Insights into the human demographic history of Africa through whole-genome sequence analysis

Gerard Serra Vidal

TESI DOCTORAL UPF / ANY 2018

DIRECTOR DE LA TESI
Dr. David Comas
Departament de Ciències Experimentals i de la Salut

Universitat Pompeu Fabra Barcelona

Instead of separation and division, all distinctions make for a rich diversity to be celebrated for the sake of the unity that underlies them. We are different so that we can know our need of one another.

— Desmond Tutu

Agraïments

Aquesta tesi és un treball col·lectiu. Són moltes les persones a qui he d'agrair l'ajuda i l'acompanyament durant aquests anys, tant a nivell científic o acadèmic com personal.

En primer lloc, gràcies al David, per la disponibilitat, el bon tracte, els ànims, la paciència i per haver estat al darrere de tot.

A la Belén per tot una mica, per guiar-me durant els primers passos, ensenyar-me com funciona el cluster, pel Bash, per acostumar-me al rigor en la recerca, i per totes les discussions, anades i vingudes sobre els africans, juntament amb l'Oscar i el Gabriel, de qui he après moltes coses.

A tots els de la 412.08 i els de la setena, especialment a la Lara per anar sempre per davant i ser una gran companya d'escriptori, a l'Àlex per les millors idees, a la Neus per l'etern bon humor, al Simone, l'Andre, la Neus i la Carla per les converses de feina i de no feina, a l'Erica pel suport computacional i al Marcel per l'ajuda mútua. I sobretot a tots per ser part d'un grup on he estat a gust i he crescut com a (espero) científic i com a (espero) persona durant aquests quatre-cinc-sis anys.

A la Rosa i al Pau pel dinars a la terrassa que sempre he acabat més animat de com començava.

Al Carles per ser un gran saltamartí.

A la Carmina per ser-hi.

Al Bernat per obrir camí. A la Maria del Mar i la Neus per ser les meves germanes preferides.

Als meus pares per donar-me el suport econòmic i moral.

I a tots els qui, de manera directa o indirecta, m'han ajudat a arribar fins aquí. Gràcies a tots.

Abstract

Africa is the source of worldwide human populations, and despite harboring the highest levels of genetic diversity, its complex demographic history has remained understudied for a long time, with several relevant questions remaining open, like the archaic introgression scenario or the continuity vs replacement debate in North Africa. To address these questions, we analyzed complete genome sequences from a geographically, linguistically, and culturally diverse panel of African individuals. We characterized the deepest human lineages, corresponding to hunter-gatherer groups, and the demographic processes that have shaped the current genetic map, including intra-continental admixture and backflow from Eurasia to northern and eastern Africa and archaic introgression in western, central and southern sub-Saharan populations. North Africa shows as an independent genetic history from the rest of the continent, with five different ancestries, including sub-Saharan, European, Middle Eastern, Caucasian and a heterogeneously persistent North African Epipalaeolithic component.

Resum

Àfrica és l'origen de totes les poblacions humanes actuals, i malgrat contenir els nivells més elevats de diversitat genètica, la seva història demogràfica ha estat relativament poc estudiada, amb moltes preguntes rellevants encara obertes, com ara els escenaris d'introgressió arcaica en poblacions modernes o el debat entre continuitat o reemplaçament en les poblacions del Nord d'Àfrica. Per tal de respondre aquestes qüestions, hem analitzat genomes complets d'un panell d'individus africans divers a nivell geogràfic, ligüístic i cultural. Hem caracteritzat els llinatges humans més divergents, corresponents a grups de caçadors-recol·lectors, així com els processos demogràfics que han donat forma al mapa genètic actual, incloent-hi flux gènic intracontinental i d'Euràsia cap al nord i est d'Àfrica, i introgressió arcaica en poblacions sub-saharianes de l'oest, centre i sud del continent. El Nord d'Àfrica té una història genètica independent de la de la resta del continent, i s'hi han identificat cinc components ancestrals: subsaharià, europeu, de l'Orient Mitjà, caucasià i un component heterogeni autòcton nordafricà persistent des de l'epipaleolític.

Preface

The origins, evolution and demographic history of human populations can be approached from several areas of knowledge, such as archaeology, palaeoanthropology, or linguistics. Genetic studies are based on the analysis of genetic differences among human individuals and populations due to demographic events such as divergence, migration, or admixture, and represent a complementary perspective which is able to provide a quantifiable and statistically powerful contribution to the study of populations origins and evolution.

Technological advances in the last decade, as well as extensive sampling of current and ancient populations, have caused a fast increase of human population genetics studies. Particularly, the availability of complete genome sequences represents a milestone in genetics studies, since for the first time a comprehensive view of the genetic diversity of populations is provided, together with the possibility of applying complex and informative methodologies.

The African continent is the original homeland of modern humankind, which implies that all human genetic diversity observed outside Africa is essentially a subset of the African diversity. However, despite its key role in understanding current variation, African populations have been classically underrepresented in human population genetics studies.

This thesis is focused on the study of human populations in Africa using whole-genome sequences and novel associated methods. Particular attention is paid to the first human lineages that diverge from the rest of populations, which harbor the highest levels of intra-population diversity, as well as to North Africa, whose human population history is essentially different from the rest of the continent due to its geographical location, isolated by the Sahara Desert and close to Europe and the Middle East.

Contents

Chapter 1

Introduction

1 Human population genetics – an overview

The collecting and analysis of genetic data has dramatically changed throughout time, in parallel to technological advances. But genetics was applied much before the discovery of genes or DNA. From prehistorical times, artificial selection has been carried out in order to improve plants and animals through selective breeding (i.e., differential reproduction), so that only plants and animals with desirable feature reproduce and as a consequence the selected characteristics become more prevalent in the population across generations.

The first formal genetic studies were performed by Gregor Mendel (Mendel, 1866), who set the inheritance laws. It was not until the 20^{th} century that the Mendelian model was widely accepted with the *Drosophila melanogaster* studies by Thomas Hunt Morgan, who set that inheritance units, named genes, occupy a specific location in chromosomes, which become the basis for genetic inheritance.

The discovery in 1944 of DNA as the carrier of genetic information was another milestone for genetic research (Avery et al., 2000), which together with the description in 1953 of the double helical structure of the DNA molecule (Watson and Crick, 1953) and the breakthrough of molecular biology from the second half of the 20^{th} century, allowed the development of molecular genetics and the determination of the sequence of a gene for the first time in 1972 (Jou et al., 1972).

Molecular genetics, i.e., the study of genetics at DNA level, has been combined with the Mendelian concepts of heredity, the Darwinian theory of evolution through natural selection, the adoption of mutation as the ultimate source for genetic variation, and the neutral theory of molecular evolution (Kimura, 1983), which states that most evolutionary changes are due to random fixation of neutral variants through random sampling drift, leading genetics research to a new paradigm: modern evolutionary synthesis.

In this framework, the first molecular population genetic studies were based on the study of particular individual loci, also called classical markers. The first used markers were protein markers including blood groups (such as the ABO, Rh and other blood group systems) and protein allomorphs, such as red cell enzymes, serum proteins and HLA antigens such as the major histocompatibility complex (MHC), albumin, acid erytrocytary phosphatase (ACP), adenosine deaminase (ADA), superoxide dismutase (SOD), transferrins, immunoglobulin allotypes (IG), or lipoprotein types (LP), among many others (see for example Stone et al. (1993); Vona et al. (2003); Muehlenbein (2010)). These markers were highly informative due to its highly polymorphic nature and have been broadly used in population genetic studies (e.g., Bosch et al. (1997); Harich et al. (2002); Chbel et al. (2003); González-Pérez et al. (2003); Gonzalez-Perez et al. (2007); El Moncer et al. (2010)), but they are unable to give a global view since they only cover specific genomic loci, which is likely to result in a biased picture and might be affected by selection (thus not accounting for population history but for a specific selective pressure).

Another kind of molecular genetic markers are the ones based on polymorphic DNA sequences. They can be short DNA sequences surrounding a single nucleotide polymorphism (SNP), or long ones, such as minisatellites or short tandem repeats (STRs) or *Alu* insertions, whose variation allows the identification of individuals, populations or species based on genetic similarity. In particular, *Alu* polymorphisms have played a key role in human population genetics due to their neutral nature and their known ancestral states and identical descent (Batzer and Deininger, 2002).

Uniparental markers are a special type of genetic markers. They include Y chromosomal DNA or mitochondrial DNA, which are transmitted from fathers or mothers to offspring, respectively, which makes them suitable to be studied for assessing paternal or maternal lin-

eages (Underhill and Kivisild, 2007). Uniparental genetic history is relatively straightforward to retrace due to the lack of recombination which provides a more direct interpretation of the demographic history in comparison with autosomal data, but the information it provides is limited to the male- or female-led history (Cummins, 2001). Sex-biased patterns of population evolution and asymmetry between male- and female-driven demographic events are often detected when studying uniparental markers. Indeed, the global distribution of Y chromosome haplogroups has been found to be generally more correlated with geography in comparison with mitochondrial DNA (Wood et al., 2005; Lao et al., 2008).

Mitochondrial DNA (mtDNA) markers are the best studied among all systems of genetic markers. This is mainly due to its maternal-specific transmission, the fact that it is present in both men and women, which makes sample collection easier than in the case of the Y chromosome; its small size, which makes its sequencing process easier; and its high mutation rate, which makes it very variable and makes it suitable for genetic population studies (Litvinov and Khusnutdinova, 2015).

On the other hand, the Y chromosome is transmitted from fathers to sons only and it contains a large non-recombining block (NRY) which makes it convenient for population genetic studies through the parental line (Litvinov and Khusnutdinova, 2015). The Y chromosome is sensitive to the drift effect, especially to the founder effect (Underhill et al., 2001), and the diversity of its sequences is the smallest in the nuclear genome (Jobling and Tyler-Smith, 2003).

In order to obtain new information about the genetic structure of human populations that classical and uniparental markers are not powerful enough to provide, human population genetics has been refocused from the study of individual loci to the genomic level (Litvinov and Khusnutdinova, 2015). This represented a big step forward since the number of polymorphisms typed increased dramatically from a a few hundreds to hundreds of thousands and up to a million SNPs in current DNA microarrays.

This kind of analysis of autosomal SNPs is different from the analyses of uniparental markers in several aspects: *(i)* an account for recombination is needed, *(ii)* their inheritance does not happen through one of the parents only, *(iii)* the high number of SNPs requires new computational tools and analytic methods, but allows the obtention

of much more complex and detailed demographic histories.

Following to microarrays, sequencing technologies have been developed. DNA sequencing is the process of determining the order of nucleotides of a DNA molecule. This has many applications, such as exome sequencing or targeted resequencing, which are used to analyze specific fractions of the genome, or whole-genome sequencing, which delivers a comprehensive view of the complete genome.

One of the main advantages of whole-genome sequencing over microarrays lies in the fact that sequencing is not affected by SNP ascertainment bias, unlike genotyping microarrays. The pre-ascertained set of SNPs contained in arrays might be biased depending on the populations of study. Commonly used arrays (such as Affymetrix 6.0) tend to be biased towards European populations, implying that many variants which are not present in these populations are not analyzed and, on the contrary, polymorphism present in Europeans is overrepresented, which could produce biases in the inferences and data analyses. It must be pointed out that efforts have been made in order to minimize this bias in population genetics studies, particularly with the design of the Human Origins Array (Patterson et al., 2012). This microarray contains 11 different SNP datasets, each one ascertained on the basis of being heterozygous in a single genome sequence from each of the 11 different populations (Pugach and Stoneking, 2015).

Other SNP array limitation lie in the fact that allele frequency distributions of typed SNPs might be artifactly shifted towards intermediate frequency alleles, estimates of linkage disequilibrium might not reflect real values, and ascertained SNPs tend to be present in many populations, which make them unfit to discover new variants or the variation in uncommon populations. Due to all these factors, results obtained by the analysis of SNP array data and whole-genome sequencing data can lead to significantly different conclusions when studying demographic history or natural selection in populations (Lachance and Tishkoff, 2013).

In addition, in comparison to DNA microarrays, whole-genome sequencing provides information that is orders of magnitude larger in terms of detail and exhaustivity. While DNA arrays provide genotypes for up to one million SNPs, full genome sequences account for the complete genome, which in the case of humans is 6 billion positions long, i.e., 3000 times more (since each position has two alle-

les). While DNA microarray genotyping types particular, previously determined positions of the genome, whole-genome sequencing determines the order of all the nucleotides in a genome, providing a high-resolution, base-by-base view of the genome, and can determine variation in any part of the DNA sequence, allowing the discovery of previously unknown SNPs and, unlike microarray genotyping, not only limiting the analyses to the a priori typed SNPs, which might miss a considerable amount of polymorphism. Moreover, whole-genome sequencing can provide information not only for SNPs but also other kind of genetic variation, such as indels, inversions, translocations, copy number variants, or large structural variants, which might otherwise be missed. In addition, sequencing technology is ideal for discovery applications, like identifying causative variants or *de novo* genome assembly.[]

Sanger sequencing was the first developed sequencing technique, although it has been gradually replaced by next-generation sequencing (NGS) (also known as shotgun sequencing), and has remained as a method for validation of NGS results and for obtaining especially long contiguous DNA sequence reads. Both techniques are similar, but NGS has a much lower cost than Sanger sequencing due to its high-throughput capacity degree, which enables a much higher degree of parallelization and much smaller reaction volumes (Ari and Arikan, 2016). In NGS, millions of DNA fragments are sequenced in a single run, while Sanger sequencing produces only one forward and reverse read. By contrast, Sanger sequencing produces long reads and has 99.999% accuracy (Shendure and Ji, 2008), whilst NGS error rate can be up to 100 times higher (error rate = 0.1% for Illumina sequencing, one of the most used NGS technologies) (Shendure and Ji, 2008; Fox et al., 2014) and delivers much shorter reads (100 - 300 base pairs).

