Y., Newman, P., 2004. On the African Origins of the Afroasiatic Language Family: A Response to Diamond and Bellwood. Science 306, 1680–1683.

Eriksson, A., Betti, L., Friend, A.D., Lycett, S.J., Singarayer, J.S., Cramon-Taubadel, N. von, Valdes, P.J., Balloux, F., Manica, A., 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc. Natl. Acad. Sci. U.S.A. 109, 16089–16094. https://doi.org/10.1073/pnas.1209494109

Eriksson, A., Manica, A., 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. U.S.A. 109, 13956–13960. https://doi.org/10.1073/pnas.1200567109

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D., Pääbo, S., 2015. An early modern human from Romania with a recent Neanderthal ancestor. Nature 524, 216–219. https://doi.org/10.1038/nature14558

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., Pääbo, S., 2013. DNA analysis of an early modern human from Tianyuan Cave, China. PNAS 110, 2223–2227. https://doi.org/10.1073/pnas.1221359110

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., Skoglund, P., Derevianko, A.P., Drozdov, N., Slavinsky, V., Tsybankov, A., Cremonesi, R.G., Mallegni, F., Gély, B., Vacca, E., Morales, M.R.G., Straus, L.G., Neugebauer-Maresch, C., Teschler-Nicola, M., Constantin, S., Moldovan, O.T., Benazzi, S., Peresani, M., Coppola, D., Lari, M., Ricci, S., Ronchitelli, A., Valentin, F., Thevenet, C., Wehrberger, K., Grigorescu, D., Rougier, H., Crevecoeur, I., Flas, D., Semal, P., Mannino, M.A., Cupillard, C., Bocherens, H., Conard, N.J., Harvati, K., Moiseyev, V., Drucker, D.G., Svoboda, J., Richards, M.P., Caramelli, D., Pinhasi, R., Kelso, J., Patterson, N., Krause, J., Pääbo, S., Reich, D., 2016. The genetic history of Ice Age Europe. Nature 534, 200–205. https://doi.org/10.1038/nature17993

Fuller, D.Q., Kingwell-Banham, E., Lucas, L., Murphy, C., Stevens, C., 2015. Comparing Pathways to Agriculture. Archaeology International 18. https://doi.org/10.5334/ai.1808

Fuller, D.Q., Willcox, G., Allaby, R.G., 2012. Early agricultural pathways: moving outside the "core area" hypothesis in Southwest Asia. J. Exp. Bot. 63, 617–633. https://doi.org/10.1093/jxb/err307

Gabunia, L., Vekua, A., Lordkipanidze, D., Swisher, C.C., Ferring, R., Justus, A., Nioradze, M., Tvalchrelidze, M., Antón, S.C., Bosinski, G., Jöris, O., Lumley, M.-A. -d., Majsuradze, G., Mouskhelishvili, A., 2000. Earliest Pleistocene Hominid Cranial Remains from Dmanisi, Republic of Georgia: Taxonomy, Geological Setting, and Age. Science 288, 1019–1025. https://doi.org/10.1126/science.288.5468.1019

Gallego-Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D.G., Bhak, J., Pinhasi, R., Manica, A., 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. Science 350, 820–822. https://doi.org/10.1126/science.aad2879

Gallego-Llorente, M., Connell, S., Jones, E.R., Merrett, D.C., Jeon, Y., Eriksson, A., Siska, V., Gamba, C., Meiklejohn, C., Beyer, R., Jeon, S., Cho, Y.S., Hofreiter, M., Bhak, J., Manica, A., Pinhasi, R., 2016. The genetics of an early Neolithic pastoralist from the Zagros, Iran. Sci Rep 6. https://doi.org/10.1038/srep31326

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Bradley, D.G., Orlando, L., 2016. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol Ecol Resour 16, 459–469. https://doi.org/10.1111/1755-0998.12470

Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T.F.G., Hofreiter, M., Bradley, D.G., Pinhasi, R., 2014. Genome flux and stasis in a five millennium transect of European prehistory. Nat. Commun. 5, 5257. https://doi.org/10.1038/ncomms6257

Gangal, K., Sarson, G.R., Shukurov, A., 2014. The Near-Eastern Roots of the Neolithic in South Asia. PLoS ONE 9, e95714. https://doi.org/10.1371/journal.pone.0095714

García-Garcerà, M., Gigli, E., Sanchez-Quinto, F., Ramirez, O., Calafell, F., Civit, S., Lalueza-Fox, C., 2011. Fragmentation of Contaminant and Endogenous DNA in Ancient Samples Determined by Shotgun Sequencing; Prospects for Human Palaeogenomics. PLOS ONE 6, e24161. https://doi.org/10.1371/journal.pone.0024161

Garrod, D.A.E., 1937. The Near east as a gateway of prehistoric migration. Bulletin of the American School of Prehistoric Research 13, 17–21.

Garrod, D.A.E., Bate, D., 1937. The Stone And of Mount Carmel. Volume 1: Excavations at the Wadi el-Mughara. Oxford University Press, Oxford.

Gebremeskel, E.I., Ibrahim, M.E., 2014. Y-chromosome E haplogroups: their distribution and implication to the origin of Afro-Asiatic languages and pastoralism. Eur. J. Hum. Genet. 22, 1387–1392. https://doi.org/10.1038/ejhg.2014.41

Gimbutas, M.A., 2001. The Living Goddesses. University of Califormia Press.

Gimbutas, M.A., Dexter, M.R., 1997. The Kurgan culture and the Indo-Europeanization of Europe : selected articles from 1952 to 1993. Institute for the Study of Man, Washington, DC.

