UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1949

Human demographic history

Insights on the human past based on genomes from Southern through Central Africa

GWENNA BRETON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020

ISSN 1651-6214
ISBN 978-91-513-0979-8
urn:nbn:se:uu:diva-416653

Dissertation presented at Uppsala University to be publicly examined in Ekmansalen, EBC, Norbyvägen 14, Uppsala, Friday, 18 September 2020 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor David Comas (Universitat Pompeu Fabra, Barcelona, Spain).

Abstract

Breton, G. 2020. Human demographic history. *Insights on the human past based on genomes from Southern through Central Africa*. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1949. 112 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0979-8.

Evidence from paleontology, archaeology and population genetics support that modern humans originated in Africa. While the out-of-Africa event and subsequent colonization of all continents are well documented, human history in Africa at that time and before is less studied. Some current-day hunter-gatherer populations trace most of their genetic lineages to populations who inhabited Sub-Saharan Africa until the arrival of farming. They are informative about human history before and after the arrival of farming.

I studied high-coverage genomes from two such groups, the Khoe-San from Southern Africa and the rainforest hunter-gatherers from Central Africa. I generated a total of 74 genomes, significantly increasing the number of genomes from Sub-Saharan African hunter-gatherers. I compared several versions of a commonly used pipeline for high-coverage genomes and showed that using standard ascertained reference datasets has no significant impact on variant calling in populations from Sub-Saharan Africa. Using the full genome information, I described the genetic diversity in the Khoe-San and in the rainforest hunter-gatherers and showed that gene flow from agropastoralist groups increased the Khoe-San genetic diversity. I also detected a signal of population size decline in the Khoe-San around the time of the out-of-Africa event, and I evaluated the power of the method to detect bottlenecks by applying it to simulated data. I investigated the history of modern humans in Africa by estimating divergence times between populations and applying an Approximate Bayesian Computation analysis. We confirmed that the earliest divergence event was between the Khoe-San ancestral lineage and the rest of modern humans, ~250-350 kya. I also showed that the possibility of high gene flow should be incorporated in models of human evolution.

I furthermore examined SNP array data for two BaTwa populations from Zambia and showed that 20-30% of their autosomal diversity is hunter-gatherer-like. The estimated times for the admixture between a presumably local hunter-gatherer population and incoming agropastoralist groups are consistent with archaeological records.

In this thesis, I investigated questions related to human history in Sub-Saharan Africa, from the emergence of modern humans ~300 kya to recent events related to the expansion of farming.

Keywords: high-coverage genomes, Sub-Saharan Africa, demography, genetic admixture, population genetics, human evolutionary genetics, Khoe-San, rainforest hunter-gatherers

Gwenna Breton, Department of Organismal Biology, Human Evolution, Norbyvägen 18 A, Uppsala University, SE-752 36 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0979-8
urn:nbn:se:uu:diva-416653 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-416653)

Pour Amandine, qu'il me tarde de rencontrer,
et pour ma Mamie, qu'il me tarde de revoir!

List of papers

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

I Schlebusch, C.M.*, Sjödin, P.*, Breton, G.*, Günther, T., Naidoo, T., Hollfelder, N., Sjöstrand, A.E., Xu, J., Gattepaille, L.M., Vicente, M., Scofield, D.G., Malmström, H., de Jongh, M., Lombard, M., Soodyall, H., Jakobsson, M. (2020) Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in *Homo sapiens*. *Molecular Biology and Evolution*, in press.

II Breton, G., Johansson, A., Sjödin, P., Schlebusch, C.M., Jakobsson, M. (-) Comparison of sequencing data processing pipelines and application to underrepresented human populations. (*Under review.*)

III Breton, G., Sjödin, P., Zervakis, P.I., Laurent, R., Sjöstrand, A.E., Hewlett, B.S., Barreiro, L.B., Perry, G.H., Soodyall, H., Heyer, E., Schlebusch, C.M., Verdu, P., Jakobsson, M. (-) Deciphering early human history using Approximate Bayesian Computation and 74 whole genomes from Central and Southern Africa. (*Manuscript.*)

IV Breton, G., Barham, L., Mudenda, G., Soodyall, H., Schlebusch, C.M., Jakobsson, M. (-) The "BaTwa" populations from remote areas in Zambia retain ancestry of past forager groups. (*Manuscript*.)

*equal contribution

Contents

Abbreviations

ABC	Approximate Bayesian Computation
AD	Anno domini
aDNA	ancient DNA
AMH	Anatomically modern humans
bp	Base pair
BQSR	Base quality score recalibration
CAR	Central African Republic
CNV	Copy-number variation
DRC	Democratic Republic of the Congo
GATK	Genome Analysis Toolkit
HGDP	Human Genome Diversity Project
HTS	High-throughput sequencing
HWE	Hardy-Weinberg equilibrium
KS	Khoe-San
kya	Thousand years ago
LD	Linkage disequilibrium
ML	Maximum likelihood
mtDNA	Mitochondrial DNA
N_e	Effective population size
OOA	Out-of-Africa event
PC	Principal component
PCA	Principal component analysis
SFS	Site frequency spectrum
SGDP	Simon's Genome Diversity Project
SNP	Single nucleotide polymorphism
RHG	Rainforest hunter-gatherer
RHGn	Rainforest hunter-gatherer neighbour
ROH	Runs of homozygosity
TMRCA	Time to the most recent common ancestor
TT	Two-Two method
VQSR	Variant quality score recalibration
WF	Wright-Fisher (model)

1. Research aims

The aim of my thesis is to describe human genetic diversity with a focus on Sub-Saharan African populations and to use it to decipher the population history of modern humans. More specifically the aims were to:

I Prepare high-coverage genome data to answer questions of relevance in populations underrepresented in genetic databases, by surveying the literature and developing and testing a pipeline. This was done for the autosomes (Paper II) and for the sex chromosomes and mitochondrial DNA (Paper III).

II Obtain estimates of genetic diversity measures and describe relationships between populations using non-ascertained markers (Papers I and III).

III Investigate the deep population history in Sub-Saharan Africa, in particular the topology of the tree of modern humans, with high-coverage genomes from southern and central Africa (Papers I and III).

IV Describe genetic diversity in key populations that have not been sequenced earlier and explore their relationships with other populations and their history (Papers III and IV).

V Perform an Approximate Bayesian Computation analysis on entire genomes to infer the population history of sub-Saharan groups (Paper III).