Another key factor that explains the success of NGS and its extended usage is the evolution of its cost across time, together with technical evolution and improvement. DNA sequencing cost decrease throughout time is often represented together with Moore's law (Moore, 2006), which states that computational power doubles every two years. This comparison shows that the sequencing technology is improving at a much faster pace than computers, which has enabled an explonential decrease in its cost. Indeed, the cost of sequencing an individual human genome is around $1,000 in 2017, while it used to be $10,000,000 ten years ago (Figure 1).

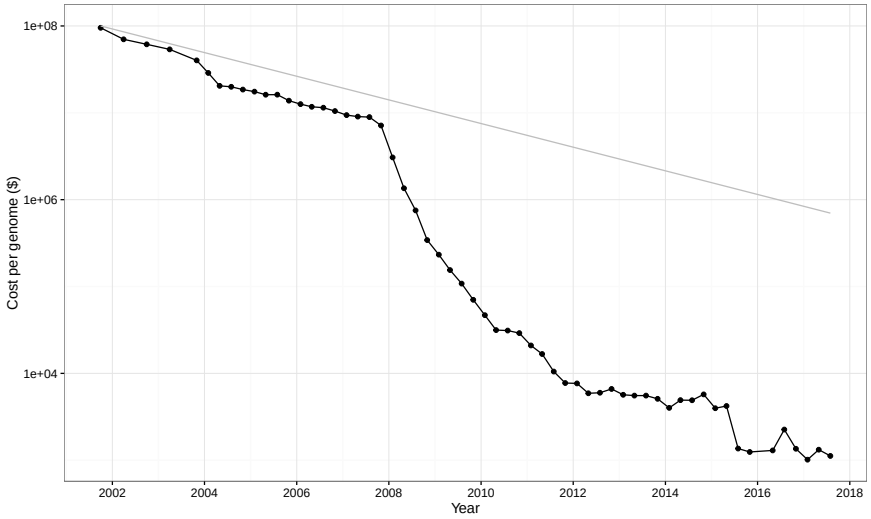

Figure 1: Evolution of the cost per genome throughout time (black dotted line). The decreasing gray line shows how the sequencing cost would decrease if sequencing technology evolution followed Moore's Law. Data from https://www.genome.gov/27541954/dna-sequencing-costs-data/.

Many computational challenges derive from the advent of NGS, due to the vast amount of data generated, whose processing and analysis has required the development of new standard computational pipelines and efficient methods able to deal with such large amounts of data. DNA sequencers produce raw reads that need to be reassembled into the complete genome sequence using a reference genome. This process is known as mapping and requires high computing power to find the most likely location on the reference genome where each read belongs. Then, variations from the reference genome are discovered during the variant calling phase and are posteriorly filtered and annotated (Roy et al., 2018). The final polymorphism data can be further downstream analyzed using population genetics tools.

Third-generation sequencing (also known as long-read sequencing) is the set of recently developed, and still on development, sequencing techniques with a different basis from NGS and consisting on reading the DNA sequence of a single molecule level instead of breaking the DNA into small fragments or reads and amplification by PCR. This technology produces long reads which can be mapped much more precisely, allowing a much higher precision and the detection of long structure variants. Examples of these technologies are nanopore

sequencing (developed by Oxford Nanopore Technology) and single molecule real time sequencing (developed by Pacific Biosciences).

Table 1 shows a summary of the different types of genetic data that have been used to study demographic history along with their advantages and disadvantages.

Data type	Uses and advantages	Limitations
Autosomal STRs	• Hundreds of independent STRs can be genotyped in many individuals, which reduces the effect of evolutionary stochasticity • Their high mutation rate is useful for inferring recent demographic events and for distinguishing between closely related populations	• Limited inference of demographic events at deep timescales • High uncertainty in mutation rates and mutation model
mtDNA	• The absence of recombination allows reconstruction of a gene tree • Smaller N_e than autosomal DNA, which allows better discrimination between populations • Samples from many thousands of individuals can be characterized at low cost • High copy number makes it amenable for ancient DNA extraction and analyses	• A single genealogy contains little information about the underlying population history • Likely to be subjected to the effects of natural selection • High uncertainty in mutation rates

NRY	• The absence of recombination allows reconstruction of a gene tree • Smaller N_e than autosomes, which allows better discrimination between populations • Samples from many thousands of individuals can be characterized at low cost	• A single genealogy contains little information about the underlying population history • Likely to be subjected to the effects of natural selection • High uncertainty in mutation rates and mutation model • Ascertainment bias results from the genotyping of specific SNPs
SNP microarrays	• Hundreds of thousands of SNPs can be genotyped in a single experiment • Unprecedented resolution of population structure	• Large ascertainment bias results from the haphazard way by which SNPs were discovered • Less powerful for making inferences in populations that are diverged from those in which SNPs were discovered
Second generation sequencing	• Massive amounts of relatively unbiased sequence data can be obtained from targeted regions or entire genomes compared with Sanger sequencing • High throughput and does not require a targeted pre-PCR step, which allows sequencing of ancient DNA • Lowest per-base cost of any current sequencing methodology	• Relatively error prone compared with Sanger sequencing • Biases may arise with regard to regions that are preferentially sequenced • Sequencing is through short reads (100-150 bp), which restricts the use of methods that require haplotype-phased data

Third generation sequencing	• Can generate long sequence reads (>10 kb) • Some methods can sequence DNA from single cells, which is particularly useful for very ancient samples • Long reads may also allow *de novo* assembly and thus reduce reference biases	• Per-base cost is currently more expensive than second-generation sequencing • Bioinformatic tools have not yet been fully developed to cope with the increased read length

Table 1: Types of genetic data that are used to infer population history (from Veeramah and Hammer (2014)). N_e, effective population size; PCR, polymerase chain reaction.

The Human Genome Project (Lander et al., 2001; Venter et al., 2001) was a milestone in human genetic research and consisted of two independent projects, with public and private funding and developed using Sanger and shotgun technologies, respectively.

The rise of the study of human genetic variation began soon after the publication of the human genome sequence in 2001, with projects such as the Human Genome Diversity Project (Rosenberg et al., 2002; Cavalli-Sforza, 2005), whose aim was to document the genetic variation of the human species, and the HapMap Project (Belmont et al., 2003, 2005; Frazer et al., 2007; Altshuler et al., 2010), whose goal was to build a haplotype map of the human genome.

The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010, 2012, 2015) represents one of the most important efforts in the field of population genetics so far. Over its three phases, 2504 individual complete genomes from worldwide populations have been sequenced and 84.4 million SNPs have been described, providing a detailed catalog of human genetic diversity. However, this project presents two main weaknesses. On the one hand, genomes were sequenced at low coverage (4-fold). This implies that a significant proportion of the variation on the genome might be overlooked, since many loci, particularly heterozygous positions, cannot be identified as polymorphic with such a low coverage. On the other hand, the 1000 Genomes Project provides an incomplete picture of the global human genetic diversity, since it omits certain geographic areas, such as Southern Africa, North Africa, the Middle East, Eastern Europe,

Siberia, Southern South America, or Oceania.

In recent years, additional studies have been performed such as the Simons Genome Diversity Project (Mallick et al., 2016), which provided 279 individual complete genomes from 130 populations, including indigenous populations from all continents and covering a wider geographical range than the 1000 Genomes Project and with a high coverage of 43-fold on average, which makes it a powerful dataset to detect rare variants and to have a more exhaustive picture of global human diversity.

The advent of NGS and advanced sequencing techniques has also boosted the analysis of ancient genomes beyond mtDNA and some classically used markers (Veeramah and Hammer, 2014; Der Sarkissian et al., 2015). The sequencing of the first ancient human genome took place in 2010 by Rasmussen et al. (2010) and many ancient human genomes have been sequenced and analized since. Ancient hominins such as Neanderthals and Denisovans have been also been sequenced (Green et al., 2010; Prüfer et al., 2014; Reich et al., 2010; Meyer et al., 2012). Ancient genomes provide a piece of great value in population genetics and especially in the study of demographic dynamics, helping researchers to track genetic variant frequency changes across space and time, and filling gaps that would be difficult to fully understand without genetic data from archaic individuals (Marciniak and Perry, 2017).

2 Genetic history of Africa

The reconstruction of the African demographic history is a key point in the study of human population genetics due to its central role in modern human origins.

Palaeoanthropological and archaeological studies support an African origin of anatomically modern humans (AMHs) that took place 200,000 - 300,000 years ago (Cann et al., 1987; Stringer, 2000; McDougall et al., 2005; Jakobsson et al., 2008; Nielsen et al., 2017). Indeed, many fossil remains with a combination of ancient and modern traits have been found all over Africa, especially in East Africa, where the 195,000-year-old fossils in Omo Kibish (Ethiopia) (McDougall et al., 2005; Liu

et al., 2006; Brown et al., 2012), the 160,000-year-old Herto skull from Middle Awash (Ethiopia) (White et al., 2003), and the 120,000-year-old remains from Ngaloba (Tanzania) (Day et al., 1980; Grove et al., 2015) are commonly regarded as the first early anatomically modern humans; as well as in South Africa, with remains such as the Florisbad skull in South Africa from about 260,000 years ago (Huphrey, 2003) and the 250,000-year-old *Homo naledi* fossils in South Africa (Berger et al., 2015; Dirks et al., 2017), which are considered an older form of *Homo sapiens* and an extinct old branch sister to the current *Homo sapiens* ancestors, respectively. More recently, early *Homo sapiens* skeletal remains found in Jebel Irhoud (Morocco) were dated to 315,000 years old (Hublin et al., 2018; Richter et al., 2017), not only pushing back the origins of modern humans but also adding northwest Africa to the list of putative origin sites and pointing to a possible pan-African origin of *Homo sapiens*. All these findings are shown in Figure 2.

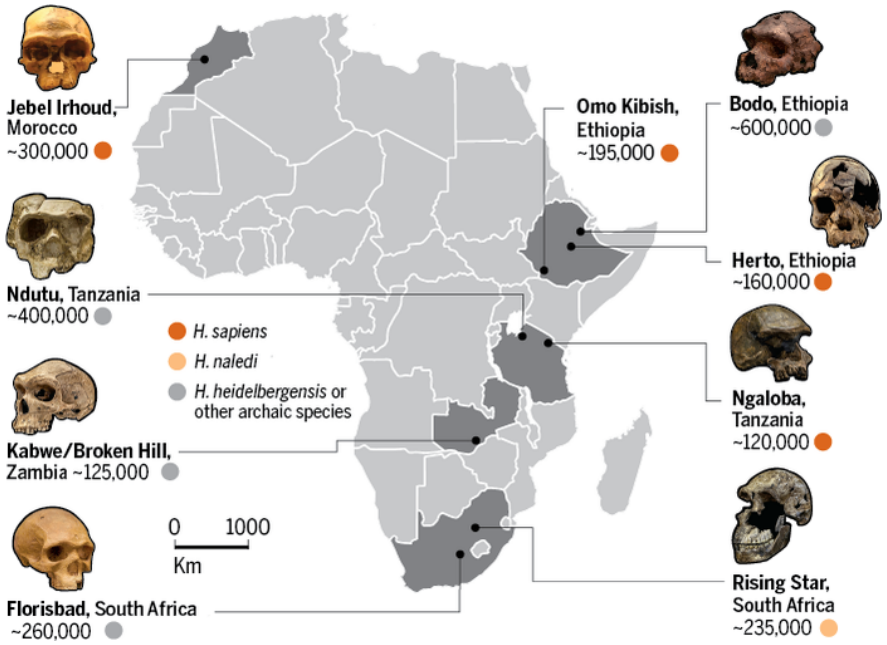

Figure 2: African sites with archaic human remains (Gibbons, 2017).

In the last decade, other human species, such as Neanderthals and Denisovans, have been identified out of the African continent, dated as far back as 400,000 ya (Nielsen et al., 2017), and whose divergence time with the ancestors of modern humans dates back to more than 500,000 years ago (Meyer et al., 2016), which could imply that previous lineages of *Homo sapiens* could have existed much before 200,000 years

ago, although no clear fossil evidence has been discovered to suport this hypothesis (Stringer and Galway-Witham, 2017).

Debate exists about whether the morphological diversity observed in ancient human fossils indicates that they belong to different human species or, on the contrary, the human species used to be more diverse than it currently is. The morphological diversity in skulls shows that different features of modern humans arose in different locations at different times, pointing to multiple African origins of AMHs, whose modern features would have evolved in a fragmented manner in several areas connected by gene flow (Nielsen et al., 2017). This theory, known as African multiregionalism, states that no single place or population gave rise to current populations, but the whole African continent has to be considered as the modern human cradle, i.e., humans originated from different populations that lived across Africa and were separated from each other by geographical barriers (Scerri et al., 2018). These barriers were not static but dynamic, changing in parallel to climate changes that reshaped the African landscape. As deserts and forests expanded or narrowed, human groups were drawn together or pulled apart, which facilitated convergent population evolution and a common melting pot among nowadays geographically, ecologically, and climatically divergent areas.