Goring-Morris, A.N., Belfer-Cohen, A., 2011. Neolithization Processes in the Levant: The Outer Envelope. Current Anthropology 52, S195–S208. https://doi.org/10.1086/658860

Green, R.E., Briggs, A.W., Krause, J., Prüfer, K., Burbano, H.A., Siebauer, M., Lachmann, M., Pääbo, S., 2009. The Neandertal genome and ancient DNA authenticity. EMBO J. 28, 2494–2502. https://doi.org/10.1038/emboj.2009.222

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., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, 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, Z., Gusic, 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., Paabo, S., 2010. A Draft Sequence of the Neandertal Genome. Science 328, 710–722.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., Pääbo, S., 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444, 330–336. https://doi.org/10.1038/nature05336

Günther, T., Valdiosera, C., Malmström, H., Ureña, I., Rodriguez-Varela, R., Sverrisdóttir, Ó.O., Daskalaki, E.A., Skoglund, P., Naidoo, T., Svensson, E.M., Castro, J.M.B. de, Carbonell, E., Dunn, M., Storå, J., Iriarte, E., Arsuaga, J.L., Carretero, J.-M., Götherström, A., Jakobsson, M., 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. PNAS 112, 11917–11922. https://doi.org/10.1073/pnas.1509851112

Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tänzer, M., Villems, R., Renfrew, C., Gronenborn, D., Alt, K.W., Burger, J., 2005. Ancient DNA from the First European Farmers in 7500-Year-Old Neolithic Sites. Science 310, 1016–1018. https://doi.org/10.1126/science.1118725

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R.G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S.L., Risch, R., Rojo Guerra, M.A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K.W., Reich, D., 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211. https://doi.org/10.1038/nature14317

Hammer, M.F., Woerner, A.E., Mendez, F.L., Watkins, J.C., Wall, J.D., 2011. Genetic evidence for archaic admixture in Africa. Proc. Natl. Acad. Sci. U.S.A. 108, 15123–15128. https://doi.org/10.1073/pnas.1109300108

Harris, D.R., Asouti, E., Bogaard, A., Charles, M., Conolly, J., Coolidge, J., Dobney, K., Gosden, C., Heathcote, J., Jaques, D., Larkum, M., Limbrey, S., Meadows, J., Schlanger, N., Wilkinson, K., 2010. Origins of Agriculture in Western Central Asia: An Environmental-Archaeological Study. University of Pennsylvania Press.

Harrower, M., McCorriston, M., D'Andrea, A., 2010. General/Specific, Local/Global: Comparing the Beginnings of Agriculture in the Horn of Africa (Ethiopia/Eritrea) and Southwest Arabia (Yemen). Am. Antiq. 75, 452–472.

Hart, K.L., Kimura, S.L., Mushailov, V., Budimlija, Z.M., Prinz, M., Wurmbach, E., 2013. Improved eye- and skin-color prediction based on 8 SNPs. Croat. Med. J. 54, 248–256.

Higham, T., Douka, K., Wood, R., Ramsey, C.B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., Barroso-Ruíz, C., Bergman, C., Boitard, C., Boscato, P., Caparrós, M., Conard, N.J., Draily, C., Froment, A., Galván, B., Gambassini, P., Garcia-Moreno, A., Grimaldi, S., Haesaerts, P., Holt, B., Iriarte-Chiapusso, M.-J., Jelinek, A., Jordá Pardo, J.F., Maíllo-Fernández, J.-M., Marom, A., Maroto, J., Menéndez, M., Metz, L., Morin, E., Moroni, A., Negrino, F., Panagopoulou, E., Peresani, M., Pirson, S., de la Rasilla, M., Riel-Salvatore, J., Ronchitelli, A., Santamaria, D., Semal, P., Slimak, L., Soler, J., Soler, N., Villaluenga, A., Pinhasi, R., Jacobi, R., 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. Nature 512, 306–309. https://doi.org/10.1038/nature13621

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978. https://doi.org/10.1002/joc.1276

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-del-Molino, D., Dorp, L. van, López, S., Kousathanas, A., Link, V., Kirsanow, K., Cassidy, L.M., Martiniano,

R., Strobel, M., Scheu, A., Kotsakis, K., Halstead, P., Triantaphyllou, S., Kyparissi-Apostolika, N., Urem-Kotsou, D., Ziota, C., Adaktylou, F., Gopalan, S., Bobo, D.M., Winkelbach, L., Blöcher, J., Unterländer, M., Leuenberger, C., Çilingiroğlu, Ç., Horejs, B., Gerritsen, F., Shennan, S.J., Bradley, D.G., Currat, M., Veeramah, K.R., Wegmann, D., Thomas, M.G., Papageorgopoulou, C., Burger, J., 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. PNAS 113, 6886–6891. https://doi.org/10.1073/pnas.1523951113

Hofreiter, M., Paijmans, J.L.A., Goodchild, H., Speller, C.F., Barlow, A., Fortes, G.G., Thomas, J.A., Ludwig, A., Collins, M.J., 2015. The future of ancient DNA: Technical advances and conceptual shifts. Bioessays 37, 284–293. https://doi.org/10.1002/bies.201400160

Housley, R.A., Gamble, C.S., Street, M., Pettitt, P., 1997. Radiocarbon evidence for the Lateglacial Human Recolonisation of Northern Europe. Proceedings of the Prehistoric Society 63, 25–54. https://doi.org/10.1017/S0079497X0000236X

Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A., Weale, M.E., Bradman, N., Bekele, E., Kivisild, T., Tyler-Smith, C., Nielsen, R., 2013. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Mol. Biol. Evol. 30, 1877–1888. https://doi.org/10.1093/molbev/mst089

Hutchison, C.A., Newbold, J.E., Potter, S.S., Edgell, M.H., 1974. Maternal inheritance of mammalian mitochondrial DNA. Nature 251, 536–538. https://doi.org/10.1038/251536a0

Jensen, T.G.K., Liebert, A., Lewinsky, R., Swallow, D.M., Olsen, J., Troelsen, J.T., 2011. The -14010*C variant associated with lactase persistence is located between an Oct-1 and HNF1α binding site and increases lactase promoter activity. Hum. Genet. 130, 483–493. https://doi.org/10.1007/s00439-011-0966-0

Jochim, M., 1998. A Hunter-Gatherer Landscape: Southwest ermany in the Late Pleistocene and Mesolithic. Plenum Press, New York.