To put these research aims into context, I first present different notions that are necessary to understand the general workflow of evolutionary genetics studies. I then highlight some of the aspects which are specific to working with human data, as compared to working with population genetic data in other organisms. Finally, I summarise the current state of knowledge about human diversity and demographic history, with a focus on Sub-Saharan Africa, and introduce the populations and questions that are investigated in this thesis. This is followed by the summary of my papers and concluding remarks.

2. Background on genetics and population genetics

Central to understanding biology and evolution is the propagation of heritable material from one generation to the next and the impact of heritable material on the individual. Population genetics and quantitative genetics are two research fields concerned with these questions. The focus of population genetics is to understand how genetic variation appears, is lost or becomes fixed in populations, and to use that knowledge to make inferences about the processes which have shaped, and continue to shape, populations. When one is concerned with humans, genetic diversity is not the only interesting aspect; a myriad of studies describe human diversity across the world, today and in the past, in terms of culture, language, or phenotype (any observable trait), to name a few. Describing and understanding this diversity is and has been the focus of much effort in various fields, including but not limited to anthropology, archaeology, ethnology, evolutionary genetics, linguistics and musicology; and it is particularly interesting to combine hypotheses and evidence from different fields to infer our past.

In this thesis, I investigate various aspects of human demography. I use here a population genetics definition of demography, understood as neutral processes that can affect the genetic diversity of a population, such as changes in population size, drift, and population structure. This is different but related to the more standard definition of demography as the study of populations' life history traits, via statistics such as birth and death rates. Natural selection is a non-neutral process which also affects genetic diversity: individuals' ability to survive and reproduce - and thus to contribute genetically to the next generation - depend on their phenotype. Demography and selection can both increase or decrease the genetic similarities between populations. In humans, culture in a broad sense, for example subsistence patterns or dominance relationships between human groups, also influences the patterns of genetic diversity (through natural selection for subsistence patterns, and mate choice *i.e.* demography for dominance relationships). These and other forces affect genomes in populations in different ways. By comparing genomic variation between different populations, we can learn about the past events that shaped it. The different steps in such a project are summarised in Figure 2.1; I will refer back to the different parts of the figure throughout this chapter.

Here I start with an introduction to genetics and population genetics. I then present different methods to generate genetic data, as well as tools and methods to describe genetic diversity and to infer demographic history (second, third and fourth boxes in Figure 2.1).

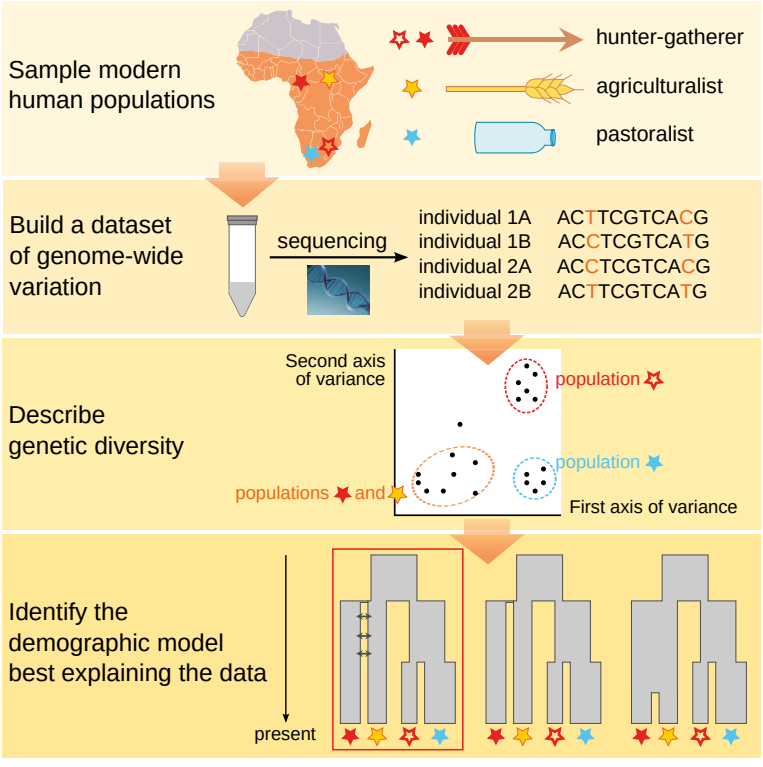

Figure 2.1. (Human) population genetics studies: a graphical abstract.

2.1 The bases of genetics, genome organization and population genetics

2.1.1 Historical vignettes

Two seminal works for understanding evolutionary biology were published in the 19th century: Charles Darwin's "The Origin of Species" in 1859 (Darwin, 1859) and Gregor Mendel's "Experiments in Plant Hybrids" in 1866 (Mendel, 1866). The first focused on selective forces shaping diversity while the second established the laws of Mendelian inheritance.

Darwin described many examples of variation within and between species to support the fact of evolution, and proposed a theory to explain it: natural selection, *i.e.* the idea that individuals with advantageous features have more offspring than other individuals and thus that the population changes from generation to generation (given some conditions: variation in the population and limited resources). His work has been refined, but many of the original ideas remain in modern theory. The accepted mechanism for heredity at that time was that parental traits were mixed and "diluted" in offspring ("blending heredity"); this represented a difficulty for Darwin, as it did not fit with his obser-

vations. This was to be solved with Mendel's work, that showed that heredity was based on discrete units that do not blend.

In modern terms, these units can be called **genes** or **loci** (singular locus). Different versions of a gene or a locus are called **alleles**. The **genotype** is the assortment of alleles at a locus, while the **phenotype** is an observable physical trait – such as eye color. One focus of genetics is to identify the genotypes underlying a specific phenotype.

Mendel formulated three laws, which describe Mendelian traits – that is, phenotypic traits that are dependent on a single gene (for example, **lactase persistence**, the ability to digest lactose as an adult). The "Law of Segregation" states that each (diploid) individual has two genetic copies of each locus and that these copies segregate during meiosis (cell divisions that result in gametes, *i.e.* sperm and egg cells). Gametes have only one genetic copy of each locus and the offspring receives one copy from each of its parents. The "Law of Independent Assortment" states that alleles for different traits are inherited independently. That is only partially true (as some alleles are "linked" to others). The "Law of Dominance" has to do with the dominance of one allele over the other – for example in the case of lactase persistence, the allele "digester" is dominant compared to the "non-digester" allele and an individual with a single "digester" allele will express the phenotype "digester". However, not all traits show this type of dominance. Many traits are not Mendelian and are called "complex traits". Height (or stature) is such a trait. It was shown that thousands of variants explain a little bit of the variation in adult height, which is also influenced by the environment, such as diet (Yengo et al., 2018).