In parallel to palaeontological evidence, genomics has also shown Africa as the cradle of modern humans. Ancient genomic data has been used to estimate very early demographic events. A 2,000-year-old South African genome, as well as four other genomes from the Iron Age published by Schlebusch et al. (2017), were used to date modern human emergence between 350,000 and 260,000 years ago through the "Two plus Two" method, based on estimation of model parameters under a pure split model using single individual samples. (Schlebusch et al., 2017).

Many genomic studies have shown that African populations harbor the highest levels of genetic diversity in the world (Tishkoff et al., 2009; Henn et al., 2011; Lachance et al., 2012; Pemberton et al., 2012; Henn et al., 2016), which agrees with an African origins of modern humans. Indeed, genetic diversity observed in Eurasia, Oceania, and America is subset of the African variation (Ramachandran et al., 2005; Schlebusch et al., 2012), with small contributions of non-African archaic hominins such as Neanderthals or Denisovans (Green et al., 2010; Meyer et al., 2012).

This is consistent with linguistic diversity in Africa, which is higher than anywhere else in the world. One in three languages (i.e., more than 2,000) is spoken in Africa, which mostly belong to four major families: Niger-Kordofanian, including Bantu languages and present in western, central, southeast and southern Africa; Afroasiatic, spoken in northern and eastern Africa; Nilo-Saharan, spoken in areas of Saharan, eastern, and northeastern Africa; and Khoisan, a family of languages defined by click consonants spoken by the San in southern Africa and the Hadza and Sandawe in eastern Africa (see Figure 3). Social and cultural diversity are also very high and include all types of lifestyles, including dietary and subsistence methods, such as agriculturalists, pastoralists, and hunter-gatherers.

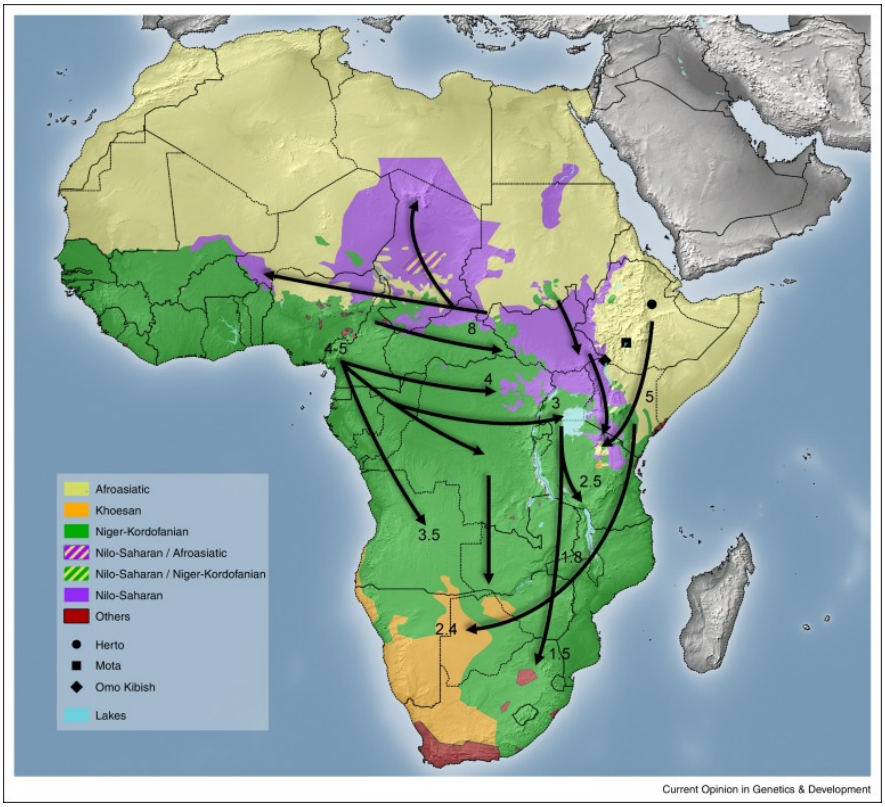

Figure 3: Map of Africa showing the distribution of the major language families, the location of hominid remains discussed in (Beltrame et al., 2016), and major migration routes of AMH through the continent within the past 10,000 years.

Uniparental markers analysis also points to Africa as the cradle of modern humans (Underhill and Kivisild, 2007). The most ancestral

Y chromosome haplogroup has been dated to 338,000 ya (Mendez et al., 2013) whilst the time to the most recent ancestor (TMRCA) for mtDNA has been estimated to 140,000-240,000 ya (Schuster et al., 2010; Behar et al., 2012), both in sub-Saharan Africa.

Analysis of mtDNA, Y chromosome, and autosomal diversity point to the Khoisan and Pygmies (hunter-gatherers from the rainforests in the Congo Basin in Central Africa) as the most divergent extant populations compared to other African groups speaking Niger-Kordofanian, Nilo-Saharan and Afroasiatic languages (Wood et al., 2005; Tishkoff et al., 2009; Gronau et al., 2011; Pickrell and Pritchard, 2012). This model is represented in Figure 4.

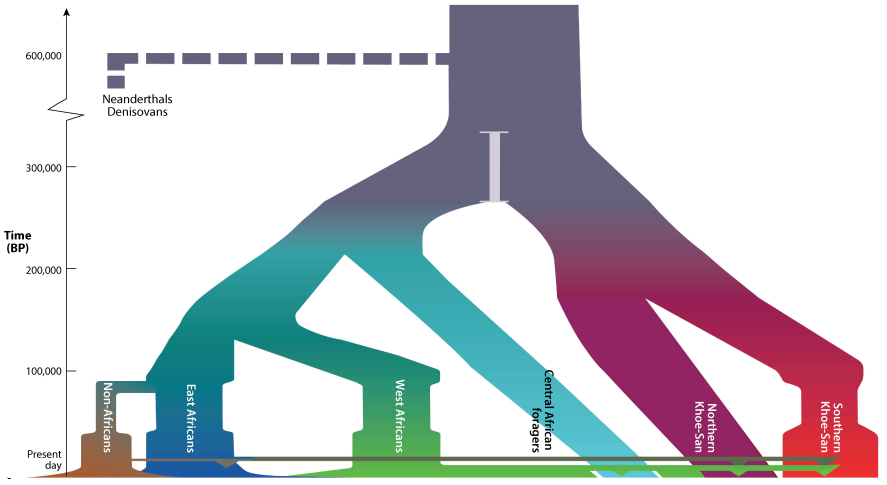

Figure 4: Population split times, hierarchy, and population sizes (width along a horizontal axis for populations). Horizontal colored lines represent migration, with down-pointing triangles representing admixture into another group (Schlebusch et al., 2017).

However, real African demographic history appears to be much more complex than a model with splits and migrations. Indeed, the study of uniparental markers suggests that forager populations related to Khoisan might have been widespread in eastern Africa and posteriorly replaced outside southern Africa, or might have originated in eastern Africa and migrated to the south of the continent around 50,000 ya (Tishkoff et al., 2007, 2009; Skoglund et al., 2017), which would agree with the presence of click-language-speaking populations in Tanzania, who despite sharing a similar language share little ancestry with Khoisan (Schlebusch et al., 2012).

Among all migrations in Africa during the last millennia, one particular demographic event has had a very high impact in the current African genetic landscape. The so-called Bantu expansion was carried out by these groups, initially living in West Africa, who early adopted agriculture and who spread across the continent, erasing a substantial portion of the genetic footprint of other African populations. This migration started around 4,000-5,000 ya from the Grassfields region in the borderland between current-day Nigeria and Cameroon, going from west to east, to the Great Lakes of Uganda by around 3,000 ya and then from east to south in the last 2,500 years, rapidly expanding into central and southern Africa, reaching Mozambique $\sim 1,800$ ya and South Africa $\sim 1,500$ ya (Beltrame et al., 2016).

Two different routes from the Grassfields region to the Great Lakes have been identified: north and south of the nowadays Gabon rainforest. The fact that current Bantu populations from eastern and southern Africa are genetically more similar to populations based south of the rainforest than those to the north suggests that the Bantu crossed the rainforest before splitting into two groups which later followed migratory routes towards eastern and southern Africa, where they came into contact with autochtonous populations of these regions.

Other dispersion waves have been identified, such as the one occurring $\sim 3,500$ ya from Cameroon to Angola, the spread of pastoralism into sub-Saharan Africa around 4,500 ya, the Afroasiatic populations migration from Ethiopia into Kenya and Tanzania within the past 5,000 years and, after admixing with Bantu groups, their spread to southern Africa around 2,400 ya (Patin et al., 2009).

These migrations are represented in Figure 3 together with the African linguistic families and the areas where they are spoken.

As previously stated, ancient genomes can be extremely helpful when studying the demographic history of populations. Nevertheless, the vast majority of ancient DNA samples come from individuals from Eurasia while Africa has a minority representation (see Figure 5), mainly due to its hot and humid weather, which does not favor the conservation of ancient remains and particularly DNA, whose deterioration increases with temperature. The first African ancient genome was sequenced in 2015 by Gallego Llorente et al. (2015) and until last year only 16 human genomes had been sequenced. All published ancient genomes from Africa are shown in Table 2.

Study	*n*	Age (ya)	Location
Gallego Llorente et al. (2015)	1	4500	Ethiopia
Schuenemann et al. (2017)	3	4300-2000	Egypt
Schlebusch et al. (2017)	7	2000-300	South Africa
Rodríguez-Varela et al. (2017)	5	1400-1000	Canary Islands
Skoglund et al. (2017)	16	8000-300	Kenya, Tanzania, Malawi, South Africa
Van De Loosdrecht et al. (2018)	7	15000	Morocco
Fregel et al. (2018)	15	7000-5000	Morocco

Table 2: African ancient DNA datasets published to date. ya: years ago.

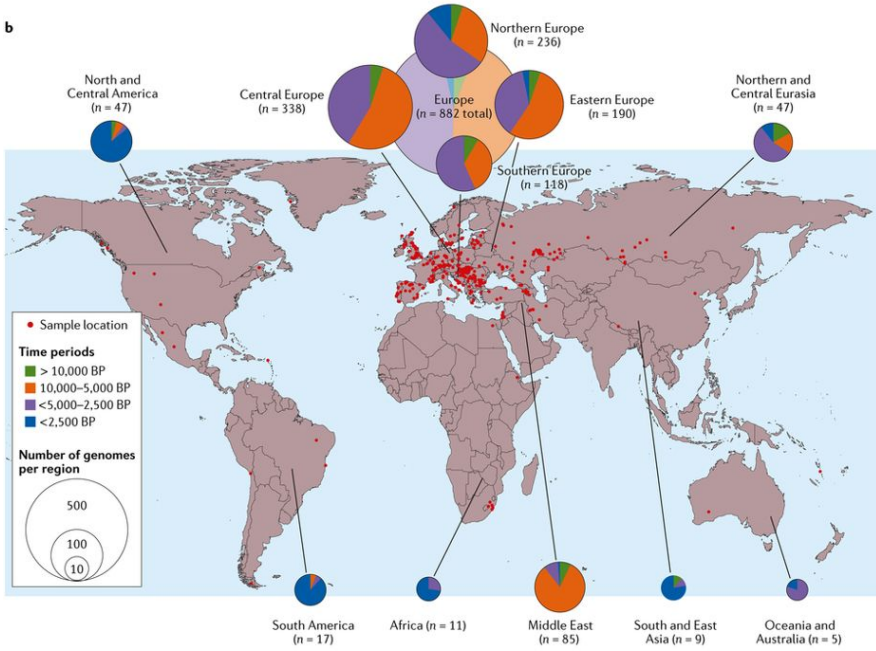

Figure 5: Ancient human and archaic hominin genome worldwide datasets (Marciniak and Perry, 2017). It does not account for the last four African datasets in table 2

The study of ancient DNA in Africa has added valuable information regarding to continental migrations in the last 5,000 years. One of the most relevant findings was a back-to-Africa migration around 3,000 years ago from the West Eurasia to East Africa that was inferred from the analysis of a 4,500-year-old genome from Mota Cave (Ethiopia) (Gallego Llorente et al., 2015), who also proved population continuity over the last 4,500 years in this region.

These findings are in agreement with Pickrell et al. (2013)), who besides identifying the same back migration to East Africa (3,000 years

ago) found signals of such backflow in South Africa (900-1,800 years ago), thanks to fossil remains from a 3,000-year-old individual from Tanzania, which carried not only East African ancestry but also from Middle Eastern Neolithic (Skoglund et al., 2017). Current pastoralists in southern Africa share genetic signatures with this ancient Tanzanian sample, pointing to a possible movement of pastoralism from East Africa to South Africa.

Insights into the Bantu expansion have also been revealed through the analysis of African ancient genomes. A 2,000-year-old individual from South Africa showed relatedness to the present-day Khoisan populations, as well as to ancient hunter-gatherers from Malawi and Tanzania from 8,100-2,500 and 1,400 years ago, respectively (Skoglund et al., 2017), but not to current East African populations. This could be explained by a migration led by West African Bantu populations to East and South Africa around 2,000 years ago that replaced local hunter-gatherers, since no hunter-gatherer ancestry from southern or eastern Africa was found in another ancient sample from Tanzania from 750 years ago. Schlebusch et al. (2017) found similar evidence in an 2,000-year-old genome from South Africa.

In addition to admixture processes within African populations and between Africans and non-Africans, another kind of events must be considered when studying population history: archaic introgression coming from other hominin species that lived in the past and persisted until 30,000-40,000 ya (Higham et al., 2014).