Johanson, D.C., White, T.D., 1979. A systematic assessment of early African hominids. Science 203, 321–330. https://doi.org/10.1126/science.104384

Jones, B.L., Raga, T.O., Liebert, A., Zmarz, P., Bekele, E., Danielsen, E.T., Olsen, A.K., Bradman, N., Troelsen, J.T., Swallow, D.M., 2013. Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. The American Journal of Human Genetics 93, 538–544. https://doi.org/10.1016/j.ajhg.2013.07.008

Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego-Llorente, M., Cassidy, L.M., Gamba, C., Meshveliani, T., Bar-Yosef, O., Müller, W., Belfer-Cohen, A., Matskevich, Z., Jakeli, N., Higham, T.F.G., Currat, M., Lordkipanidze, D., Hofreiter, M., Manica, A., Pinhasi, R., Bradley, D.G., 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat Commun 6, 8912. https://doi.org/10.1038/ncomms9912

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., Orlando, L., 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Jostins, L., Xu, Y., McCarthy, S., Ayub, Q., Durbin, R., Barrett, J., Tyler-Smith, C., 2014. YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. arXiv:1407.7988 [q-bio].

Keller, A., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., Maixner, F., Leidinger, P., Backes, C., Khairat, R., Forster, M., Stade, B., Franke, A., Mayer, J., Spangler, J., McLaughlin,

S., Shah, M., Lee, C., Harkins, T.T., Sartori, A., Moreno-Estrada, A., Henn, B., Sikora, M., Semino, O., Chiaroni, J., Rootsi, S., Myres, N.M., Cabrera, V.M., Underhill, P.A., Bustamante, C.D., Vigl, E.E., Samadelli, M., Cipollini, G., Haas, J., Katus, H., O'Connor, B.D., Carlson, M.R.J., Meder, B., Blin, N., Meese, E., Pusch, C.M., Zink, A., 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3, 698. https://doi.org/10.1038/ncomms1701

Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M., Wilson, J.F., 2010. Genomic Runs of Homozygosity Record Population History and Consanguinity. PLOS ONE 5, e13996. https://doi.org/10.1371/journal.pone.0013996

Kitchen, A., Ehret, C., Assefa, S., Mulligan, C.J., 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proceedings of the Royal Society of London B: Biological Sciences 276, 2703–2710. https://doi.org/10.1098/rspb.2009.0408

Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., Villems, R., 2004. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. Am. J. Hum. Genet. 75, 752–770.

Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., Knijff, P. de, Feldman, M., Cavalli-Sforza, L.L., Oefner, P.J., 2006. The Role of Selection in the Evolution of Human Mitochondrial Genomes. Genetics 172, 373–387. https://doi.org/10.1534/genetics.105.043901

Klein, R.G., 1995. Anatomy, behavior, and modern human origins. J World Prehist 9, 167–198. https://doi.org/10.1007/BF02221838

Klein, R.G., 1989. The human career: human biological and cultural origins. University of Chicago Press.

Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinform. 15, 356. https://doi.org/10.1186/s12859-014-0356-4

Kosse, K., 1979. Settlement ecology of the Early and Middle Neolithic Körös and Linear Pottery Cultures in Hungary /. B.A.R., Oxford, England :

Lachance, J., Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays 35, 780–786. https://doi.org/10.1002/bies.201300014

Lam, Y.M., Chen, X., Pearson, O.M., 1999. Intertaxonomic Variability in Patterns of Bone Density and the Differential Representation of Bovid, Cervid, and Equid Elements in the Archaeological Record. American Antiquity 64, 343–362. https://doi.org/10.2307/2694283

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L.,

Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., Bastide, M. de la, Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E.R., Roodenberg, S.A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J.M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S.M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D.A., O'Reilly, S., Richards, M.B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J.F., Yengo, L., Hovhannisyan, N.A., Patterson, N., Pinhasi, R., Reich, D., 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536, 419–424. https://doi.org/10.1038/nature19310

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K.I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H.A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C.M., Brisighelli, F., Busby, G.B.J., Cali, F., Churnosov, M., Cole, D.E.C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S.A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B.M., Hervig, T., Hodoglugil, U., Jha, A.R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R.W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A.,

Starikovskaya, E.B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C.A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M.G., Ruiz-Linares, A., Tishkoff, S.A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E.E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., Krause, J., 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413. https://doi.org/10.1038/nature13673

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. https://doi.org/10.1038/nature10231

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Schlebusch, C., Jakobsson, M., 2014. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc. Biol. Sci. 281. https://doi.org/10.1098/rspb.2014.1448

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schröder, R., Stoneking, M., 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig Genet 5, 13. https://doi.org/10.1186/2041-2223-5-13

Lockard, C.A., 2010. Societies, Networks, and Transitions, Volume 1: To 1500. Cengage Learning, India.

Lowery, R.K., Uribe, G., Jimenez, E.B., Weiss, M.A., Herrera, K.J., Regueiro, M., Herrera, R.J., 2013. Neanderthal and Denisova genetic affinities with contemporary humans: Introgression versus common ancestral polymorphisms. Gene 530, 83–94. https://doi.org/10.1016/j.gene.2013.06.005

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N.K., Raja, J.M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H.-J., Oppenheimer, S., Torroni, A., Richards, M., 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034–6.

MacHugh, D.E., Edwards, C.J., Bailey, J.F., Bancroft, D.R., Bradley, D.G., 2000. The extraction and analysis of ancient DNA from bone and teeth: a survey of current methodologies. Ancient Biomolecules 3, 81–102.