2.1.2 An introduction to the (human) genome

Deoxyribonucleic acid, or **DNA**, was shown to be the heritable material in 1944 (Avery et al., 1944) and its double-helical structure was described less than ten years later (Watson and Crick, 1953). DNA consists of a sequence of molecules called "nucleotides" (often designated as one of four letters: A, T, C or G). DNA exists mostly as a paired double strand, hence the unit **base pair** or "bp" (a base pair is either "A-T", "T-A", "C-G" or "G-C").

In eukaryotes – organisms whose cells contain a nucleus -, DNA is organised in chromosomal units, which are more or less condensed during the cell cycle and are located in the nucleus. How exactly the genetic information is organised varies across eukaryotic species. In humans, there are 23 pairs of chromosomes in almost all cells; 22 of these pairs are two versions of highly similar chromosomes. Humans are **diploid** (two copies of the chromosomes) and chromosomes 1 to 22 are the **autosomes**. The 23rd pair of chromosomes corresponds to the sex chromosomes; females have two X chromosomes while males have one X and one Y chromosome. Additionally, the mitochondria, which is involved in essential metabolic activities and is present in many copies

in human cells, also contains DNA (the mitochondrial DNA, **mtDNA**). Together, the autosomes, the sex chromosomes and the (very short) mtDNA form the "human genome". The total length of a single set of autosomes is about 3 billion base pairs. Chromosomes vary in size and composition; chromosomes 1 and 2 are about five times larger than the shortest chromosomes, 21 and 22.

While a child receives one copy of each of the 23 chromosomes from her/his biological mother and one from her/his biological father, the mtDNA is inherited from the maternal side. Moreover, a Y chromosome can only be inherited from the biological father. These monoparental loci are called **haploid** (a single variant is inherited from the parents). One characteristic of the uniparental loci is that they do not recombine (apart from the pseudoautosomal regions on the Y chromosome). **Recombination** is a process in which the two chromosomes of a pair exchange fragments; it results in reshuffling of the genetic information. Because of the differences in transmission and the presence or absence of recombination, the autosomes, the sex chromosomes, and the mtDNA carry information across generations in a different way: Y chromosomes represent the male lineage, mtDNA the female lineage. The X chromosome is not exclusive to one of the sexes, but it retains different information than the autosomes: two-thirds of the X chromosomes are in females and thus the X chromosomes spend two-thirds of their evolutionary time in females.

2.1.3 Genetic variation

The DNA contains genes, most of which code for proteins, as well as non-coding regions, some of which are important for the regulation of gene expression. In humans, $\sim$2% of the genome is occupied by protein-coding genes (Jobling et al., 2013). Variation in protein-coding genes, but also in other genetic regions, impacts the phenotype of an individual. While a gene generally corresponds to several hundreds of bases, a locus can correspond to a single base.

When, at a given position, a chromosome has an "A" while another has a "C", we call it a **SNP**: "single nucleotide polymorphism". A **genetic polymorphism** is a locus for which there are at least two variants. It is biallelic if there are two alleles and multiallelic if there are more than two. Insertions and deletions - together called indels - are another type of variant in which an allele has one or several base pairs inserted or deleted compared to another allele at the same locus. **Microsatellites** are short sequences of nucleotides which are repeated in a row; the exact number of repeats differ between alleles. **Structural variation** designates large scale (larger than 1,000 base pairs) genome rearrangements, such as inversions or copy-number variation (**CNV**). CNVs are regions of the genome which are repeated and the number of repeats varies between individuals. An example is the amylase gene (*AMY1*) in humans; the amylase is an enzyme present in the saliva and involved in the degradation of

starch, and it has been shown that the number of copies of the amylase gene varies between populations and that it is higher in populations who traditionally had a starch-rich diet (Perry et al., 2007).

New variants appear by processes such as mutation, recombination or duplication, for example during DNA replication (when DNA is copied). In **the infinite sites model**, a commonly used model in the population genetic field, a new mutation produces a new allele: the assumption is that there are enough mutable sites that two mutations will not occur at the same locus.

2.1.4 The fate of a mutation

A new mutation (a change in the DNA sequence) is at first present in a single copy in a population. If this new mutation is **neutral**, *i.e.* if it does not impact the fitness of its carrier (it has no effect on the number of fertile offspring of the carrier), the probability of the mutation to be transmitted to the next generation is equal to the probability of it not being transmitted (*i.e.* lost). If it is not neutral, *i.e.* it impacts the carrier's fitness, the chances of it being transmitted or lost are not equal anymore. If the mutation is transmitted, the same happens at the next generation. Depending on the number of offspring carrying the mutation, the number of copies of the new mutation increases or decreases in the population. Eventually, it can become fixed (all individuals have this specific mutation -or allele- at the locus) or lost.

New mutations are the prerequisite for genetic variation and genetic variation is the prerequisite for evolution. I will now present how population genetics can help us to understand genetic variation and evolution.

2.1.5 Population genetics in a nutshell

Population genetics studies how mutations appear, spread and go to fixation, or are lost in populations. Given a number of alleles at a generation, what will be the number at the next generation? And ultimately, will a new allele become fixed, or lost, or will it stay at intermediate frequencies in the population? Population genetics can be predictive: it makes predictions regarding the future of alleles. It can also be retrospective: which past processes could explain what we observe today? In this thesis, I focus on the retrospective aspect: I use methods based on population genetic principles to infer past histories.

The relationship between genotype frequencies and allele frequencies is a central concept in population genetics. At equilibrium, the Hardy-Weinberg principle describes this relationship. If we consider a biallelic locus with alleles "A" and "a", there are three possibly genotypes: an individual can have two "A" alleles (genotype AA), two "a" alleles (aa), or one of each (Aa). Genotypes AA and aa are **homozygous** while genotype Aa is **heterozygous**. We define p as the frequency of allele "A" in the population; because there are only two

alleles, the frequency of "a" is $q = 1 - p$. At Hardy-Weinberg equilibrium (**HWE**), the expected frequency of the genotype AA is $p \times p = p^2$; the expected frequency of the genotype aa is $q^2 = (1 - p)^2$; and the expected frequency of the genotype Aa is $2pq = 2p(1 - p)$. A common procedure is to compare the observed and the expected genotype frequencies at a locus; if they differ, it means that the population is not at equilibrium at this locus, which can be due to different reasons such as demographic or adaptive processes.