Introgression from Neanderthals and Denisovans into the ancestors of current populations has been widely documented in North Africa, the Middle East, Europe, and parts of Asia and Oceania (Mallick et al., 2016). A rich archaeological, palaeontological and genetic record exists for these two hominin species, but this is not the case with other ancient hominins that might have inhabited other areas.

Despite the fact that numerous archaic hominin lineages are known to have existed in Africa (Wolf and Akey, 2018; Bräuer, 2008) and might have lived in the same time and space with modern humans (Dirks et al., 2017), studies of archaic admixture in African populations have been limited compared to other areas, mainly due to the underrepresentation of African genetic data in large genomic datasets and the scarcity of ancient DNA samples (see above).

Several studies, however, have made an effort to investigate the likelihood of archaic admixture in African populations, using indirect methods (i.e., in the absence of any recovered ancient DNA sample) of two types – linkage-disequilibrium-based methods such as the S^* statistics (Lachance et al., 2012; Vernot and Akey, 2014; Hsieh et al., 2016) and demographic-model-based methods like approximate bayesian computation (ABC) approaches (Bertorelle et al., 2010) or hidden Markov models (HMM) approaches (Skov et al., 2018) – in order to detect signals of archaic introgression coming from from ghost populations, i.e., populations that might have existed in the past and whose genetic traces can be found in current or ancient samples but are not unknown since no supporting genetic nor fossil record has been found (Racimo et al., 2015).

Evidence from these studies indicates archaic introgression signals in several African hunter-gatherer populations such as Khoisan (Hammer et al., 2011), Hadza and Sandawe (Lachance et al., 2012), Baka (Lachance et al., 2012; Hsieh et al., 2016), Biaka (Hammer et al., 2011; Hsieh et al., 2016), and Mbuti (Hammer et al., 2011), as well as in non-hunter-gatherers, such as Yoruba (Plagnol and Wall, 2005). Adaptative introgression has been identified at locus level, such as the salivary *MUC7* locus (Xu et al., 2017). Ancient genomes have also been used in order to identify putative ghost populations. An extinct lineage contributing to the genetic pool of some current western Africans has been described by Skoglund et al. (2017) (Mende) and Durvasula and Sankararaman (2018) (Yoruba) separately. However, it should be taken into account that despite the evidence of unknown archaic introgression in sub-Saharan Africa, no exhaustive analyses have been done comparing the putative ghost populations with other known extinct hominins such as Neanderthals and Denisovans.

2.1 Out of Africa

While the African origin of modern humans is widely accepted, several hypothesis exist regarding the expansion of the ancestors of current populations around the world. Several out-of-Africa migrations have been described through time, as early as 220,000 years ago via the Sinai Peninsula (Hershkovitz et al., 2018), 120,000 years ago via the Arabian Peninsula (Bae et al., 2017), 110,000 years ago via North Africa (Balter, 2011), and 70,000 - 50,000 years ago via the Middle East

(Stringer, 2000; Mellars, 2006; Liu et al., 2015; Clarkson et al., 2017).

Several hypotheses have been proposed. According to Bae et al. (2017), the classical "out of Africa" model consisting of a single dispersal wave from Africa into Eurasia around 60,000 years ago needs to be revisited since it cannot explain all the genetic diversity observed out of Africa. For example, modern human fossils have been found at multiple locations in central and southern China dated between 70,000 and 120,000 years ago (Bae et al., 2017). Other findings show AMHs reached Southeast Asia and Australia prior to 60,000 years ago (Westaway et al., 2017). Another study performed by Pagani et al. (2015) claims that *Homo sapiens* left Africa in at least two waves, since at least 2% of the genomes of Papua New Guineans comes from an early human dispersal, who would have left Africa around 120,000 years ago. However, other recent studies state that all previous dispersal waves may have been overridden by the later dispersal wave occurring approximately 60,000 years ago (Mallick et al., 2016) into the Middle East and ultimately to Asia, Australia, and Europe. According to this point of view, all present-day non-African populations branched off from a single ancestral population that went out of Africa. This would explain why the coalescence time of all non-African current populations is around 60,000 ya.

While the first hypothesis states that despite recent dispersal contributing the genetic make-up of present-day non-Africans, the earlier dispersals are still evident; the second one affirms that current non-African populations are descended from a single founding population almost completely and an earlier migration can essentially be ruled out since its genetic footprints have been erased in current-day populations.

2.2 Genetic history of North Africa

The demographic history of North Africa is to a large extent independent from the rest of the continent, mainly because of its geographic isolation due to the presence of the Sahara Desert, which traces its southern limit. The Sahara has been a biological corridor during humid periods, the latest of which, occurred during the Holocene climatic optimum favored the formation of the so-called last Green Sahara between 12,000 years ago to about 6,000 years ago. Favourable

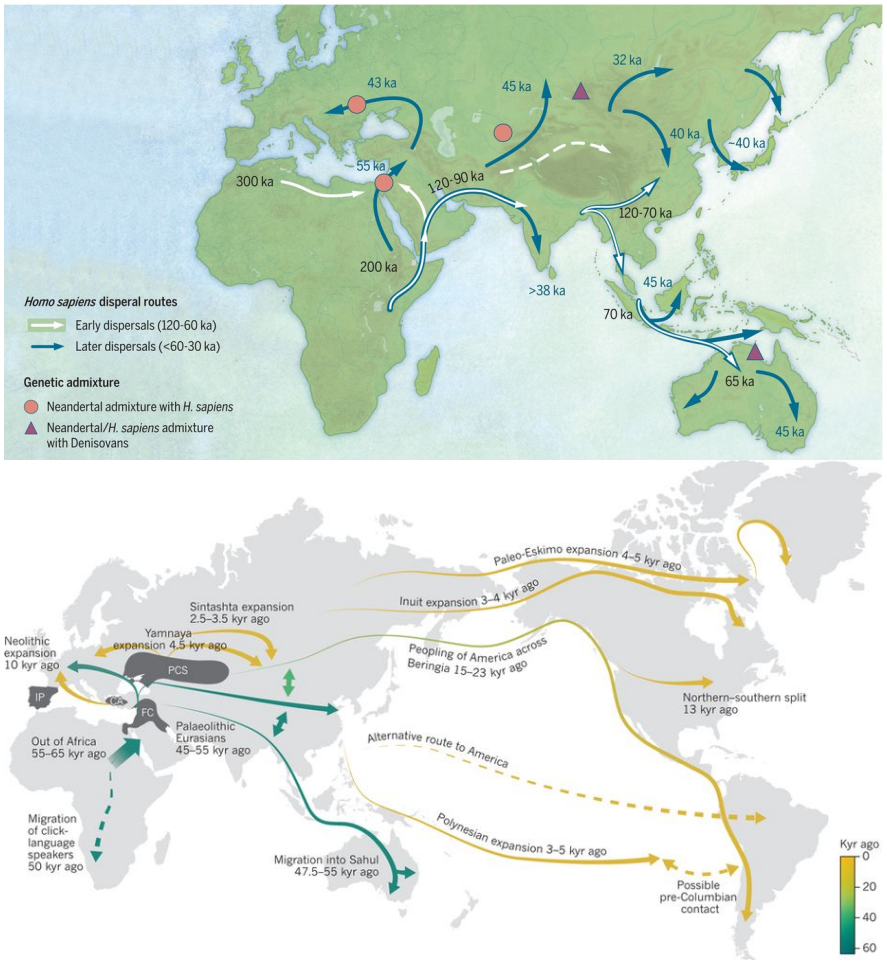

Figure 6: Human dispersal routes out of Africa. Multiple-waves model (top) and single-wave model (bottom) as shown by Bae et al. (2017) and Nielsen et al. (2017), respectively.

climatic conditions and a fertile environment promoted the occupation and dispersal of trans-Saharan human populations, whose coalescence age dates back to this last Green Sahara period, while most North African and sub-Saharan clades expanded locally in the following arid phase (D'Atanasio et al., 2018). The present desert as we know it stabilized about 2,700 years ago after a gradual transition from humid to dry that started 6,000 years ago according to palaeoclimatic evidences (Kröpelin et al., 2008). This desert nature has dramatically shaped its demographic history apart from sub-Saharan Africa.

The Mediterranean Sea is the northern limit of North Africa, which together with the Nile river have turned North Africa into a very demographically dynamic through time. A great variety of cultures has inhabited North Africa since long ago: the Aterian during the Middle Palaeolithic (145,000 to 40,000 years ago) (Nespoulet et al., 2008), the Iberomaurusian during the Epipalaeolithic (22,000-9000 ya) (Newman, 1995), the Capsian during the Mesolithic (10,000-6000 ya) (Rahmani, 2004) or the transition Neolithic from 5,500 ya (Fadhlaoui-Zid et al., 2011). Several of the greatest West Eurasian civilizations have had settlements in North Africa: Egyptians, Phoenicians, Carthaginians, Greeks, Romans, Vandals, Byzantines, Arabs, Ottomans (Camps, 1974; Hunt et al., 2010; Belcastro et al., 2010; Barton et al., 2013). Therefore, North Africa has historically harbored a complex amalgam of cultures, populations and, as a consequence, genetic backgrounds.

Current North African populations belong to two main ethnicities: Berbers (or Amazighs), whose ancestors evolved from prehistoric Iberomaurusian, Capsian and the Neolithic-intruded communities (Desanges, 1980; Fadhlaoui-Zid et al., 2011); and Arabs, whose ancestors entered North Africa during the Muslim conquest of North Africa. It must be noted that these two categories are merely linguistic categories and that this classification is not exhaustive, since other communities exist, such as the Copts, who inhabited eastern North Africa since the 1st century during the Roman period; or the Jew communities, some of which date back to pre-Roman times (Maghrebi Jews), who lived under Arab rule during the Middle Age and mixed with Sephardic Jews coming from Iberia between the 13th and 16th centuries, before collapsing in the mid-20th century during the Jewish exodus from the Arab countries. After the Islamization process of North Africa (beginning in the 7th century), many non-Arab communities have adopted Arab language and culture as their own. As a consequence, current Arab and Berber communities do not have neither demographic nor genetic independent histories (Arauna et al., 2017; Arauna and Comas, 2017).

Due to its geographic isolation from the rest of the continent and its more prevalent contact with Europe and the Middle East than with the rest of the African continent, North Africa has frequently been considered as a different region from sub-Saharan Africa and, unlike eastern or Southern Africa, has not been considered a key location when studying modern human origins. However, recent studies by Richter et al. (2017) and Hublin et al. (2018) have put North

Africa, particularly the Jebel Irhoud site in Morocco, on the map as yet another African location the first modern humans could have inhabited as back in time as 315,000 years ago. These findings point to a pan-African origin of *Homo sapiens* (see above).

Before these discoveries, human presence in North Africa had been dated back to 160,000 ya (Smith et al., 2007). The Middle Stone Age in North Africa (300,000 - 24,000 years ago) has not been exhaustively studied, and many different cultural and industrial nomenclatures have been proposed but they do not cover the real underlying variability (Scerri, 2017). In addition, palaeontological findings (see Figure 7) are biased towards non-desert regions, which shows an incomplete picture of the North African Middle Stone Age cultural landscape.

Figure 7: Map of some of the most relevant North African hominin sites (Balter, 2011).

.

From the Aterian to the Capsian cultures, the main open questions regarding North African prehistory are related to the population continuity or replacement through time, as well as the origin and development of the subsequent cultures.

A demic diffusion model with possible interactions with local groups is commonly accepted for the North African Neolithic (Mulazzani et al., 2016), being the Middle East the most likely origin of newcomer populations (Morales et al., 2013). Although several studies have pointed to trans-Gibraltar demic or technological diffusion from Iberia into North Africa (Linstädter and Kehl, 2012; Mulazzani et al., 2016), the most accepted hypothesis points to a contemporary and demographically similar Neolithic expansion from the Middle East to

both Europe and North Africa from the Middle East (Mulazzani et al., 2016; Pimenta et al., 2017).

Genetic studies point out that current North African populations are the result of a complex history of admixture involving at least three main admixture events: a back-to-Africa migration earlier than 12,000 (Henn et al., 2012),a wave coming from the Middle East around 1,400 ya and several migrations coming from sub-Saharan Africa, mainly due to slave trade, from 1,200 ya (Harich et al., 2010; Arauna et al., 2017).

Classical and uniparental genetic markers were the first to be studied. Mitochondrial DNA diversity in North African has been identified as the oldest genetic evidence, since some haplogroups with North African origins (U6 and M1) are related to the North African Middle Stone Age (Secher et al., 2014). Other studies have claimed U6 originated in Western Asia and expanded to North Africa around 22,000 ya (Pereira et al., 2010). These two haplogroups, considered the autochthonous North African mitochondrial haplogroups, have opposite frequency gradients: U6 decreases from West to East, while M1 decreases from East to West (Olivieri et al., 2006; Pennarun et al., 2012).

A heterogeneous mtDNA haplogroup distribution in North Africa has been reported (Fadhlaoui-Zid et al., 2011), poiting to a highly admixed scenario. Indeed, according to recent studies (Font-Porterias et al., 2018), the frequency of autochthonous haplogroups in North Africa U6 and M1 is around 10%, whilst 20% belong sub-Saharan L lineages and the remaining 70% have an Eurasian origin (H, HV, R0, J, T, U, W). Kefi et al. (2016) dated the presence of these Eurasian lineages in North Africa from at least 20,000 ya and highlighted the existence of gene flow between Southern and Northern coast of the Mediterranean.