Manica, A., Amos, W., Balloux, F., Hanihara, T., 2007. The effect of ancient population bottlenecks on human phenotypic variation. Nature 448, 346–348.

Marshall, F., Steward, K., Barthelme, J., 1984. Early Domestic Stock at Dongodien in Northern Kenya. Azania 19, 120–127.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Martiniano, R., Cassidy, L.M., Ó'Maoldúin, R., McLaughlin, R., Silva, N.M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M.J., Serra, M., Burger, J., Parreira, R., Moran, E., Valera, A.C., Porfirio, E., Boaventura, R., Silva, A.M., Bradley, D.G., 2017. The population genomics of archaeological transition in west Iberia: Investigation of ancient

substructure using imputation and haplotype-based methods. PLOS Genetics 13, e1006852. https://doi.org/10.1371/journal.pgen.1006852

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E.R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J.L., de Castro, J.M.B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M.A.R., Roodenberg, J., Vergès, J.M., Krause, J., Cooper, A., Alt, K.W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., Reich, D., 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499–503. https://doi.org/10.1038/nature16152

Matthews, R., Nashli, H.F., 2013. The Neolithisation of Iran, BANEA monograph Series. Oxbow Books.

McDougall, I., Brown, F.H., Fleagle, J.G., 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433, 733–736. https://doi.org/10.1038/nature03258

McEvoy, B.P., Powell, J.E., Goddard, M.E., Visscher, P.M., 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res 21, 821–829. https://doi.org/10.1101/gr.119636.110

McHenry, H.M., 1992. How big were early hominids? Evol. Anthropol. 1, 15–20. https://doi.org/10.1002/evan.1360010106

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

McPherson, K., 2009. Trading Encounters: from the Euphrates to the Indus in the Bronze Age. International Journal of Nautical Archaeology 38, 429–430. https://doi.org/10.1111/j.1095-9270.2009.00244_4.x

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

Mellars, P., 1989. Major Issues in the Emergence of Modern Humans. Current Anthropology 30, 349–385.

Merrett, D.C., 2004. Bioarchaeology in Early Neolithic Iran: assessment of health status and subsistence strategy. Unpublished Ph.D. Dissertation. (Unpublished Ph.D. thesis). University of Manitoba.

Merriwether, D.A., Clark, A.G., Ballinger, S.W., Schurr, T.G., Soodyall, H., Jenkins, T., Sherry, S.T., Wallace, D.C., 1991. The structure of human mitochondrial DNA variation. J. Mol. Evol. 33, 543–555.

Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J.M.B., Carbonell, E., Viola, B., Kelso, J., Prüfer, K., Pääbo, S., 2016. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531, 504–507. https://doi.org/10.1038/nature17405

Meyer, M., Kircher, M., 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. 2010, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C. de, Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon,

A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2012a. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338, 222–226. https://doi.org/10.1126/science.1224344

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C. de, Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2012b. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338, 222–226. https://doi.org/10.1126/science.1224344

Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L., Reich, D., 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 7, e1001373. https://doi.org/10.1371/journal.pgen.1001373

Morris, A.G., Heinze, A., Chan, E.K.F., Smith, A.B., Hayes, V.M., 2014. First Ancient Mitochondrial Human Genome from a Prepastoralist Southern African. Genome Biol. Evol. 6, 2647–2653. https://doi.org/10.1093/gbe/evu202

Olalde, I., Allentoft, M.E., Sánchez-Quinto, F., Santpere, G., Chiang, C.W.K., DeGiorgio, M., Prado-Martinez, J., Rodríguez, J.A., Rasmussen, S., Quilez, J., Ramírez, O., Marigorta, U.M., Fernández-Callejo, M., Prada, M.E., Encinas, J.M.V., Nielsen, R., Netea, M.G., Novembre, J., Sturm, R.A., Sabeti, P., Marquès-Bonet, T., Navarro, A., Willerslev, E., Lalueza-Fox, C., 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 507, 225–228. https://doi.org/10.1038/nature12960

Olalde, I., Schroeder, H., Sandoval-Velasco, M., Vinner, L., Lobón, I., Ramirez, O., Civit, S., Borja, P.G., Salazar-García, D.C., Talamo, S., Fullola, J.M., Oms, F.X., Pedro, M., Martínez, P., Sanz, M., Daura, J., Zilhão, J., Marquès-Bonet, T., Gilbert, M.T.P., Lalueza-Fox, C., 2015. A common genetic origin for early farmers from Mediterranean Cardial and Central European LBK cultures. Mol Biol Evol msv181. https://doi.org/10.1093/molbev/msv181

Omrak, A., Günther, T., Valdiosera, C., Svensson, E.M., Malmström, H., Kiesewetter, H., Aylward, W., Storå, J., Jakobsson, M., Götherström, A., 2016. Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. Current Biology 26, 270–275. https://doi.org/10.1016/j.cub.2015.12.019

Oppenheimer, S., 2012. A single southern exit of modern humans from Africa: Before or after Toba? Quaternary International, The Toba Volcanic Super-eruption of 74,000 Years Ago: Climate Change, Environments, and Evolving Humans 258, 88–99. https://doi.org/10.1016/j.quaint.2011.07.049

Ovchinnikov, I.V., Götherström, A., Romanova, G.P., Kharitonov, V.M., Lidén, K., Goodwin, W., 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 404, 490–493. https://doi.org/10.1038/35006625

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., Bekele, E., Bradman, N., Balding, D.J., Tyler-Smith, C., 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. 91, 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., Mekonnen, E., Luiselli, D., Bradman, N., Bekele, E., Zalloua, P., Durbin, R., Kivisild, T., Tyler-Smith, C., 2015. Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am. J. Hum. Genet. 96, 986–991. https://doi.org/10.1016/j.ajhg.2015.04.019

Pankhurst, R., 1998. The Ethiopians. Blackwell Publishers Ltd, Oxford.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J.B., Fernandes, V., Pereira, L., Veen, L.V. der, Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J.-M., Quintana-Murci, L., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356, 543–546. https://doi.org/10.1126/science.aal1988

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient Admixture in Human History. Genetics 192, 1065–1093.