The field of population genetics started in the 1920s as a theoretical field, since no appropriate data was available at the time to test the formulated hypotheses. Consequently, population genetic models are often developed for "ideal" (in a mathematical sense) populations. The Wright-Fisher model (**WF model**) (Fisher, 1923; Wright, 1931), is an example of such a model. It is widely used to describe how allele frequencies change from generation to generation; in particular, it describes **genetic drift**. I mentioned earlier that Darwin's theory of evolution had been refined; in fact, it became clear that natural selection is not the only force shaping (genetic) diversity; there are other forces as well, such as mutation, recombination, migration and genetic drift. Genetic drift is the "random change of allele frequencies in populations of finite size" (Nielsen and Slatkin, 2013). The relative importance of selection and drift in shaping genetic diversity within and between species has been the subject of much debate. According to the neutral theory of molecular evolution (Kimura, 1968), selection alone is not sufficient to explain the patterns of genetic variation; in fact the majority of mutations are neutral or slightly deleterious.

In this thesis, I focus mostly on neutral variation, though selection provides important insights into the history of populations and species, including humans (Fan et al., 2016). By describing how allele frequencies change over time in a WF population, properties of genetic drift are exemplified; for example, given the same initial frequency, it takes less time for an allele to become fixed or lost in a population with a small number of individuals than in a larger population. The exact probabilities of, for example, the time to fixation, can be calculated. This is possible because the WF model makes simplifying assumptions about the population: haploid; constant and finite population size; random mating; no selection or no mutation; non-overlapping generations. Note that the model can be modified for diploid species with sexual reproduction. No natural population fulfills these assumptions. However, predictions based on the WF model are very useful for understanding natural populations. It is nevertheless important, now that data is available to test population genetic predictions, to keep in mind the assumptions and limitations of the different models and methods. For example, several models (not including the WF model) consider infinite population sizes and/or make predictions for an entire population. In reality, natural populations are not infinite in size, and it is very unlikely to sample all individuals of a natural population. It is common to use a sample, whose size is a tiny fraction of the actual population size.

An important way in which the WF model is used, is to define the **effective population size** (N_e). Given a natural population with a certain census size (number of individuals), the effective population size is the (census) size of a WF population with the same amount of genetic variation given the same mutation rate. (For Wright-Fisher populations, census and effective population size are identical.) In general, the census size is larger than the effective population size, as processes such as non-random mating reduce the number of individuals who effectively contribute to the next generation.

The **coalescent** (Kingman, 1982) is another key concept in population genetics. It quickly became central because it efficiently relates theory and data, which is more and more abundant (see Section 2.2 about obtaining genetic data). The concept of the coalescent is powerful because it functions backward in time; given two lineages observed at present, it gives the probability that these two lineages coalesce – *i.e.* have a common ancestor – a given number of generations ago. It is very convenient because, contrary to forward-in-time approaches, one does not have to consider an entire population to make predictions; one can derive results from a sample of a few individuals. Consequently, the coalescent leads to very efficient algorithms for simulating genetic data. The simulation softwares used in Papers I and III are based on such algorithms. By looking at the coalescent tree of populations undergoing events such as reduction or expansion of the population size, one can more easily understand how some quantities of interest vary.

Genetic diversity is often described by **summary statistics**, quantities which have been defined in a theoretical framework (such as the WF model) and which are informative about different aspects of a population. In order to test, for example, whether a population had a constant population size in the past, the values of summary statistics calculated in a natural population can be compared to the expected values obtained in ideal populations under different models of evolution. Standardised summary statistics can also be used to compare natural populations or species. Depending on the questions addressed and the type of data available, different statistics are more or less relevant. Some summary statistics that I used in my thesis are described in Section 2.3.

2.1.6 Telling things apart in population genetics

A recurring difficulty in population genetic analyses is that demographic processes (*e.g.* population declines, expansions, or founder effects) and selective processes (*e.g.* an advantageous derived allele rising in frequency) have similar effects on the genome, *e.g.* a loss or a gain of genetic diversity (Li et al., 2012). Thus, the same deviation of summary statistics from expectations under a specific model can be due to different processes. One way around this issue is to combine summary statistics that differ in sensitivity to demographic and selective processes. Another way is to consider the scale of the effect: signatures

of selection are local (around the locus under selection) while demography impacts the genome more generally. In the past, much effort has been put into selecting neutral regions for demographic inferences (Rosenberg et al., 2002; Verdu et al., 2009). Today we have easier access to genome-wide data, and thus the issue might seem less relevant. However many recent studies have shown that the situation is not solved, with *i)* increasing evidence that most of the selection processes do not occur as "strong sweeps" (where a single, new advantageous allele rises to high frequency) but rather from standing variation (where one or several alleles rise in frequency from a pool of existing alleles) and *ii)* evidence of "long-range" effects of selection (*i.e.* it is difficult to find regions of the genome which are truly neutral) (Hernandez et al., 2011; Torres et al., 2018).

Another difficulty is to untangle parameters. One example is given by the parameters "divergence time" and "gene flow". We can consider the simple case of two populations that have diverged some time in the past, at T_{div} and that have a given amount of genetic differentiation today. If the two populations did not exchange any migrants since they diverged, all the generations since divergence have contributed to their differentiation. But if they have exchanged migrants, this has made the two populations more similar again. Consequently, for a given amount of genetic differentiation, T_{div} will be most recent in the absence of gene flow, and will increase with the amount of gene flow. If we do not consider gene flow (and it has, in fact, happened), we will underestimate the divergence time.

2.2 How to read genetic information

This is the step in a population genetic study where a dataset of genome-wide variation is built from DNA samples (Figure 2.1). This can be done in several ways.

2.2.1 Some history

The ABO blood group system is the first described example of a human phenotypic variation directly linked to an underlying genetic polymorphism. It was discovered in 1900 based on precipitation of serum and blood samples (Landsteiner, 1900). Numerous other polymorphisms were described afterwards, in particular using protein electrophoresis (different proteins migrating at different speed in an electric field) or restriction enzymes (enzymes cutting the DNA at a particular sequence of nucleotides). Other major advances were the invention of Sanger sequencing in 1977 (Sanger et al., 1977), which allowed to decipher the sequence of nucleotides, and the polymerase chain reaction (PCR), a method to amplify DNA, in 1985 (Mullis et al., 1986). The

first genome to be sequenced, in 1980, was that of a virus, the φX174 bacteriophage. A combination of different techniques culminated in the publication of two draft human genomes in 2001, the publicly founded "mosaic" reference genome (Lander et al., 2001), and a private initiative led by J. Craig Venter (Venter et al., 2001). Since then, whole genomes have been sequenced from a large number of organisms.