On the other hand, a strong structure has been reported in the Y chromosomal diversity in North Africa (Arredi et al., 2004). However, this geographical structure is somehow diluted when the most frequent Y chromosome haplogroup, E-M183 (M81), is analyzed. This haplogroup shows a global frequency of more than 50% in North African populations and it is almost absent elsewhere, with the exception of southern Europe (the Iberian Peninsula and Sicily). A recent origin of this lineage (between 3,000 and 2,000 years ago) was revealed by

Solé-Morata et al. (2017). This haplogroup descends from the East African E-M35 haplogroup, it is distributed in a decreasing (but not homogeneous) gradient from eastern to western North Africa, being particularly high (more than 80% and up to 100% in particular groups) in Berbers and showing its highest frequency in Morocco (67.37%) and its lowest frequency in Egypt (11.9%) (Cruciani et al., 2004; Fadhlaoui-Zid et al., 2013). The lack of inner population structure, together with the observed frequency gradient, is compatible with a western North African origin of the haplogroup and a rapid demic expansion through all the area.

Interestingly, despite the origin of mtDNA haplogroups in North Africa being older than that of the Y chromosome, historical events such as the trans-Saharan Islamic slave trade mainly contributed to the mtDNA and autosomal gene pool, whereas the North African paternal gene pool of sub-Saharan origin (whose coalescence age dates back to the last Green Sahara) was mainly shaped earlier and suffered only a marginal effect.

The first genome-wide study of North African human populations was perfomed by Henn et al. (2012), who used Affymetrix 6.0 array data to reveal a very complex genetic composition of North African populations. A native North African genetic component was identified, together with three other ancestral components: a European one, a Middle Eastern one, and a sub-Saharan one. This structure was confirmed by Arauna et al. (2017), who used haplotype-based methods in order to characterize the North African gene pool.

According to the Henn et al. (2012), North African populations diverged from Eurasian populations between 40,000 and 12,000 ya. This implies that current North African groups are not the direct descendants of the populations located in the same area more than 40,000 years ago but continuity might have occurred during the last 12,000 years. The fact that the autochthonous component is clearly different from the sub-Saharan one implies that its origin is not African, but from a group of Eurasians that returned to North Africa after the out of Africa migration.

More recent genetic contact between North Africans and their surrounding populations has also been described, influencing both North African populations (Henn et al., 2012) and their neighbors from the Southern Europe Botigue et al. (2013). In particular, a relatively homo-

geneous European component is observed across all North African groups, whereas the Middle Eastern influence has a differential impact which is high in eastern North Africa and decreases as the distance from the Arabian Peninsula increases. The sub-Saharan African admixture in North Africa occurred much more recently, between 24 and 41 generations ago, and it does not affect all North African groups.

A clear genetic differentiation between Arabs and Berbers is not found (Arauna et al., 2017). Berbers include from heterogeneous communities and isolated groups with high inbreeding and low effective population sizes, to admixed groups with high frequencies of sub-Saharan and Middle Eastern components (Arauna et al., 2017) (Figure 8).

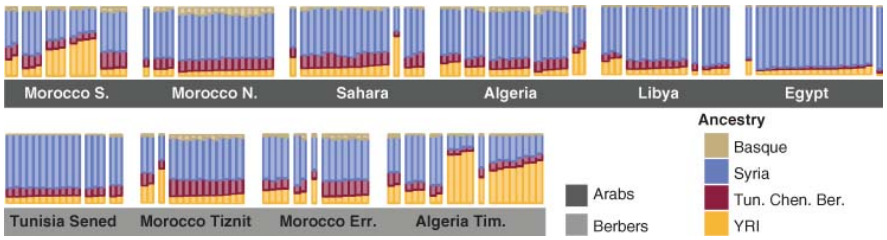

Figure 8: Analyses of autosomal markers. Each bar shows the proportion of the genome that each individual share with each of the four ancestral populations according to ChromoPainter analyses. The separation between bars within each population represents clusters of individuals with similar ancestry proportions (Arauna and Comas, 2017).

Geographic distance and genetic diversity were not found to be correlated in North African populations, probably because of heterogeneous or unbalanced admixture (Arauna et al., 2017).

Another ancestral component was identified in some eastern North African populations, particularly in the Copts from Egypt (Dobon et al., 2015). Copts have a distinct ancestry and, unlike Egyptian Arab-speakers, they have not been influenced by the Middle Eastern component, which points that their genetic composition could be similar to an ancestral Egyptian population, without the present strong Arab influence. Other North African ethnic minorities, like the Jews, also show a distinct history from their Arab and Berber neighbours, mainly due to the genetic isolation of their communities (Campbell et al., 2012).

North African populations have also experienced archaic hominin

introgression. In particular, Sánchez-Quinto et al. (2012) tested admixture with Neanderthals. The excess of derived alleles shared with Neanderthals when compared to sub-Saharan Africans could not be attributed to recent gene flow with Middle Easterns and Europeans, since Neanderthal genetic signal was found to be higher in populations with a local, pre-Neolithic North African ancestry, which points to introgression from Neanderthals to the North African ancestors of current North Africans.

During the last decade, the lack of ancient DNA from North Africa has been a setback in the study of its demographic history. However, earlier this year, two relevant studies have been published including ancient DNA data from North Africa.

The first one, developed by Van De Loosdrecht et al. (2018), explored the genomes of ancient individuals belonging to the Iberomaurusian culture found in the Grotte des Pigeons (Taforalt) in Morocco. These samples, which provide the most African ancient DNA analyzed so far (15,100 - 13,900 ya), share about two-thirds of their genetic ancestry with Middle Eastern Natufians, who lived 14,500 to 11,000 ya; and one-third with sub-Saharan Africans, particularly with the putative ancestors of today West Africans and the Hadza from Tanzania, a higher proportion than the one found in current North Africans. No signals from Palaeolithic Europeans were detected in the Taforalt samples, thus refuting a previously proposed European origin for the Iberomaurusian culture (Ferembach, 1985). The Taforalt individuals and Natufians might have inherited their shared DNA from a North African or Middle Eastern ancestral population living more than 15,000 years ago. As for the sub-Saharan DNA in their genome, the Iberomaurusians may have admixed with southern migrants.

The second study added two extra ancient datasets, one from the Early Neolithic site of Ifri n'Amr or Moussa in Morocco, dated to approximately 7,000 ya, and another one of Late Neolithic individuals from the Kelif el Boroud site in Morocco, dated to approximately 3,000 ya. Fregel et al. (2018) found that Early Neolithic individuals had a similar ancestry pattern to the Epipalaeolithic individuals from the Taforalt site, while Late Neolithic individuals shared only around half of their ancestry with Early Neolithic and Later Stone Age North Africans. The other half of their ancestry was shared with Early Neolithic individuals from southern Iberia, pointing to trans-Gibraltar Strait migration during the Neolithic. These observations point to

an adoption of technological innovations by North African populations at early stages of the Neolithic transition, whilst subsequent migrations from Europe influenced further Neolithic developments.

Objectives

The objective of this thesis is to study the human African genetic landscape through whole-genome sequencing data, and put it into a worldwide context.

Specifically, the goals of this thesis are:

1. Characterize the internal human diversity in Africa and North Africa overcoming ascertainment bias-related problems by using complete genomes.
2. Characterize the deepest splits in the human lineage and the processes (such as migrations, admixture, and archaic introgression) that have shaped the current genetic map.
3. Study the gene flow between Africans groups and populations living in surrounding areas.
4. Address the continuity vs replacement debate in North Africa.
5. Study the demographic patterns that have shaped the genetic diversity of current North African populations.
6. Revisit North African ancestral components as a result of gene flow coming from different surrounding areas through time.

Chapter 2

Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from unknown Neanderthal sister taxa into sub-Saharan populations

Authors: Belen Lorente-Galdos[1,2,*], Oscar Lao[3,4,*], Gerard Serra-Vidal[1,*], Gabriel Santpere[1,2], Lukas F.K. Kuderna[1], Lara R. Arauna[1], Karima Fadhlaoui-Zid[5], Ville N. Pimenoff[6,7], Himla Soodyall[8], Pierre Zalloua[9], Tomas Marques-Bonet[1,3,10], David Comas[1]

[1] Institut de Biologia Evolutiva (UPF/CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, 08003, Spain

[2] Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA

[3] CNAG/CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona 08028, Spain

[4] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[5] University Tunis El Manar, Laboratory of Genetics, Immunology and human Pathology, Tunis, 2092, Tunisia

[6] Department of Cancer Epidemiology and Prevention, Bellvitge Institute of Biomedical Research (IDIBELL), Catalan Institute of Oncology,

Barcelona, Spain

[7] Department of Archaeology, University of Helsinki, Finland

[8] Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand and National Health Laboratory Service, Johannesburg, South Africa

[9] School of Medicine, The Lebanese American University, Beirut, 1102-2801, Lebanon

[10] Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, 08003, Spain

[*] These authors contributed equally to this work

Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biol. 2019 Apr 26;20(1). DOI: 10.1186/s13059-019-1684-5

Chapter 3

Whole-genome sequences reveal heterogeneity in the Palaeolithic population continuity and Neolithic expansion in North Africa

Authors: Gerard Serra-Vidal[1], Marcel Lucas-Sanchez[1], , Karima Fadhlaoui-Zid[2], Asmahan Bekada[3], Pierre Zalloua[4], David Comas[1]

[1] Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Departament de Ciencies Experimentals i de la Salut, Barcelona, 08003, Spain.

[2] University Tunis El Manar, Department of Genetics, Immunology and human Pathology, Tunis, 2092, Tunisia.

[3] Département de Biotechnologie, Faculté des Sciences de la Nature et de la Vie, Université Oran 1 (Ahmad Ben Bella), Oran, Algeria.

[4] The Lebanese American University, Chouran, School of Medicine, Beirut, 1102-2801, Lebanon.

Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D. Heterogeneity in Palaeolithic Population Continuity and Neolithic Expansion in North Africa. Curr Biol. 2019 Nov 18;29(22):3953-3959.e4. DOI: 10.1016/j.cub.2019.09.050

Chapter 4

Discussion

Approaching the African genetic diversity through whole-genome sequencing data analyses has been a good opportunity to better decipher its complexity and answer or refine relevant open questions, such as the archaic introgression and admixture landscape in Africa or the population continuity vs replacement debate in North Africa.

Population structure and demographic history, both at local and continental scales, can be studied with a higher level of deepness and accuracy than ever before thanks to the next-generation sequencing data and novel methodologies, such as multiple sequentially Markovian coalescent (MSMC), haplotype-based methods such as ChromoPainter, approximate bayesian computation coupled to deep learning (ABC-DL) or S^*, which have been used together with population genetics classical methods such as principal component analysis, ADMIXTURE or f-statistics.

Many relevant studies regarding African genetic diversity have been published since the beginning of this work, which has forced me to adjust the focus of the study.

1 The African human genetic landscape

African human populations harbor the higher level of genetic diversity than any other world region due to the African origin of modern

humans, which have inhabited Africa longer than anywhere else in the world (Cann et al., 1987; Tishkoff et al., 2009). However, its internal genetic diversity has been classically understudied compared to other regions. Here, we have tried to characterize this diversity. Four main groups have been identified in terms of genetic differentiation and ancestral components: Khoisan (southwestern Africa), Pygmies (Central Africa), North Africans, and sub-Saharan non-hunter-gatherer groups. It must be noted that these groups are not uniform but very heterogeneous, as shown by the pairwise differences between individuals.

These groups are further confirmed by the study of population size evolution, which reveals a widespread population size decrease affecting all populations after the divergence of ancestral population of all present day humans around 200,000 years ago (in accordance to Mallick et al. (2016)). Hunter-gatherer groups, however, show a softer decline, followed by Pygmy groups, with all other sub-Saharan populations showing a more pronounced decrease. For their part, non-African populations, as well as North Africans, show a sharp N_e decrease, in accordance with their ancestors' migration from Africa to Eurasia (Pickrell et al., 2012; Petersen et al., 2013; Kim et al., 2014).

The observed genetic diversity shows a relative correlation with geography, especially in the north-south axis, i.e., geographically closer populations tend to be genetically more similar. However, this correlation is not strong and it has many nuances. Linguistic and lifestyle diversity also appear to be in association with genetic diversity, since hunter-gatherer communities are much differentiated from their surrounding agropastoralist populations, while Niger-Kordofanian-speaking populations show relative similarity to each other in spite of inhabiting very distant geographic areas. These genetic homogeneity across sub-Saharan Africa can be attributed to the Bantu expansion that took place $\sim 5,000 - 3,000$ years ago and swept most of the preexisting diversity in the southern half of the continent (Alves et al., 2011; Li et al., 2014; Schlebusch et al., 2012). Back-to-Africa gene flow to sub-Saharan Africa was also identified in some particular groups, showing its highest levels in Eastern African groups (Toubou from North Chad and Kenyan Bantu) (in accordance with Gallego Llorente et al. (2015)) and detecting weak but significant traces in the Khoisan (according to Schuster et al. (2010); Pickrell et al. (2013)).

The deepest human splits, corresponding to African hunter-gatherer

groups (Khoisan, Mbuti and Baka) (Schlebusch et al., 2017) were tested for admixture with their corresponding neighboring populations in order to be better characterized. While the Khoisan and Mbuti share significant amounts of derived alleles with some of their closer neighbors (Bantu and Dinka/Laal, respectively), no signature of extensive admixture could be detected between Mbuti and Baka and their Bantu neighboring groups, pointing to a relative genetic isolation of the Pygmies.