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2, e190. https://doi.org/10.1371/journal.pgen.0020190

Patterson, N.J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient Admixture in Human History. Genetics 192, 1065–1093. https://doi.org/10.1534/genetics.112.145037

Pearson, O.M., 2008. Statistical and biological definitions of "anatomically modern" humans: Suggestions for a unified approach to modern morphology. Evol. Anthropol. 17, 38–48. https://doi.org/10.1002/evan.20155

Peltenburg, E., Colledge, S., Croft, P., Jackson, A., McCartney, C., Murray, M.A., 2000. Agro-pastoralist colonization of Cyprus in the 10th millennium BP: initial assessments. Antiquity 74, 844–853. https://doi.org/10.1017/S0003598X0006049X

Perlès, C., 2001. The Early Neolithic in Greece. Cambridge University Press, Cambridge.

Phillipson, D., 2005. African Archaeology. Cambridge University Press, Cambridge, UK.

Phillipson, D.W., 1998. Ancient Ethiopia: Aksum, Its Antecedents and Successors. British Museum Press, London.

Phillipson, D.W., 1993. The antiquity of cultivation and herding in Ethiopia., in: T. Shaw, P. Sinclair, B. Andah, A. Okpoko (Eds.) The Archaeology of Africa: Food, Metals and Towns. Routledge, London, pp. 344–357.

Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Gueldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P.-R., Lachance, J., Mountain, J., Bustamante, C.D., Berger, B., Tishkoff, S.A., Henn, B.M., Stoneking, M., Reich, D., Pakendorf, B., 2012. The genetic prehistory of southern Africa. Nat. Commun. 3, 1143. https://doi.org/10.1038/ncomms2140

Pickrell, J.K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., Reich, D., 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. U.S.A. 111, 2632–2637. https://doi.org/10.1073/pnas.1313787111

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., Raczky, P., Anders, A., Pietrusewsky, M., Rollefson, G., Jovanovic, M., Trinhhoang, H., Bar-Oz, G., Oxenham, M., Matsumura,

H., Hofreiter, M., 2015. Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. PLOS ONE 10, e0129102. https://doi.org/10.1371/journal.pone.0129102

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49. https://doi.org/10.1038/nature12886

Prugnolle, F., Manica, A., Balloux, F., 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15, R159-60.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795

Qin, P., Stoneking, M., 2015. Denisovan Ancestry in East Eurasian and Native American Populations. Mol. Biol. Evol. 32, 2665–2674. https://doi.org/10.1093/molbev/msv141

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Development Core Team, 2001. R: A Language and Environment for Statistical Computing. the R Foundation for Statistical Computing, Vienna, Austria.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford Jr, T.W., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Mägi, R., Campos, P.F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L.P., Fedorova, S.A., Voevoda, M.I., DeGiorgio, M., Sicheritz-Ponten, T., Brunak, S., Demeshchenko, S., Kivisild, T., Villems, R., Nielsen, R., Jakobsson, M., Willerslev, E., 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505, 87–91. https://doi.org/10.1038/nature12736

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspinas, A.-S., Eriksson, A., Moltke, I., Metspalu, M., Homburger, J.R., Wall, J., Cornejo, O.E., Moreno-Mayar, J.V., Korneliussen, T.S., Pierre, T., Rasmussen, M., Campos, P.F., Damgaard, P. de B., Allentoft, M.E., Lindo, J., Metspalu, E., Rodríguez-Varela, R., Mansilla, J., Henrickson, C., Seguin-Orlando, A., Malmström, H., Stafford, T., Shringarpure, S.S., Moreno-Estrada, A., Karmin, M., Tambets, K., Bergström, A., Xue, Y., Warmuth, V., Friend, A.D., Singarayer, J., Valdes, P., Balloux, F., Leboreiro, I., Vera, J.L., Rangel-Villalobos, H., Pettener, D., Luiselli, D., Davis, L.G., Heyer, E., Zollikofer, C.P.E., Ponce de León, M.S., Smith, C.I., Grimes, V., Pike, K.-A., Deal, M., Fuller, B.T., Arriaza, B., Standen, V., Luz, M.F., Ricaut, F., Guidon, N., Osipova, L., Voevoda, M.I., Posukh, O.L., Balanovsky, O., Lavryashina, M., Bogunov, Y., Khusnutdinova, E., Gubina, M., Balanovska, E., Fedorova, S., Litvinov, S., Malyarchuk, B., Derenko, M., Mosher, M.J., Archer, D., Cybulski, J., Petzelt, B., Mitchell, J., Worl, R., Norman, P.J., Parham, P., Kemp, B.M., Kivisild, T., Tyler-Smith, C., Sandhu, M.S., Crawford, M., Villems, R., Smith, D.G., Waters, M.R., Goebel, T., Johnson, J.R., Malhi, R.S., Jakobsson, M., Meltzer, D.J., Manica, A., Durbin, R., Bustamante, C.D., Song, Y.S., Nielsen, R.,

Willerslev, E., 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. https://doi.org/10.1126/science.aab3884