Once a reference sequence is available for one individual of a species, it is possible to describe the diversity of the species by comparing other individuals (of the same species) to the reference. By counting **reference** and alternative alleles one can quantify how different a given sample is from the reference. An **alternative allele** is a position of the genome where the new sample is different from the reference. This is different from the ancestral / derived duality, which has an evolutionary dimension and requires knowledge of the ancestral state of each position, *i.e.* which nucleotide was present in the ancestor of the species. For a diploid variant, the **minor allele frequency** is the frequency of the rarer of the two alleles. Interestingly, once more polymorphisms became available for humans, it was shown that some of the first polymorphisms ever described, like the ABO system, allowed for an accurate description of human worldwide diversity (Lewontin, 1972). It is worth noting that differences to a reference genome are relative, as reference genomes are not equally related to all representatives of a species.

2.2.2 SNP arrays

Following the HapMap Project (Gibbs et al., 2003), whole-genome **SNP arrays** (or chips) were developed by various companies. Based on the human reference and on known polymorphisms, markers are chosen along the genome and chips are constructed with specific primers for each marker. After binding to the right primer, the fragment containing the position of interest is sequenced. SNP arrays represent an affordable way to describe diversity in a dataset and are widely used. However they have limitations. First, the specific marker that you are interested in is often not on the array (but it can be inferred if it is in linkage with a marker on the array, *i.e.* it is generally inherited together with a marker on the array). Second, SNP arrays suffer from ascertainment bias: the polymorphisms are *a priori* chosen based on a predefined and limited set of individuals (often, but not always individuals of European descent), leading to an over-representation of the genetic variation from the population used as a reference, and therefore under-representing variation from other populations. It is a problem for African populations in particular as a lot of their variation is missed, and thus variation is underestimated in Africa (Lachance and Tishkoff, 2013). This in turn reduces the power of analyses (Bergström et al., 2020). Third, SNP arrays are mostly meant, as their name suggest, to study SNPs; but as we have seen earlier, there are many other types of ge-

netic variation. The first limitation can be partially overcome with a technique called "imputation" (described later in this subsection). The development of SNP arrays that target variation in populations of diverse ethnic background, such as the Human Origin (Patterson et al., 2012) or the H3Africa (Ramsay et al., 2016) arrays (the latter used in Paper IV), address the second issue and allow for better estimates of genetic diversity. Nevertheless, some regions, such as the mitochondria, remain poorly described by SNP arrays (Paper IV, Lankheet et al. *in prep.*). While SNP arrays remain popular in population genetics because of their low cost, which allows for generation of information for sample sizes of tens or hundreds of individuals in a population, as well as the availability of many comparative datasets, other techniques are becoming increasingly popular and are described in the next paragraphs.

2.2.3 High-throughput sequencing methods

From 2004 onward, several new sequencing technologies, here referred to as **high-throughput sequencing methods** (HTS) (but often called "next-generation sequencing", NGS) were developed (reviewed in (Morey et al., 2013; Reuter et al., 2015; Goodwin et al., 2016)). These techniques rely on the fragmentation of the genome into short fragments (often in the order of a hundred base pairs) and on the fixation of these fragments to a support before parallel sequencing using different methods. They permit an enormous increase of genome sequencing and a steep decline of the cost of sequencing. Moreover, because (almost) the entire genome is sequenced, these methods alleviate the issue of ascertainment bias. Parts of the human genome that remain difficult to access are the highly repetitive regions, such as the centromeres and the telomeres (middle and extremities of chromosomes). Other regions prove hard to assemble once sequenced such as Y-chromosme palyndromic regions or HLA regions.

Illumina technology (Illumina/Solexa company) is one of the most widely used techniques. It is based on ligation of short fragments to a glass slide, followed by amplification and sequencing with fluorescent nucleotides (which allows to determine the sequence) (Reuter et al., 2015). It outputs billions of reads, *i.e.* short strings (usually of one or a few hundred base pairs) of nucleotides. The HiSeq X Ten system, which was used for most sequencing in this thesis, is specifically tuned for sequencing human genomes at a coverage - or depth - of 30 X (*i.e.* each position of the autosomal genome is covered by an average of 30 reads). A sample sequenced at 30 X or more is considered a "high-coverage" genome; coverages around 4-8 X (or less) are "low-coverage". While the latest Illumina machine, the NovaSeq Series, can sequence a human genome in an hour, the raw data needs processing before it can be used for analyses.

Ongoing technological challenges focus on sequencing from single DNA molecules or on sequencing longer fragments (Reuter et al., 2015), which would allow for better calling of insertions, deletions and structural variants and alleviate the issue of phasing (discussed below). Moreover, while the cost of sequencing a human genome has decreased tremendously, associated costs such as data processing and storage, are often overlooked (Lightbody et al., 2019). Another practical issue is the standardization of processing pipelines and merging data generated with different sequencing platforms and processing pipelines (Baichoo et al., 2018; Regier et al., 2018).

2.2.4 An introduction to processing HTS data

A FASTQ file - the common format for raw sequencing data – contains several lines for each read, including a sequence of nucleotides and a quality line, where each position from the sequence has a base quality score which gives an indication on how certain it is that a specific nucleotide has been sequenced correctly (Cock et al., 2010). In this section, I will describe the main steps and file formats used for discovery of SNPs and indels in HTS data from human individuals. While a similar pipeline can be applied to other well characterised species (**model organisms**), specific steps require reference datasets that are not available for most species (see Section 3.2). Other features of the genomes, such as large structural variants or CNVs, require specific procedures which I will not discuss.

The processing pipeline generally begins with aligning the reads stored in the FASTQ to a reference genome, for example with the software bwa (Li and Durbin, 2009). This is the "mapping step": the reads are attributed a position in the reference genome. Optional steps such as adapter removal (adapters are short sequences added to the DNA fragments prior to sequencing) can be performed prior to mapping. After mapping, instead of many short reads, one obtains longer stretches of sequence. bwa outputs files in SAM (Sequence Alignment/Map) format (typically compressed in the binary format BAM) (Li et al., 2009). A SAM file consists of a header section containing various information and an alignment section containing the reads' sequences together with, among other information, the coordinates where they map in the reference genome, a mapping quality for each read, and a string describing how each position of each read aligns to the reference.