This contrasts with the uniparental phylogenetic analyses, particularly the Y chromosome, for which Baka and Mbuti Pygmies share haplogroup E with most sub-Saharan samples, including Bantus, whereas the deepest uniparental lineages (haplogroup A) correspond to the Khoisan. The latter also have the most basal mtDNA haplogroup (L0), like Mbuti Pygmies, while L1 is found in the Baka. The high homogeneity found in the case of the Y-chromosome, together with the lack of correspondence for the mtDNA, points to a mostly male-driven Bantu expansion affecting Central and South Africa and a higher demographic impact on the Pygmies than on the Khoisan groups.

Evidence for archaic introgression in African populations

Interbreeding between archaic and modern humans has happened several times across time, involving Neanderthals, Denisovans, as well as several unidentified hominins for which no DNA samples have been found.

The archaic introgression question in Africa has been object of great debate in recent years (Hammer et al., 2011; Lachance et al., 2012; Hsieh et al., 2016; Xu et al., 2017; Durvasula and Sankararaman, 2018). For a long time, it has remained the only continent without genetic evidences for archaic hominin presence, due to the rapid degradation of fossils in sub-Saharan African environments, except for North Africa, which is the only region where Neanderthal introgression has been detected (Sánchez-Quinto et al., 2012).

However, several methodologies have been developed in order to infer archaic introgression when no ancient sample is available (Racimo et al., 2015). Using these techniques, Hammer et al. (2011) inferred archaic introgression in 61 non-coding regions from two hunter-gatherer

groups (Biaka Pygmies and Khoisan). 2% of the genome was estimated to have been introgressed in these populations approximately 35,000 years ago from archaic hominins diverging from the modern human lineage around 700,000 years ago. Lachance et al. (2012) used whole-genome sequences of three Sub-Saharan hunter-gatherers (Western Africa Pygmies, Hadza and Sandawe), estimating archaic introgression with one or more archaic hominin populations about 40,000 years ago. Hsieh et al. (2016) inferred one admixture event from an unknown archaic population into the ancestors of Western Pygmies during the last 30,000 yr. According to Skoglund et al. (2017), an extinct basal western African lineage has contributed to current West Africans' gene pool, specifically, 13% for Mende and 9% for Yoruba. Durvasula and Sankararaman (2018) estimated 8% of Yoruba population gene pool traces its origin to an unidentified, archaic population.

Our study contributes to the complexity of the archaic introgression landscape in Africa, using an ABC-DL approach on whole-genome sequences not only discarding a scenario without archaic introgression, whose likelihood is almost zero, but also estimating extensive archaic admixture across Africa, with ancient traces detected in West Africa, East Africa and South Africa, into both hunter-gatherer and non-hunter-gatherer groups. In particular a proportion of 8.8% in Khoisan, 5.8% in Mbuti Pygmies, and 3.3% in Mandenka were estimated. Our analysis also pointed to the nature of the ghost population involved, which is likely to be an extinct sister lineage of Neanderthals from whom they diverged 450,000 years ago.

2 The North African genetic landscape

North Africa has a particular demographic structure and an essentially different genetic history from the rest of Africa and it needs to be considered apart due to its high affinity with its non-African surrounding regions (Jakobsson et al., 2008; Henn et al., 2012; Mallick et al., 2016). Despite displaying the lowest levels of genetic diversity on a continental scale, North Africa presents high levels of heterogeneity and diversity between and within populations when compared to most non-African populations (Bosch et al., 1997; Henn et al., 2012; Arauna et al., 2017). As we have seen, two main factors have

contributed to shape its internal diversity: on the one hand, the different ancestral components present in the North African gene pool, namely, sub-Saharan, European and Anatolian Neolithic, Middle Eastern, Iranian-Caucasian, and autochthonous Iberomaurusian North African, which are present in North African populations at different frequencies; and, on the other hand, the differential admixture, inbreeding and drift that has affected particular populations.

In particular, the autochthonous North African component shows a West-East cline in current populations, being more prevalent in western North Africa and in Berber-speaking groups than in Arab populations. The effect of genetic drift was confirmed through f_3 (Patterson et al., 2012), maximum likelihood (Pickrell and Pritchard, 2012), and runs of homozygosity, all of which point to differential genetic drift between and within populations, which reinforces the heterogeneous conception of North African populations.

The role of ancient DNA in North African population genetics studies

The lack of ancient DNA (aDNA) samples has been a handicap in genetic studies, especially for those focusing on the demographic history of populations. In the last decade, technical improvements have allowed the collection of ancient DNA samples, and the first whole-genome ancient sequences have been published. These datasets can be studied and merged with current populations data, after applying specific quality controls and filters accounting for post-mortem damage, which produces an excess of cytosine deamination and degradation of aDNA (Molak and Ho, 2011; Dabney et al., 2013).

Thanks to the addition of ancient DNA datasets to our data, we were able to confirm population continuity in North Africa during the last 15,000 years, as well as to describe differences among regions regarding to the ancient native component. In particular, three recently published ancient datasets from North Africa (Rodríguez-Varela et al., 2017; Van De Loosdrecht et al., 2018; Fregel et al., 2018) (aboriginal Canary Islanders, Moroccan Iberomaurusian and Moroccan Neolithic, respectively) have been very useful to shed light on the demographic history of North Africa, in particular to the long-standing debate about population continuity or replacement. These data provide a

time transect with samples from four different periods (Epipaleolithic, Early Neolithic, Late Neolithic, and 7th - 11th centuries), which represents an excellent framework to study the evolution of populations in this area during the last 15,000 years.

Together with whole-genome sequences from current populations, we were able to describe population continuity in North Africa from the Iberomaurusian period until the present era. The North African autochthonous ancestry found in current samples is heterogeneous: it is more prevalent in western North Africa and Berber groups than in eastern North Africa populations, particularly Egyptians and Libyans, who have had a higher influence from the Middle East in recent times (from the Arab conquest) and show a more diluted autochthonous component as a result.

We also saw that the Neolithic transition and its demic diffusion through North Africa was the migration with the highest demographic impact in the current genetic landscape, since present-day populations share a similar genetic pattern with populations from the Late Neolithic (3,000 ya) and aboriginal Canary Islanders coming from North Africa 2,500 ya, whereas Early Neolithic samples (5,000 ya) are shown to be more similar to Iberomaurusian Epipalaeolithic ones (15,000 ya), which points to a major shift in the North African gene pool during the Middle-Late Neolithic expansion.

High-coverage, whole-genome sequences from Neanderthals and Denisovans (Prüfer et al., 2014; Meyer et al., 2012) were also used to confirm the presence of Neanderthal signature in North Africa, as stated by Sánchez-Quinto et al. (2012), contrary to the rest of the continent, where no Neanderthal nor Denisova genetic signals were found.

Haplotype-based methods as a new standard

Haplotype-based methods account for linkage disequilibrium (LD) between SNPs, contrary to methods that are based on allele frequencies. Since variants in LD are transmitted together as a single block, haplotype-based methods provide for better, more reliable and higher-resolution estimates than SNPs considered as independent units. In addition, haplotypes are less affected by the fixation of derived alleles

by drift, which makes them suitable to study past events.

In our study, ChromoPainter and fineSTRUCTURE (Lawson et al., 2012) were used to test the population continuity hypothesis, providing extra evidence for what had already been observed through PCA and ADMIXTURE. PSMC and MSMC were used to infer population size evolution and separation history.

ChromoPainter has power enough to detect the North African Epipalaeolithic signal in current North Africans, while the signal is not detected in out-of-Africa populations, which reinforces the continuity hypothesis observed using SNP-based methods and the previous study of Fregel et al. (2018). fineSTRUCTURE, which defines clusters of individuals using haplotype information and a Markov chain Monte Carlo algorithm, groups North Africans together and apart from neighboring populations, and confirms that no geographic structure can be established in the region, as it had been previously reported (Arauna et al., 2017).

On the other hand, PSMC and MSMC recapitulate the out-of-Africa bottleneck in non-African and North African populations, and at the same time outline the historically higher effective population size of hunter-gatherer groups and the lower N_e reduction in North Africans compared to Eurasians, notably due to their strong non-African ancestral component (Haber et al., 2016).

3 Whole-genome sequences as a new paradigm for population genetics

Technological advances and computational power increase in the last decade have changed genetic studies. As a consequence, the cost of genotyping and sequencing has plummeted and whole-genome sequencing (WGS) is becoming a new standard. This new technology has the power to overcome many of the handicaps of the previous approaches, particularly the ascertainment bias typical of genotyping arrays (Lachance and Tishkoff, 2013), whose SNPs have been identified in discovery panels (which tend to have higher minor allele frequencies than random SNPs), and are spaced in order to tag common variation across the genome (Novembre and Ramachandran,

2011). As a consequence, these SNPs tend to be biased to SNPs that are related to genome-wide association studies and to coding variation in European populations, which have been typically overrepresented in genetic studies such as the 1,000 Genomes Project. In population genetics, the usage of such biased datasets tend to come along with problems when applying specific statistical methods that assume a random subset of all variants is being surveyed.

Moreover, different whole-genome sequence datasets can be merged keeping most of the variants and, as a consequence, keeping its high statistical power, contrary to what happens when dealing with SNP array data from different arrays.

The lack of SNP ascertainment bias in WGS has been critical to increase the accuracy of population genetic analyses (Lachance and Tishkoff, 2013). However, the quality control filters applied to next-generation sequencing (NGS) data can also induce ascertainment bias that, despite being subtle, can be challenging for population analyses (Zhang and Dolan, 2010).

Whole-genome sequencing analysis have confirmed some results based on previous technologies, whilst other inferences have been refined or reinterpreted (Veeramah and Hammer, 2014).

In our studies, a total of more than 22 million variants were described, including many previously unknown SNPs (particularly rare variants, which are often missed in genotyping experiments), and the catalog of African genetic variability has been greatly increased, in particular, the North African SNPs catalog, thanks to the first study of the area at whole-genome sequence level.

In addition to the refinement of population genetic studies and the discovery of millions of previously unknown variants, whole-genome sequencing allows the usage of new methodologies that take profit of the complete sequence, which could not be applied to genome-wide sets of pre-ascertained SNPs. This is the case of sequentially Markovian coalescent methods, which use sequences from one or more individuals to infer genetic separation history and population size changes, admixture, or migration. These methods include PSMC (Li and Durbin, 2011), which uses a Hidden Markov Model to infer local time to the most recent common ancestor (TMRCA) based on the heterozygous density in one diploid genome; MSMC (Schiffels

and Durbin, 2014), which extends PSMC to multiple individuals, focusing on the first coalescence event for any pair of haplotypes and locally infers branch lengths and coalescence times from the observed pattern of mutations in multiple individuals; or diCal (Steinrücken et al., 2015)), based on a sequentially Markov conditional sampling distribution framework, providing an accurate approximation of the probability of observing a newly sampled haplotype given a model and a set of previously sampled haplotypes.

Alternative methods to the coalescent approach have been developed to analyse WGS, such as *∂a∂i* (Gutenkunst et al., 2009), which allows the modeling of three simultaneous populations and estimates growth rates, bottlenecks, and migration rates. Simulation-based methods, such as approximate bayesian computation (ABC) approaches, as well as statistics like S^*, have previously been applied to genotyping datasets, but their application on whole-genome datasets has allowed not only a higher exhaustivity and an unbiased estimation at a global scale but also at a local scale, providing the power of identifying specific genomic regions subject to particular demographic effects (bottlenecks, expansions, selective sweeps, introgression, etc). Having complete sequences multiplies their applications and provides tools of great power and usefulness to study population genetics.

Other methods, like classical frequency-based methods, such as multidimensional statistics (such as principal component analysis), maximum likelihood estimation of individual ancestry (such as STRUCTURE or ADMIXTURE), F_{ST} or f-statistics), can be applied to SNP array data and are not exclusive to WGS data, but WGS provides a full and unbiased genetic picture.

4 Challenges and future studies

A more extensive geographical coverage of the geographic, linguistic and cultural diversity is needed to study the global genetic landscape more deeply. Not only because the high diversity observed between African populations requires an exhaustive sampling to account for most of its genetic variability, but also because many cultural and linguistic isolates exist in the African continent, (such as the Laal in our study, a language spoken by less than 1,000 people, or some Khoisan

groups such as the ‡Khomani), which harbor a significant genetic diversity that is left out if not considered in continental studies.

Not only more sampled populations would be helpful, but also more samples per population are needed to account for the genetic heterogeneity at a continental, local or population scales. In African populations, within-population diversity tends to be very high. This has two main consequences for studies with few samples per population: on the one hand, population variability might not be entirely represented, and, on the other hand, some methods might lack statistical power to produce significant inferences. For example, methods based on the site frequency spectrum (SFS) have more power when a medium-high sample size is considered. Heterogeneity of certain populations, as seen in some North African groups, also presents a challenge, since they were found to be very diverse, with some of their populations (such as Berber- and Arab- speaking Moroccan groups) needing more representatives in order to gather a complete overview of their internal variation.

The debate about archaic introgression in Africa needs further investigation and probably a global, continental approach in order to summarize all studies to this date, which are not comprehensive from both a methodological and a population perspective. The recovering of ancient DNA from the inferred sub-Saharan African hominins would shed light into this question and complement indirect estimates of ghost population introgression. In North Africa, genomes older than the Iberomaurusians could help to trace history back to the Aterian culture and to test for longer population continuity.