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., Kivisild, T., Zhai, W., Eriksson, A., Manica, A., Orlando, L., De La Vega, F.M., Tridico, S., Metspalu, E., Nielsen, K., Ávila-Arcos, M.C., Moreno-Mayar, J.V., Muller, C., Dortch, J., Gilbert, M.T.P., Lund, O., Wesolowska, A., Karmin, M., Weinert, L.A., Wang, B., Li, J., Tai, S., Xiao, F., Hanihara, T., Van Driem, G., Jha, A.R., Ricaut, F.-X., De Knijff, P., Migliano, A.B., Gallego Romero, I., Kristiansen, K., Lambert, D.M., Brunak, S., Forster, P., Brinkmann, B., Nehlich, O., Bunce, M., Richards, M., Gupta, R., Bustamante, C.D., Krogh, A., Foley, R.A., Lahr, M.M., Balloux, F., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Wang, J., Willerslev, E., 2011a. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94–98. https://doi.org/10.1126/science.1211177

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., Kivisild, T., Zhai, W., Eriksson, A., Manica, A., Orlando, L., Vega, F.M.D.L., Tridico, S., Metspalu, E., Nielsen, K., Ávila-Arcos, M.C., Moreno-Mayar, J.V., Muller, C., Dortch, J., Gilbert, M.T.P., Lund, O., Wesolowska, A., Karmin, M., Weinert, L.A., Wang, B., Li, J., Tai, S., Xiao, F., Hanihara, T., Driem, G. van, Jha, A.R., Ricaut, F.-X., Knijff, P. de, Migliano, A.B., Romero, I.G., Kristiansen, K., Lambert, D.M., Brunak, S., Forster, P., Brinkmann, B., Nehlich, O., Bunce, M., Richards, M., Gupta, R., Bustamante, C.D., Krogh, A., Foley, R.A., Lahr, M.M., Balloux, F., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Wang, J., Willerslev, E., 2011b. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94–98. https://doi.org/10.1126/science.1211177

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M.T.P., Wang, Y., Raghavan, M., Campos, P.F., Kamp, H.M., Wilson, A.S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E.D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T.L., Grønnow, B., Meldgaard, M., Andreasen, C., Fedorova, S.A., Osipova, L.P., Higham, T.F.G., Ramsey, C.B., Hansen, T. v O., Nielsen, F.C., Crawford, M.H., Brunak, S., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., Willerslev, E., 2010a. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463, 757–762. https://doi.org/10.1038/nature08835

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M.T.P., Wang, Y., Raghavan, M., Campos, P.F., Kamp, H.M., Wilson, A.S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E.D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T.L., Gr\onnow, B., Meldgaard, M., Andreasen, C., Fedorova, S.A., Osipova, L.P., Higham, T.F.G., Ramsey, C.B., Hansen, T.V.O., Nielsen, F.C., Crawford, M.H., Brunak, S., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., Willerslev, E., 2010b. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463, 757–62.

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., Pääbo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053–1060. https://doi.org/10.1038/nature09710

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H.-V., Parik, J., Loogväli, E.-L., Derenko, M., Malyarchuk, B., Bermisheva, M., Zhadanov, S., Pennarun, E., Gubina, M., Golubenko, M., Damba, L., Fedorova, S., Gusar, V., Grechanina, E., Mikerezi, I., Moisan, J.-P., Chaventré, A., Khusnutdinova, E., Osipova, L., Stepanov, V., Voevoda, M., Achilli, A., Rengo, C., Rickards, O., De Stefano, G.F., Papiha, S., Beckman, L., Janicijevic, B., Rudan, P., Anagnou, N., Michalodimitrakis, E., Koziel, S., Usanga, E., Geberhiwot, T., Herrnstadt, C., Howell, N., Torroni, A., Villems, R., 2003. Origin and Diffusion of mtDNA Haplogroup X. Am J Hum Genet 73, 1178–1190.

Renaud, G., Slon, V., Duggan, A.T., Kelso, J., 2015. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol 16. https://doi.org/10.1186/s13059-015-0776-0

Renaud, G., Stenzel, U., Kelso, J., 2014. leeHom: adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Res. 42, e141. https://doi.org/10.1093/nar/gku699

Renfrew, A.C., 1987. Archaeology and Language: The Puzzle of Indo-European Origins. Pimlico, London.

Richards, M., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.J., Sykes, B., 1996. Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am. J. Hum. Genet. 59, 185–203.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Gölge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Calì, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Nørby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozari, R., Torroni, A., Bandelt, H.J., 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67, 1251–1276.

Riehl, S., Zeidi, M., Conard, N.J., 2013. Emergence of Agriculture in the Foothills of the Zagros Mountains of Iran. Science 341, 65–67. https://doi.org/10.1126/science.1236743

Rightmire, G.P., 1993. The Evolution of Homo Erectus: Comparative Anatomical Studies of an Extinct Human Species. Cambridge University Press.

Rightmire, G.P., 1992. Homo erectus: Ancestor or evolutionary side branch? Evol. Anthropol. 1, 43–49. https://doi.org/10.1002/evan.1360010204

Robertshaw, P., 1987. Prehistory in the Upper Nile Basin. Jouenal of African History 28, 177–189.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat Biotech 29, 24–26. https://doi.org/10.1038/nbt.1754

Rohland, N., Reich, D., 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 22, 939–946. https://doi.org/10.1101/gr.128124.111

Rowley-Conwy, P., 2004. How the West was lost : a reconsideration of agricultural origins in Britain, Ireland and southern Scandinavia. Current anthropology. 45, 83–113. http://dx.doi.org/10.1086/422083

Sánchez-Quinto, F., Botigué, L.R., Civit, S., Arenas, C., Ávila-Arcos, M.C., Bustamante, C.D., Comas, D., Lalueza-Fox, C., 2012a. North African Populations Carry the Signature of Admixture with Neandertals. PLoS ONE 7, e47765. https://doi.org/10.1371/journal.pone.0047765

Sánchez-Quinto, F., Schroeder, H., Ramirez, O., Ávila-Arcos, M.C., Pybus, M., Olalde, I., Velazquez, A.M.V., Marcos, M.E.P., Encinas, J.M.V., Bertranpetit, J., Orlando, L., Gilbert, M.T.P., Lalueza-Fox, C., 2012b. Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. Curr. Biol. 22, 1494–1499. https://doi.org/10.1016/j.cub.2012.06.005