Before variants can be identified ("variant calling"), several steps have to be performed on BAM files. The Genome Analysis Toolkit (GATK) (DePristo et al., 2011) is a collection of tools and guidelines developed at the Data Sciences Platform at the Broad Institute which enable the processing of BAM files and calling of variants. In particular, the GATK Best Practices (DePristo et al., 2011; Van der Auwera et al., 2013) are standard workflows in the analyses of HTS data. Paper II focuses on applying the "Germline short variant dis-

covery (SNPs + Indels)" workflow to populations underrepresented in genetic datasets, while data prepared following different versions of this workflow is analysed in Papers I and III. Some steps commonly included are marking of duplicate reads (*i.e.* reads which correspond to the sequencing of the same DNA fragment and thus carry redundant information); realignment around indels (the mapping of reads close to common indels is improved); and a recalibration of the base quality score ("BQSR").

Variant calling is performed on fully processed BAM files. In this thesis, variant calling was performed with two different GATK variant calling tools, "UnifiedGenotyper" for Paper I and "HaplotypeCaller" for Papers II and III. An advantage of HaplotypeCaller is that it calls SNPs and indels simultaneously and performs local realignment in regions of the genome that seem to contain variants. It is also built to facilitate the addition of new samples. Other variant callers exist such as FreeBayes (Garrison and Marth, 2012) and SAMtools mpileup (Li et al., 2009). For a review of alignment and variant calling algorithms, see (Mielczarek and Szyda, 2016).

The output format of variant callers is a VCF file (Variant Call Format) (Danecek et al., 2011). VCF files, like SAM files, consists of a header and a data section. The data section can contain records for *i)* each position in the genome, an "all sites VCF"; or *ii)* only for variant positions. Besides the position and alleles, each record line contains a quality score and a filtering field, as well as various annotations. Often, variant callsets have to be filtered before they are analysed. Indeed, variant callers are very sensitive and some of the variants are false positives. Filtering can be done with GATK tools. Common filtering steps include filtering based on: variant annotations (there are two main approaches described in the papers of this thesis, hard filtering and a GATK's specific approach, Variant Quality Score Recalibration - "VQSR"); genotyping missingness; minor allele frequency; departure from HWE. Entire regions which are known to be difficult to sequence can also be filtered out. The choice of filtering steps depends on the purpose of the project; for example, medical studies are often concerned about false positives and filter the callsets heavily, while studies aiming at describing human diversity avoid a too stringent filtering since it would typically bias diversity estimates downwards.

2.2.5 Further processing of genome-wide variant data

I have focused on (pre-)processing of HTS data, as it is a more complex process than processing SNP array data. This section presents some processing steps which can be applied to variant data, whether it comes from a HTS or a SNP array study. Note that pre-processing steps are necessary for SNP array data as well; Paper IV gives a pipeline example. The callset filtering steps described in the preceding section resemble the filtering steps applied to SNP array data.

In humans, each autosomal position is present in two copies, one from each parent. Variants at different genomic positions that are inherited together represent a **haplotype**. In the absence of recombination -that breaks down haplotypes- and mutations -that create new haplotypes-, each chromosome would represent a haplotype. In humans, there is about one recombination event by chromosomal arm per generation (Dumont and Payseur, 2008). Consequently, haplotypes are shorter than (autosomal) chromosomes. Moreover selection forces affect haplotypes. For instance, in the case of strong positive selection (selective sweep), markers close to the variant under selection will tend to be inherited together with the selected variant (hitchhiking process), which leads to a local loss of diversity and a long haplotype (Wollstein and Stephan, 2015). Markers that are inherited together are in **linkage disequilibrium** (LD).

Knowing the **phase** of a variant – *i.e.* which haplotype it occurs on – is important for a number of analyses. Unfortunately, the most commonly used HTS technologies do not provide that information: reads are too short to contain several variants (except for regions with high variant density). This information is not available from SNP array data either. As a result, variant data has to be phased. There are two main categories of phasing approaches. The first category is experimental phasing (reviewed in (Huang et al., 2017)). For example, some HTS technologies, such as PacBio (Rhoads and Au, 2015), output long reads encompassing several variants; haplotypes can be constructed by overlapping such reads. The drawbacks of these approaches are their cost and their higher sequencing error rate (compared to short read sequencing). However, it is likely that as technologies improve they will become more and more popular. Sperm-typing is another experimental alternative that produces phase directly as the germ cells are haploid. The second category is statistical phasing - some softwares are reviewed in (Miar et al., 2017). In this case, phasing occurs after the sequencing. It is based on the idea that within a population, there are clusters of linked variants. When choosing a phasing approach, several aspects have to be considered, such as computing time, accuracy of phasing and availability of appropriate reference datasets (if a reference dataset is needed). The latter aspect is particularly relevant for HTS data, as variants that are not present in the reference dataset will be more difficult to phase (see (Bergström et al., 2020) for an approach to circumvent that issue). In general, the quality of the inferences increases with the sample size (Porcu et al., 2013).

Statistical phasing is often performed together with **imputation** – that is, inferring variants which were not called (for example due to low coverage, or because the variants are absent from the SNP array). Again, the quality of the reference dataset (phased individuals with high density of markers) is instrumental in obtaining a good imputation. Consequently, it is easier to impute data in populations of European ancestry than of Sub-Saharan African ancestry. Imputation is particularly useful when combining samples from different datasets: comparisons are often restricted to the overlapping positions when merging with SNP arrays or even low-coverage genomes. If the non-

overlapping positions are imputed, it is possible to perform the analyses on a larger number of markers.

2.2.6 DNA from extant and ancient populations

Because genetic material is transmitted from parent to offspring, the genome of a given individual is informative about its ancestors. This is why genomes from extant individuals can be used to investigate past population history. Additionally, the last few decades have seen the development of ancient DNA (**aDNA**) sequencing techniques to investigate past populations history. aDNA designates DNA isolated from "ancient" specimens, who died in a more or less recent past. It provides a window into specific times (ranging from a couple of generations to hundred of thousands of years ago). It allows, for example, to describe the genetic diversity prior to a population migration event which might have partially or totally replaced local populations. aDNA is also very valuable in learning about species that are not extant today. A fascinating example is the aDNA obtained from an extinct human subspecies, named "Denisovan" after the cave where its remains were first found (Reich et al., 2010). In that case, aDNA was extremely valuable as the few remains (a phalanx, some teeth) were not sufficient to distinguish it from other human remains. It was later shown that modern humans, in particular in Oceania and Island South East Asia, have substantial (a few percent) Denisovan-like ancestry.