From a methodological point of view, next-generation sequencing data mapping and variant calling present certain limitations derived from the short length of the reads, such as the impossibility of aligning reads within repetitive regions in the reference genome or in regions that may not exist in the reference genome due to structural variants in the analyzed sequence (Metzker, 2010).

There are other relevant challenges for demographic inference related to NGS besides SNP calling uncertainty (Nielsen et al., 2011), such as reference genome bias and phase uncertainty. Reference genome bias (that is, the tendency to induce mapping errors of reads that differ from their homologous location in the reference genome at polymorphic sites) (Sousa and Hey, 2013), provokes an underestimation of

the differences between the mapped data and the reference genome, leading to more mapping errors in samples that are more divergent from the reference genome, which is not equally representative of all human populations (Ross et al., 2013). This is improved in posterior reference assemblies, such as GRCh38, which offers more accurate genomic analysis, better annotation, and fewer false positive structural variants (Guo et al., 2017).

Phase uncertainty (i.e., the difficulty to know whether two reads come from the same or from different haplotypes) (Sousa and Hey, 2013), challenges the assessment of linkage disequilibrium over longer distances, which can affect demographic estimates, particularly for haplotype-based or hidden Markov model-based methods, which use LD information. Current unphased NGS data requires computational phasing before applying these methods, which introduces computational errors in the data and requires a reference panel which tends to be incomplete and biased. Consequently, the demographic inferences made with this data are more prone to be inaccurate and contain errors.

Third-generation sequencing technologies, which are based on direct sequencing of individual nucleic acid molecules (Schadt et al., 2010; Ari and Arikan, 2016), have higher throughput and promise great advantages over NGS data, mainly due to the length of the obtained reads, which help to reduce reference genome bias and phase uncertainty, as well as the coverage of repeated regions and the identification of structural variants.

Bibliography

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

Alves, I., Coelho, M., Gignoux, C., Damasceno, A., Prista, A., and Rocha, J. (2011). Genetic Homogeneity Across Bantu-Speaking Groups from Mozambique and Angola Challenges Early Split Scenarios between East and West Bantu Populations. *Human Biology*, 83(1):13–38.

Arauna, L. R. and Comas, D. (2017). Genetic Heterogeneity between Berbers and Arabs. In *eLS*, pages 1–7. John Wiley & Sons, Ltd, Chichester, UK.

Arauna, L. R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhlaoui-Zid, K., Zalloua, P., Hellenthal, G., and Comas, D. (2017). Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa. *Molecular biology and evolution*, 34(2):318–329.

Ari, Å. and Arikan, M. (2016). Next-Generation Sequencing: Advantages, Disadvantages, and Future. In *Plant Omics: Trends and Applications*, pages 109–135. Springer International Publishing, Cham.

Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., Pascali, V. L., Novelletto, A., and Tyler-Smith, C. (2004). A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa. *The American Journal of Human Genetics*, 75(2):338–345.

Avery, O. T., Macleod, C. M., McCarty, M., and Peltier, L. F. (2000). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a

desoxyribonucleic acid fraction isolated from Pneumococcus type III. *Clinical Orthopaedics and Related Research*, 79(379 SUPPL.):137–58.

Bae, C. J., Douka, K., and Petraglia, M. D. (2017). On the origin of modern humans: Asian perspectives. *Science*, 358(6368):eaai9067.

Balter, M. (2011). Was North Africa the Launch Pad for Modern Human Migration? *Science*, 331(6013):20–23.

Barton, R. N., Bouzouggar, A., Hogue, J. T., Lee, S., Collcutt, S. N., and Ditchfield, P. (2013). Origins of the iberomaurusian in NW Africa: New AMS radiocarbon dating of the middle and later stone age deposits at taforalt cave, Morocco. *Journal of Human Evolution*, 65(3):266–281.

Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370–379.

Behar, D. M., Van Oven, M., Rosset, S., Metspalu, M., Loogväli, E. L., Silva, N. M., Kivisild, T., Torroni, A., and Villems, R. (2012). A "copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics*, 90(4):675–684.

Belcastro, M. G., Condemi, S., and Mariotti, V. (2010). Funerary practices of the Iberomaurusian population of Taforalt (Tafoughalt, Morocco, 11-12,000 BP): the case of Grave XII. *Journal of Human Evolution*, 58(6):522–532.

Belmont, J. W., Boudreau, A., Leal, S. M., Hardenbol, P., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'Ang, L. Y., Huang, W., Liu, B., Shen, Y., Tam, P. K. H., Tsui, L. C., Waye, M. M. Y., Wong, J. T. F., Zeng, C., Zhang, Q., et al. (2003). The international HapMap project. *Nature*, 426(6968):789–796.

Beltrame, M. H., Rubel, M. A., and Tishkoff, S. A. (2016). Inferences of African evolutionary history from genomic data. *Current Opinion in Genetics and Development*, 41:159–166.

Berger, L. R., Hawks, J., de Ruiter, D. J., Churchill, S. E., Schmid, P., Delezene, L. K., Kivell, T. L., Garvin, H. M., Williams, S. A., DeSilva, J. M., Skinner, M. M., Musiba, C. M., Cameron, N., Holliday, T. W.,

Harcourt-Smith, W., et al. (2015). Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. *eLife*, 4(September2015).

Bertorelle, G., Benazzo, A., and Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, 19(13):2609–2625.

Bosch, E., Calafell, F., Pérez-Lezaun, A., Comas, D., Mateu, E., and Bertranpetit, J. (1997). Population History of North Africa: Evidence from Classical Genetic Markers. *Human Biology*, 69(3):295–311.

Bräuer, G. (2008). The origin of modern anatomy: By speciation or intraspecific evolution? *Evolutionary Anthropology: Issues, News, and Reviews*, 17(1):22–37.

Brown, F. H., McDougall, I., and Fleagle, J. G. (2012). Correlation of the KHS Tuff of the Kibish Formation to volcanic ash layers at other sites, and the age of early Homo sapiens (Omo I and Omo II). *Journal of Human Evolution*, 63(4):577–585.

Campbell, C. L., Palamara, P. F., Dubrovsky, M., Botigue, L. R., Fellous, M., Atzmon, G., Oddoux, C., Pearlman, A., Hao, L., Henn, B. M., Burns, E., Bustamante, C. D., Comas, D., Friedman, E., Pe'er, I., et al. (2012). North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. *Proceedings of the National Academy of Sciences*, 109(34):13865–13870.

Camps, G. (1974). Les Civilisations Prehistoriques De L'Afrique Du Nord Et Du Sahara. *L'Homme*, 15(1):145–147.

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

Chbel, F., de Pancorbo, M. M., Martinez-Bouzas, C., Azeddoug, H., Alvarez-Alvarez, M., Rodriguez-Tojo, M. J., and Nadifi, S. (2003). [Polymorphism of six Alu-insertions in residents of Morocco: comparative study in Arab and Berber populations and residents of Casablanca]. *Genetika*, 39(10):1398–1405.

Clarkson, C., Jacobs, Z., Marwick, B., Fullagar, R., Wallis, L., Smith, M., Roberts, R. G., Hayes, E., Lowe, K., Carah, X., Florin, S. A., McNeil, J., Cox, D., Arnold, L. J., Hua, Q., et al. (2017). Human occupation of northern Australia by 65,000 years ago. *Nature*, 547(7663):306–310.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E. B., Zaharova, B., Lavinha, J., Vona, G., Aman, R., Cali, F., Akar, N., et al. (2004). Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *American journal of human genetics*, 74(5):1014–22.

Cummins, J. (2001). Mitochondrial DNA and the Y chromosome: parallels and paradoxes. *Reproduction, fertility, and development*, 13(7-8):533–42.

Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor perspectives in biology*, 5(7):a012567.

D'Atanasio, E., Trombetta, B., Bonito, M., Finocchio, A., Di Vito, G., Seghizzi, M., Romano, R., Russo, G., Paganotti, G. M., Watson, E., Coppa, A., Anagnostou, P., Dugoujon, J.-M., Moral, P., Sellitto, D., et al. (2018). The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. *Genome Biology*, 19(1):20.

Day, M. H., Leakey, M. D., and Magori, C. (1980). A new hominid fossil skull (L.H. 18) from the Ngaloba Beds, Laetoli, northern Tanzania. *Nature*, 284(5751):55–56.

Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., Ermini, L., Fernández, R., da Fonseca, R., Ginolhac, A., Hansen, A. J., Jónsson, H., Korneliussen, T., Margaryan, A., Martin, M. D., et al. (2015). Ancient genomics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130387.

Desanges, J. (1980). *The Proto-Berbers*. Heihemann Publishers, Berkeley, university edition.

Dirks, P. H., Roberts, E. M., Hilbert-Wolf, H., Kramers, J. D., Hawks, J., Dosseto, A., Duval, M., Elliott, M., Evans, M., Grun, R., Hellstrom, J., Herries, A. I., Joannes-Boyau, R., Makhubela, T. V., Placzek, C. J., et al. (2017). The age of homo naledi and associated sediments in the rising star cave, South Africa. *eLife*, 6.

Dobon, B., Hassan, H. Y., Laayouni, H., Luisi, P., Ricaño-Ponce, I., Zhernakova, A., Wijmenga, C., Tahir, H., Comas, D., Netea, M. G., and Bertranpetit, J. (2015). The genetics of East African populations: a

Nilo-Saharan component in the African genetic landscape. *Scientific Reports*, 5(1):9996.

Durvasula, A. and Sankararaman, S. (2018). Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv*, page 285734.

El Moncer, W., Esteban, E., Bahri, R., Gay-Vidal, M., Carreras-Torres, R., Athanasiadis, G., Moral, P., and Chaabani, H. (2010). Mixed origin of the current Tunisian population from the analysis of Alu and Alu/STR compound systems. *Journal of Human Genetics*, 55(12):827–833.

Fadhlaoui-Zid, K., Haber, M., Martínez-Cruz, B., Zalloua, P., Elgaaied, A. B., and Comas, D. (2013). Genome-wide and paternal diversity reveal a recent origin of human populations in north africa. *PLoS ONE*, 8(11):e80293.

Fadhlaoui-Zid, K., Martinez-Cruz, B., Khodjet-El-Khil, H., Mendizabal, I., Benammar-Elgaaied, A., and Comas, D. (2011). Genetic structure of Tunisian ethnic groups revealed by paternal lineages. *American Journal of Physical Anthropology*, 146(2):271–280.

Ferembach, D. (1985). On the origin of the iberomaurusians (Upper palaeolithic: North Africa). A new hypothesis. *Journal of Human Evolution*, 14(4):393–397.

Font-Porterias, N., Solé-Morata, N., Serra-Vidal, G., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Calafell, F., and Comas, D. (2018). The genetic landscape of Mediterranean North African populations through complete mtDNA sequences. *Annals of Human Biology*, 45(1):98–104.

Fox, E. J., Reid-bayliss, K. S., Emond, M. J., and Loeb, L. a. (2014). Next Generation : Sequencing & Applications. *Next generation, sequencing & applications*, 1(1):1–4.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

Fregel, R., Méndez, F. L., Bokbot, Y., Martín-Socas, D., Camalich-Massieu, M. D., Santana, J., Morales, J., Ávila-Arcos, M. C., Underhill, P. A., Shapiro, B., Wojcik, G., Rasmussen, M., Soares, A.

E. R., Kapp, J., Sockell, A., et al. (2018). Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proceedings of the National Academy of Sciences*, 115(26):6774–6779.

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., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*, 350(6262):820–822.

Gibbons, A. (2017). World's oldest Homo sapiens fossils found in Morocco. *Science*.

Gonzalez-Perez, E., Moral, P., Via, M., Vona, G., Varesi, L., Santamaria, J., Gaya-Vidal, M., and Esteban, E. (2007). The ins and outs of population relationships in west-Mediterranean islands: Data from autosomal Alu polymorphisms and Alu/STR compound systems. *Journal of Human Genetics*, 52(12):999–1010.

González-Pérez, E., Via, M., Esteban, E., López-Alomar, A., Mazieres, S., Harich, N., Kandil, M., Dugoujon, J.-M., and Moral, P. (2003). Alu insertions in the Iberian Peninsula and north west Africa–genetic boundaries or melting pot? *Coll Antropol*, 27(2):491–500.

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., et al. (2010). A draft sequence of the neandertal genome. *Science*, 328(5979):710–722.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10):1031–1035.

Grove, M., Lamb, H., Roberts, H., Davies, S., Marshall, M., Bates, R., and Huws, D. (2015). Climatic variability, plasticity, and dispersal: A case study from Lake Tana, Ethiopia. *Journal of Human Evolution*, 87:32–47.

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.

Haber, M., Mezzavilla, M., Bergström, A., Prado-Martinez, J., Hallast, P., Saif-Ali, R., Al-Habori, M., Dedoussis, G., Zeggini, E., Blue-Smith, J., Wells, R. S., Xue, Y., Zalloua, P. A., and Tyler-Smith, C. (2016). Chad Genetic Diversity Reveals an African History Marked by Multiple Holocene Eurasian Migrations. *American Journal of Human Genetics*, 99(6):1316–1324.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., and Wall, J. D. (2011). Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences*, 108(37):15123–15128.

Harich, N., Esteban, E., Chafik, A., Lopez-Alomar, A., Vona, G., and Moral, P. (2002). Classical polymorphisms in Berbers from Moyen Atlas (Morocco): Genetics, geography, and historical evidence in the Mediterranean peoples. *Annals of Human Biology*, 29(5):473–487.