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., Reich, D., 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507, 354–357. https://doi.org/10.1038/nature12961

Scheu, A., Powell, A., Bollongino, R., Vigne, J.-D., Tresset, A., Çakırlar, C., Benecke, N., Burger, J., 2015. The genetic prehistory of domesticated cattle from their origin to the spread across Europe. BMC Genetics 16, 54. https://doi.org/10.1186/s12863-015-0203-2

Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Steyn, M., Soodyall, H., Lombard, M., Jakobsson, M., 2017. Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago. bioRxiv 145409. https://doi.org/10.1101/145409

Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J.D., Nigst, P.R., Foley, R.A., Lahr, M.M., Nielsen, R., Orlando, L., Willerslev, E., 2014. Genomic structure in Europeans dating back at least 36,200 years. Science 346, 1113–1118. https://doi.org/10.1126/science.aaa0114

Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P.J., Zhivotovsky, L.A., King, R., Torroni, A., Cavalli-Sforza, L.L., Underhill, P.A., Santachiara-Benerecetti, A.S., 2004. Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area. Am. J. Hum. Genet. 74, 1023–1034.

Semino, O., Santachiara-Benerecetti, A.S., Falaschi, F., Cavalli-Sforza, L.L., Underhill, P.A., 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am. J. Hum. Genet. 70, 265–268. https://doi.org/10.1086/338306

Shapiro, B., Hofreiter, M., 2014. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. Science 343, 1236573. https://doi.org/10.1126/science.1236573

Shea, J.J., 2011. Homo sapiens Is as Homo sapiens Was: Behavioral Variability versus Behavioral Modernity in Paleolithic Archaeology. Current Anthropology 52, 1–35. https://doi.org/10.1086/658067

Sikora, M., Carpenter, M.L., Moreno-Estrada, A., Henn, B.M., Underhill, P.A., Sánchez-Quinto, F., Zara, I., Pitzalis, M., Sidore, C., Busonero, F., Maschio, A., Angius, A., Jones, C., Mendoza-Revilla, J., Nekhrizov, G., Dimitrova, D., Theodossiev, N., Harkins, T.T., Keller, A., Maixner, F., Zink, A., Abecasis, G., Sanna, S., Cucca, F., Bustamante, C.D., 2014. Population Genomic Analysis of Ancient and Modern Genomes Yields New Insights into the Genetic Ancestry of the Tyrolean Iceman and the Genetic

Structure of Europe. PLoS Genet 10, e1004353. https://doi.org/10.1371/journal.pgen.1004353

Siska, V., Jones, E.R., Jeon, S., Bhak, Y., Kim, H.-M., Cho, Y.S., Kim, H., Lee, K., Veselovskaya, E., Balueva, T., Gallego-Llorente, M., Hofreiter, M., Bradley, D.G., Eriksson, A., Pinhasi, R., Bhak, J., Manica, A., 2017. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. Sci Adv 3, e1601877. https://doi.org/10.1126/sciadv.1601877

Skoglund, P., Malmström, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., Hall, P., Tambets, K., Parik, J., Sjögren, K.-G., Apel, J., Willerslev, E., Storå, J., Götherström, A., Jakobsson, M., 2014. Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. Science 344, 747–750. https://doi.org/10.1126/science.1253448

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M.T.P., Götherström, A., Jakobsson, M., 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. Science 336, 466–469. https://doi.org/10.1126/science.1216304

Skoglund, P., Mittnik, A., Sirak, K., Hajdinjak, M., Rohland, N., Mallick, S., Salie, T., Heinze, A., Meyer, M., Peltzer, A., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Roman, J., Chiumia, C., Crowther, A., Gomani-Chindebvu, E., Helm, R., Horton, M., Morris, A.G., Parkington, J., Prendergast, M.E., Ramesar, R., Shipton, C., Thompson, J., Tibesasa, R., Hayes, V.M., Pääbo, S., Patterson, N., Boivin, N., Pinhasi, R., Krause, J., Reich, D., 2017. Reconstructing prehistoric African population structure and adaptation. Presented at the Society for Molecular Biology and Evolution, 2017, Austin, Texas.

Skoglund, P., Storå, J., Götherström, A., Jakobsson, M., 2013. Accurate sex identification of ancient human remains using DNA shotgun sequencing. J. Archaeol. Sci. 40, 4477–4482. https://doi.org/10.1016/j.jas.2013.07.004

Smith, C.I., Chamberlain, A.T., Riley, M.S., Stringer, C., Collins, M.J., 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. Journal of Human Evolution 45, 203–217. https://doi.org/10.1016/S0047-2484(03)00106-4

Smith, F.H., 1994. Samples, Species, and Speculations in the Study of Modern Human Origins, in: Origins of Anatomically Modern Humans, Interdisciplinary Contributions to Archaeology. Springer, Boston, MA, pp. 227–249. https://doi.org/10.1007/978-1-4899-1507-8_11

Smith, P.E.L., 1990. Architectural Innovation and Experimentation at Ganj Dareh, Iran. World Archaeology 21, 323–335.

Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilova, E., Macaulay, V., Richards, M.B., Cerny, V., Pereira, L., 2012. The expansion of mtDNA haplogroup L3 within and out of Africa. Mol. Biol. Evol. 29, 915–927. https://doi.org/10.1093/molbev/msr245

Soffer, O., Gamble, C.S., 1990. The World at 18,000 BP. Unwin Hyman, London.

Spoor, F., Wood, B., Zonneveld, F., 1994. Implications of early hominid labyrinthine morphology for evolution of human bipedal locomotion. Nature 369, 645–648. https://doi.org/10.1038/369645a0

Straus, L.G., Eriksen, B.V., Erlandson, J.M., Yesner, D.R., 1996. Humans at the end of the Ice Age. Plenum, London.