A number of challenges characterise aDNA studies. First, it is difficult to find good samples: both environmental conditions favoring the preservation of DNA and excavation efforts are necessary. Second, it is difficult and costly to sequence: *i)* the samples are often contaminated with extrageneous DNA, *ii)* the DNA is degraded into short fragments, and *iii)* physical, chemical and biological processes can have modified the sequence. Third, it is not entirely clear how the typical features of aDNA, such as short fragments, low coverage, sequencing errors and difficulty of obtaining diploid sequences impact the population genetic frameworks that are commonly used and have been developed for modern DNA. Moreover, it is rare that more than a few samples are obtained for a given site at a given time. This makes it difficult to apply population based analyses and to know whether the samples are representative of the population (but this is changing rapidly, with increasing number of genomes sequenced for a given time period and/or location, allowing for longitudinal studies, *e.g.* (Antonio et al., 2019)).

For these different reasons, studies based on modern DNA remain (and probably will remain) essential to our understanding of the past. This is particularly true with studies of human history in Sub-Saharan Africa, as human remains are relatively rare in Sub-Saharan Africa and successful aDNA studies even rarer (Vicente and Schlebusch, 2020).

2.3 Describing genetic diversity

Once in possession of a well-filtered set of variants, possibly phased and imputed, researchers want to extract information from it (third row in Figure 2.1). As a first step we want to find ways to "summarise" the data, to identify its most salient features. In this section, I present several ways to do that.

2.3.1 Summary statistics

Summary statistics can be used to capture different aspects of genetic diversity. They can be computed at many different levels: for an individual or a population sample; for the entire genome, for specific regions or by scanning the genome using sliding windows; or for specific functional categories, for example exons, introns or intergenic regions (as in Paper I). I will present examples of statistics that are calculated on independent SNPs and that I used in my thesis.

Heterozygosity measures the probability for a sample of two gene copies to have two different variants at a site. This probability is higher in populations with high genetic diversity (if the populations are in HWE). One way to estimate heterozygosity is to divide the number of heterozygous sites by the total number of sites. This is called observed heterozygosity (see examples in Paper I). Another measure of heterozygosity is expected heterozygosity. It is based on the Hardy-Weinberg equilibrium (Section 2.1), according to which the proportion of heterozygous individuals is $2p(1-p)$ where p is the frequency of an allele at a diploid locus. This quantity can be calculated locus by locus and averaged. It is possible to correct for sample size by multiplying the statistic by $\frac{n}{n-1}$ where n is the number of gene copies in the sample. This is the unbiased estimator of heterozygosity (Nei and Roychoudhury, 1974) and it is used for example in Paper III.

In the infinite sites model and for a population at equilibrium, expected heterozygosity is equal to Θ (theta), the "population mutation parameter". Θ represents the diversity in a population at mutation-drift equilibrium; it is a function of the effective population size N_e and of the mutation rate μ ($\Theta = 4N_e\mu$ for diploid loci). Expected heterozygosity is an **estimator** of Θ, *i.e.* a quantity measured from a sample to estimate a parameter that cannot be measured directly. If we have an estimate of the mutation rate and of Θ, it is possible to estimate the effective population size.

π, or the **mean number of pairwise differences**, or the **nucleotide diversity**, is another estimator of Θ. To calculate π in a sample, one has to compare sequences pairwise and count the number of differences between the sequences in each pair.

The **site frequency spectrum** (SFS) (or allele frequency spectrum) is a series of summary statistics ($n-1$ summary statistics for the unfolded SFS and approximately $\frac{n}{2}$ for the folded SFS, where n is the number of gene copies in

the sample). In the unfolded SFS (which requires knowledge of the ancestral state of the variants), each statistic is the count (or the proportion) of loci that have i derived alleles (and $n-i$ ancestral alleles). The folded SFS can be calculated on data for which the ancestral state is not known; in that case, each class represents the count or proportion of loci with a given minor allele frequency (from $\frac{1}{n}$ to $\frac{1}{2}$). The expected SFS under the ideal Wright-Fisher population for a given value of n is known and deviations from it can be used to formulate hypotheses regarding demographic or selective processes. The **joint SFS** can be calculated for two populations and used to infer demographic parameters (Excoffier et al., 2013). The **conditional SFS** is another derivative of the SFS where the SFS is calculated based on only some variants of a population, for example the variants that have the derived allele in another population (Durvasula and Sankararaman, 2020).

Tajima's D (D_T) is a summary of the SFS and an indicator of neutrality. Its formula is:

$$D_T = \frac{\pi - \frac{S}{a_n}}{\sqrt{Var(\pi - \frac{S}{a_n})}} \tag{2.1}$$

where π is the statistics described previously; S is the number of segregating sites; and a_n is the sum of the sequence $1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n-1}$ with n the number of gene copies in the sample. $\frac{S}{a_n}$, or Watterson's Θ, is yet another estimator of Θ (Watterson, 1975). In an ideal (*e.g.* a WF) population under the infinite sites model, the two estimators are equal and the value of D_T is zero. Deviations from zero suggest non neutral processes or a demographic history that is not well approximated by an ideal population model. A positive value of D_T indicates an excess of intermediate frequency variants (which contribute more to π than to S), that can be due to a population size contraction, population structure, or balancing selection. A negative value of D_T indicates an excess of rare variants (which contribute more to S) and can be due to a population size expansion, or positive selection.

F_{ST} is, in contrast to the statistics described so far, an inter-population summary statistic; it is an index of genetic differentiation between populations. Its most extreme values correspond to panmixia ($F_{ST} = 0$, the populations are mating randomly) and to complete genetic isolation ($F_{ST} = 1$). It is calculated between two or more populations by comparing the expected heterozygosity for the populations taken separately to the expected heterozygosity assuming that these two populations were in fact a single (well-mixed or "panmictic") population. It is based on allele frequencies in the populations; a practically speaking unobservable quantity but that can be estimated from a population sample. Larger sample sizes will give more accurate results.

I discuss one more category of summary statistics which is based on stretches of variants but do not require knowledge of the phase: **runs of homozygosity** (ROH) (Broman and Weber, 1999; Kirin et al., 2010). ROH analysis builds on

identifying stretches in the genome of an individual where all positions are homozygous. While short ROH can be due to chance, longer stretches are more likely to reflect phenomena such as inbreeding and/or small population size. The definition of ROH – for example the minimal length of a stretch to qualify as ROH, or the number of "mistakes" that one allows – depends on the type of data used. In Papers I, III and IV, I used different parameters depending on the type of data.

While each summary statistics has a precise definition, there may be various ways and formula to estimate each of them from real data, and it is important to keep track of the methods used in a given study. In this thesis, summary statistics have been computed with custom scripts, with the software plink (Purcell et al., 2007), or with Python scripts following (Jay et al., 2019) and references therein.