Henn, B. M., Botigué, L. R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J. K., Fadhlaoui-Zid, K., Zalloua, P. A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C. D., and Comas, D. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, 8(1):e1002397.

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., and Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, 113(4):E440–E449.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodriguez-Botigue, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences*, 108(13):5154–5162.

Hershkovitz, I., Weber, G. W., Quam, R., Duval, M., Grün, R., Kinsley, L., Ayalon, A., Bar-Matthews, M., Valladas, H., Mercier, N., Arsuaga, J. L., Martinón-Torres, M., Bermúdez de Castro, J. M., Fornai,

C., Martín-Francés, L., et al. (2018). The earliest modern humans outside Africa. *Science (New York, N.Y.)*, 359(6374):456–459.

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., et al. (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512(7514):306–309.

Hsieh, P. H., Woerner, A. E., Wall, J. D., Lachance, J., Tishkoff, S. A., Gutenkunst, R. N., and Hammer, M. F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Research*, 26(3):291–300.

Hublin, J. J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., and Gunz, P. (2018). Author Correction: New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens (Nature (2017) DOI: 10.1038/nature22336). *Nature*, 558(7711):E6.

Hunt, C., Davison, J., Inglis, R., Farr, L., Reynolds, T., Simpson, D., El-Rishi, H., and Barker, G. (2010). Site formation processes in caves: The Holocene sediments of the Haua Fteah, Cyrenaica, Libya. *Journal of Archaeological Science*, 37(7):1600–1611.

Huphrey, L. T. (2003). *The Human Fossil Record Volume 1. Terminology and Craniodental Morphology of Genus Homo (Europe)*, volume 37. Wiley.

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

Jobling, M. A. and Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. *Nature Reviews Genetics*, 4(8):598–612.

Jou, W. M., Haegeman, G., Ysebaert, M., and Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88.

Kefi, R., Hechmi, M., Naouali, C., Jmel, H., Hsouna, S., Bouzaid, E., Abdelhak, S., Beraud-Colomb, E., and Stevanovitch, A. (2016). On the origin of Iberomaurusians: new data based on ancient mitochondrial DNA and phylogenetic analysis of Afalou and Taforalt populations. *Mitochondrial DNA Part A*, 29(1):147–157.

Kim, H. L., Ratan, A., Perry, G. H., Montenegro, A., Miller, W., and Schuster, S. C. (2014). Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature Communications*, 5(1):5692.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kröpelin, S., Verschuren, D., Lézine, A.-M., Eggermont, H., Cocquyt, C., Francus, P., Cazet, J.-P., Fagot, M., Rumes, B., Russell, J. M., Darius, F., Conley, D. J., Schuster, M., von Suchodoletz, H., and Engstrom, D. R. (2008). Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science (New York, N.Y.)*, 320(5877):765–8.

Lachance, J. and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, 35(9):780–786.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J. M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M., and Tishkoff, S. A. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, 150(3):457–469.

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., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., et al. (2008). Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18(16):1241–1248.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453.

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

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

Linstädter, J. and Kehl, M. (2012). The Holocene archaeological sequence and sedimentological processes at Ifri Oudadane, NE Morocco. *Journal of Archaeological Science*, 39(10):3306–3323.

Litvinov, S. S. and Khusnutdinova, E. K. (2015). Current state of research in ethnogenomics: Genome-wide analysis and uniparental markers. *Russian Journal of Genetics*, 51(4):418–429.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *The American Journal of Human Genetics*, 79(2):230–237.

Liu, W., Martinón-Torres, M., Cai, Y. J., Xing, S., Tong, H. W., Pei, S. W., Sier, M. J., Wu, X. H., Edwards, R. L., Cheng, H., Li, Y. Y., Yang, X. X., De Castro, J. M. B., and Wu, X. J. (2015). The earliest unequivocally modern humans in southern China. *Nature*, 526(7575):696–699.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

Marciniak, S. and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, 18(11):659–674.

McDougall, I., Brown, F. H., and Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433(7027):733–736.

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

Mendel, G. (1866). Versuche über Pflanzenhybriden. *Versuche über Pflanzenhybriden*, 4:3–47.

Mendez, F. L., Krahn, T., Schrack, B., Krahn, A. M., Veeramah, K. R., Woerner, A. E., Fomine, F. L. M., Bradman, N., Thomas, M. G., Karafet, T. M., and Hammer, M. F. (2013). An African American paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree. *American Journal of Human Genetics*, 92(3):454–459.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46.

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., and Pääbo, S. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, 531(7595):504–507.

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226.

Molak, M. and Ho, S. Y. W. (2011). Evaluating the Impact of Post-Mortem Damage in Ancient DNA: A Theoretical Approach. *Journal of Molecular Evolution*, 73(3-4):244–255.

Moore, G. E. (2006). Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Newsletter*, 20(3):33–35.

Morales, J., Pérez-Jordà, G., Peña-Chocarro, L., Zapata, L., Ruíz-Alonso, M., López-Sáez, J. A., and Linstädter, J. (2013). The origins of agriculture in North-West Africa: Macro-botanical remains from Epipalaeolithic and Early Neolithic levels of Ifri Oudadane (Morocco). *Journal of Archaeological Science*, 40(6):2659–2669.

Muehlenbein, M. P., editor (2010). *Human Evolutionary Biology*. Cambridge University Press, Cambridge.

Mulazzani, S., Belhouchet, L., Salanova, L., Aouadi, N., Dridi, Y., Eddargach, W., Morales, J., Tombret, O., Zazzo, A., and Zoughlami, J. (2016). The emergence of the Neolithic in North Africa: A new model for the Eastern Maghreb. *Quaternary International*, 410:123–143.

Nespoulet, R., El Hajraoui, M. A., Amani, F., Ben Ncer, A., Debénath, A., El Idrissi, A., Lacombe, J.-P., Michel, P., Oujaa, A., and Stoetzel, E. (2008). Palaeolithic and Neolithic Occupations in the Témara Region (Rabat, Morocco): Recent Data on Hominin Contexts and Behavior. *African Archaeological Review*, 25(1-2):21–39.

Newman, J. L. (1995). *The peopling of Africa : a geographic interpretation*. Yale University Press.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451.

Novembre, J. and Ramachandran, S. (2011). Perspectives on Human Population Structure at the Cusp of the Sequencing Era. *Annual Review of Genomics and Human Genetics*, 12(1):245–274.

Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D. M., Dugoujon, J.-M., Coudray, C., Santachiara-Benerecetti, A. S., Semino, O., Bandelt, H.-J., and Torroni, A. (2006). The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science (New York, N.Y.)*, 314(5806):1767–70.

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., et al. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *American journal of human genetics*, 96(6):986–91.

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Der Veen, L. V., Hombert, J. M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., and Quintana-Murci, L. (2009). Inferring the demographic history of

African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics*, 5(4):e1000448.

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

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., and Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, 91(2):275–292.

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

Pereira, L., Silva, N. M., Franco-Duarte, R., Fernandes, V., Pereira, J. B., Costa, M. D., Martins, H., Soares, P., Behar, D. M., Richards, M. B., and Macaulay, V. (2010). Population expansion in the North African Late Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evolutionary Biology*, 10(1):390.

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

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P. R., Lachance, J., Mountain, J., Bustamante, C. D., et al. (2012). The genetic prehistory of southern Africa. *Nature Communications*, 3(1):1143.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2013). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 111(7):2632–7.

Pickrell, J. K. and Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967.

Pimenta, J., Lopes, A. M., Comas, D., Amorim, A., and Arenas, M. (2017). Evaluating the neolithic expansion at both shores of the mediterranean sea. *Molecular Biology and Evolution*, 34(12):3232–3242.

Plagnol, V. and Wall, J. D. (2005). Possible ancestral structure in human populations. *PLoS Genetics*, preprint(2006):e105.

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., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371.

Rahmani, N. (2004). Technological and Cultural Change Among the Last Hunter-Gatherers of the Maghreb: The Capsian (10,000-6000 B.P.). *Journal of World Prehistory*, 18(1):57–105.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–7.

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., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–762.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., et al. (2010). Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*, 468(7327):1053–1060.

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., Fernandes, P., Raynal, J. P., Geraads, D., Ben-Ncer, A., Hublin, J. J., and McPherron, S. P. (2017). The age of the hominin fossils

from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, 546(7657):293–296.

Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T. H., MacCallum, M., Arsuaga, J. L., Dobney, K., Valdiosera, C., Jakobsson, M., Götherström, A., and Girdland-Flink, L. (2017). Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Current Biology*, 27(21):3396–3402.e5.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51.

Sánchez-Quinto, F., Botigué, L. R., Civit, S., Arenas, C., Ávila-Arcos, M. C., Bustamante, C. D., Comas, D., and Lalueza-Fox, C. (2012). North African Populations Carry the Signature of Admixture with Neandertals. *PLoS ONE*, 7(10):e47765.

Scerri, E. M., Thomas, M. G., Manica, A., Gunz, P., Stock, J. T., Stringer, C., Grove, M., Groucutt, H. S., Timmermann, A., Rightmire, G. P., D'Errico, F., Tryon, C. A., Drake, N. A., Brooks, A. S., Dennell, R. W., et al. (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology & Evolution*, 33(8):582–594.

Scerri, E. M. L. (2017). The North African Middle Stone Age and its place in recent human evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, 26(3):119–135.

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., and Jakobsson, M. (2017). Southern African ancient

genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358(6363):652–655.

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

Schuenemann, V. J., Peltzer, A., Welte, B., van Pelt, W. P., Molak, M., Wang, C.-C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K., Teßmann, B., Francken, M., Harvati, K., Haak, W., Schiffels, S., et al. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications*, 8:15694.

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature*, 463(7283):943–947.

Secher, B., Fregel, R., Larruga, J. M., Cabrera, V. M., Endicott, P., Pestano, J. J., and González, A. M. (2014). The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC Evolutionary Biology*, 14(1):109.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., et al. (2017). Reconstructing Prehistoric African Population Structure. *Cell*, 171(1):59–71.e21.

Skov, L., Hui, R., Hobolth, A., Scally, A., Schierup, M. H., and Durbin, R. (2018). Detecting archaic introgression without archaic reference genomes. *bioRxiv*, page 283606.

Smith, T. M., Tafforeau, P., Reid, D. J., Grun, R., Eggins, S., Boutakiout, M., and Hublin, J.-J. (2007). Earliest evidence of modern human life history in North African early Homo sapiens. *Proceedings of the National Academy of Sciences*, 104(15):6128–6133.

Solé-Morata, N., García-Fernández, C., Urasin, V., Bekada, A., Fadhlaoui-Zid, K., Zalloua, P., Comas, D., and Calafell, F. (2017). Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81). *Scientific Reports*, 7(1):15941.

Sousa, V. and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, 14(6):404–414.

Steinrücken, M., Kamm, J. A., and Song, Y. S. (2015). Inference of complex population histories using whole-genome sequences from multiple populations. *bioRxiv*, page 026591.

Stone, W. H., Ely, J. J., Manis, G. S., and VandeBerg, J. L. (1993). Classical genetic markers and DNA markers: A commensal marriage. *Primates*, 34(3):365–376.

Stringer, C. (2000). Coasting out of Africa. *Nature*, 405(6782):24–27.

Stringer, C. and Galway-Witham, J. (2017). On the origin of our species. *Nature*, 546(7657):212–214.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Tishkoff, S. A., Gonder, M. K., Henn, B. M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T. B., Ramakrishnan, U., Reed, F. A., and Mountain, J. L. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular Biology and Evolution*, 24(10):2180–2195.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., et al. (2009). The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–1044.

Underhill, P. A. and Kivisild, T. (2007). Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics*, 41(1):539–564.

Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazón Lahr, M., Foley, R. A., Oefner, P. J., and Cavalli-Sforza, L. L. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics*, 65(1):43–62.

Van De Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., Amzazi, S., Hublin, J. J., Pääbo, S., Schiffels, S., Meyer, M., et al. (2018). Pleistocene north african genomes link near eastern and sub-saharan african human populations. *Science*, 360(6388):548–552.

Veeramah, K. R. and Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics*, 15(3):149–162.

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., et al. (2001). The Sequence of the Human Genome. *Science*, 291(5507):1304–1351.

Vernot, B. and Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, 343(6174):1017–1021.

Vona, G., Moral, P., Memmì, M., Ghiani, M. E., and Varesi, L. (2003). Genetic structure and affinities of the Corsican population (France): Classical genetic markers analysis. *American Journal of Human Biology*, 15(2):151–163.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Westaway, K. E., Louys, J., Awe, R. D., Morwood, M. J., Price, G. J., Zhao, J. X., Aubert, M., Joannes-Boyau, R., Smith, T. M., Skinner, M. M., Compton, T., Bailey, R. M., Van Den Bergh, G. D., De Vos, J., Pike, A. W., et al. (2017). An early modern human presence in Sumatra 73,000-63,000 years ago. *Nature*, 548(7667):322–325.

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

Wolf, A. B. and Akey, J. M. (2018). Outstanding questions in the study of archaic hominin admixture. *PLOS Genetics*, 14(5):e1007349.

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

Xu, D., Pavlidis, P., Taskent, R. O., Alachiotis, N., Flanagan, C., Degiorgio, M., Blekhman, R., Ruhl, S., and Gokcumen, O. (2017). Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. *Molecular Biology and Evolution*, 34(10):2704–2715.

Zhang, W. and Dolan, M. E. (2010). Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. *Pharmacogenomics*, 11(2):249–56.