Street, M., Terberger, T., 1999. The last Pleniglacial and the human settlement of Central Europe: new information from the Rhineland site of Wiesbaden-Igstadt. Antiquity 73, 259–272. https://doi.org/10.1017/S0003598X00088232

Stringer, C., Gamble, C., 1993. Review of In Search of the Neanderthals: Solving the Puzzle of Human Origins.

The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

The MathWorks, Inc., 2016. MATLAB and Bioinformatics Release 2016b. Natick, Massachusetts, United States.

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14, 178–192. https://doi.org/10.1093/bib/bbs017

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., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A., Deloukas, P., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genet. 39, 31–40. https://doi.org/10.1038/ng1946

Torroni, A., Achilli, A., Macaulay, V., Richards, M., Bandelt, H.-J., 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet. 22, 339–345. https://doi.org/10.1016/j.tig.2006.04.001

Torroni, A., Lott, M.T., Cabell, M.F., Chen, Y.S., Lavergne, L., Wallace, D.C., 1994. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am. J. Hum. Genet. 55, 760–776.

Torroni, A., Schurr, T.G., Cabell, M.F., Brown, M.D., Neel, J.V., Larsen, M., Smith, D.G., Vullo, C.M., Wallace, D.C., 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53, 563–590.

Trombetta, B., Cruciani, F., Sellitto, D., Scozzari, R., 2011. A New Topology of the Human Y Chromosome Haplogroup E1b1 (E-P2) Revealed through the Use of Newly Characterized Binary Polymorphisms. PLOS ONE 6, e16073. https://doi.org/10.1371/journal.pone.0016073

Tsuneki, A., 2013. Proto-Neolithic Caves and Neolithisation in the Southern Zagros, in: : : R Matthews and H. Fazeli Nashili (Eds.), The Neolithisation of Iran: The Formation of New Societies. Oxbow Books, Oxford, pp. 94–96.

Udpa, N., Ronen, R., Zhou, D., Liang, J., Stobdan, T., Appenzeller, O., Yin, Y., Du, Y., Guo, L., Cao, R., Wang, Y., Jin, X., Huang, C., Jia, W., Cao, D., Guo, G., Claydon, V.E., Hainsworth, R., Gamboa, J.L., Zibenigus, M., Zenebe, G., Xue, J., Liu, S., Frazer, K.A., Li, Y., Bafna, V., Haddad, G.G., 2014. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. Genome Biol. 15, R36. https://doi.org/10.1186/gb-2014-15-2-r36

van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., Tarekegn, A., Thomas, M.G., Bradman, N., Hellenthal, G., 2015. Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. PLoS Genet 11, e1005397. https://doi.org/10.1371/journal.pgen.1005397

van Zeist, W., Smith, P.E.L., Palfenier-Vegter, R.M., Suwijn, M., Casparie, W.A., 1984. An archaeobotanical study of Ganj Dareh Tepe, Iran. Palaeohistoria 26, 201–224.

Vansina, J., 1995. New Linguistic Evidence and "the Bantu Expansion." The Journal of African History 36, 173–195. https://doi.org/10.2307/182309

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The Sequence of the Human Genome. Science 291, 1304–1351. https://doi.org/10.1126/science.1058040

Vianello, D., Sevini, F., Castellani, G., Lomartire, L., Capri, M., Franceschi, C., 2013. HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. Hum. Mutat. 34, 1189–1194. https://doi.org/10.1002/humu.22356

Wall, J.D., Kim, S.K., 2007. Inconsistencies in Neanderthal Genomic DNA Sequences. PLOS Genetics 3, e175. https://doi.org/10.1371/journal.pgen.0030175

Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Stevison, L.S., Gignoux, C., Woerner, A., Hammer, M.F., Slatkin, M., 2013. Higher Levels of Neanderthal Ancestry in East Asians Than in Europeans. Genetics genetics.112.148213. https://doi.org/10.1534/genetics.112.148213

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., Kayser, M., 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci. Int. Genet. 7, 98–115. https://doi.org/10.1016/j.fsigen.2012.07.005

Weaver, T.D., 2012. Did a discrete event 200,000-100,000 years ago produce modern humans? J. Hum. Evol. 63, 121–126. https://doi.org/10.1016/j.jhevol.2012.04.003

Weeks, L., Alizadeh, K., Niakan, L., Alamdari, K., Zeidi, M., Khosrowzadeh, A., McCall, B., 2006. The Neolithic Settlement of Highland SW Iran: New Evidence from the Mamasani District. Iran 44, 1–31.

White, T.D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G.D., Suwa, G., Clark Howell, F., 2003. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature 423, 742–747. https://doi.org/10.1038/nature01669

Whitlam, J., Ilkhani, H., Bogaard, A., Charles, C., 2013. The plant macrofossil evidence from Seikh-e Abad: First impressions, in: The Earliest Neolithic of Iran: 2008 Excavations at Sheikh-e Abad and Jani (Eds Matthews, R., Matthews, W. & Mohammadifar, Y.). p. Ch. 15, 175-184.

Wood, B., 1992. Origin and evolution of the genus Homo. Nature 355, 783–790. https://doi.org/10.1038/355783a0

Yang, D.Y., Eng, B., Waye, J.S., Dudar, J.C., Saunders, S.R., 1998. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am. J. Phys. Anthropol. 105, 539–543.

Zeder, M.A., 2011. The Origins of Agriculture in the Near East. Curr Anthrop 52, 221–235.

Zeder, M.A., 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. PNAS. https://doi.org/10.1073/pnas.0801317105

Zeder, M.A., Hesse, B., 2000. The Initial Domestication of Goats (Capra hircus) in the Zagros Mountains 10,000 Years Ago. Science 287, 2254–2257. https://doi.org/10.1126/science.287.5461.2254