2.3.2 Visual exploration of genetic data

Another set of common approaches derives from multivariate statistics and aim at identifying the independent axes maximizing variation in a dataset (and to visualise it). Principal Component Analysis, or **PCA** (Pearson, 1901; Menozzi et al., 1978) is particularly popular in population genetics. Given a set of individuals and markers, running a PCA corresponds to finding the axes which maximise the variance of the sample. There are as many axes as there are markers (minus one); the first axis (PC1) explains the largest percentage of the total variance, PC2 is orthogonal to PC1 and explains the next largest percentage of total variance and so on. One advantage of PCA is that it does not need filtering of the callset besides basic quality filters (the variants do not need to be filtered for LD or minor allele frequency for example). PCA were popularised in the 90's by Luigi Luca Cavalli-Sforza in his book "The history and geography of human genes" (Cavalli-Sforza et al., 1994). More recently, Novembre et al. (2008) showed that the projections of the main axes of variation in a set of genetic markers correlated with geography for European populations. Ancient samples can be added to PC analyses; the common procedure in that case is to use reference modern samples to define the PC axis, and to "project" the ancient samples into the modern diversity.

An alternative approach to PCA is to do a Multidimensional Scaling Plot (MDS) based on a quantity of interest, for example pairwise allele sharing distances or F_{ST}. Depending on the chosen quantity, this method can be very fast. Moreover, once an interesting pattern is observed, it is possible to go back to the biologically meaningful statistics (which is not as straightforward for the vectors underlying PCA axes).

2.3.3 Clustering methods

Model-based clustering methods that infer clusters maximizing allele frequencies differences, like ADMIXTURE (Alexander et al., 2009) and STRUCTURE (Pritchard et al., 2000), are also very popular in human population genetics. Both infer genetic memberships for individuals to predefined numbers of virtual clusters (referred to as "K") based on genetic similarity. Based on the genetic membership proportions of individuals in different populations, it is possible to make hypotheses about the ancestry of the clusters. The results can be plotted in visually attractive plots that illustrate population relationships effectively, however these plots are often over-interpreted (Novembre, 2016; Lawson et al., 2018). Thus, when using such methods one has to: *i)* Use them correctly: run a sufficient number of repeats for each value of K, stop when new clusters are uninformative, for example when each modern population gets its own cluster, and *ii)* Be cautious with over-interpretation. Like PCA or MDS plots, these methods are merely a way to describe the data and formulate hypotheses, which next need to be formally tested. Alone, they should not be used to make inferences, for example about admixture events. Specific hypotheses about admixture can be further investigated with more direct tests of admixture, *e.g.* f-statistics (Weir and Cockerham, 1984; Wright, 1950).

2.4 Inferring demographic history

Various processes shape the demographic history of populations and species. Some demographic processes, like population splits and isolation-by-distance, contribute to separate lineages; others bring them together, for example through admixture events or repeated migrations. Another aspect of demography are changes in population size.

By combining descriptors of genetic diversity such as those described in Section 2.3 and knowledge of the population, it is possible to formulate hypotheses as to what demographic processes might be relevant for understanding a given observed pattern. Then one can apply methods to test these hypotheses and infer the value of parameters (for example, the past size of a population). Methods in population genetics are based on expectations of the genetic diversity (or on summary statistics of the diversity) under different models. Given observed data D, and a parameter of interest Θ, we are interested in the likelihood function, which gives the probability of the observed data given the parameter value. The higher the likelihood, the more likely it is that the parameter value is correct; thus, the goal of many methods is to find the **maximum likelihood** (ML). The limitation of such approaches is that one needs to explore the entire parameter space, *i.e.* compute, for each possible value of the parameter, the probability of the data. This is often done via simulations. It becomes very time-consuming for complex models (with large numbers of parameters), even if techniques such as Markov Chain Monte Carlo

allow efficient exploration of the parameter space. **Bayesian approaches** take a different angle with parameters having a probability distribution (in ML approaches parameters have a finite value). Based on prior knowledge, one defines a prior distribution for the parameter (for example, for population sizes, the prior distribution could be the range of sizes observed in natural populations). Then, one gathers data, or performs an experiment, and based on that obtains the posterior distribution of the parameter (*i.e.* for a range of values, the probability that the parameter takes a value given the data).

The following sections present different demographic processes and methods that aim to characterise them. Most of these methods are based on ML or on Bayesian approaches. In the end I introduce a flexible framework, Approximate Bayesian Computation, that can be applied to test any demographic process against a set of observed data by generating simulated data under possible scenarios. This can then be used to choose between scenarios and to estimate parameters (see Paper III).

2.4.1 Divergence

"Divergence" or "split" is the process in which an ancestral population gives birth to two or more populations which are subsequently genetically isolated from each other (they no longer reproduce together after the split). Ultimately, it can lead to speciation, *i.e.* to distinct species (with enough time and degree of isolation). Populations can diverge for various reasons: *e.g.* because they are on either side of geographical barriers (mountains, sea); because of changes in lifestyle resulting in groups occupying different niches; in humans, because of cultural rules defining groups of individuals allowed to reproduce together excluding other groups. It is rare that a population splits "clean" into two or more groups from one generation to the next without any subsequent contact between the populations, even if that is often assumed in models. Often, divergence is progressive; it might start with population structure, *i.e.* within a population, individuals reproduce preferably with some individuals. It is also common that diverging populations continue to exchange migrants after the split.

One way to characterise divergence is to estimate the **time to the most recent common ancestor** (TMRCA), *i.e.* the youngest individual who is an ancestor for all individuals in two (or more) populations. Note that estimating the TMRCA is an upper bound of the divergence time (*i.e.* the divergence time will in general be younger that the TMRCA) (Jobling et al., 2013). (Zhou and Teo, 2016) compared several methods which estimate the divergence time, such as MSMC (Schiffels and Durbin, 2014) and G-PhoCS (Gronau et al., 2011), by running them on simulated data for classical demographic models (such as divergence with and without migration). These methods differ in the type of data they use, how they summarise that data, and the underlying statistical frame-

work. One of the method giving the most accurate results was MSMC. Here I highlight two methods used in this thesis.

G-PhoCS (Gronau et al., 2011) (used in Paper I) is a coalescent-based approach that estimates ancestral population size and divergence times (migration rate estimates are also possible) by examining many ($\sim$30,000) relatively short sequences ($\sim$1,000 bp) in a sample. Each sequence is assumed to be sufficiently short to be in complete linkage (no recombination between sites in the sequence) and has some information about the genealogy, and G-PhoCS integrates that information across loci to estimate the parameters. One advantage of G-PhoCS is that is does not require the genomes to be phased.

The **Two-two** or **TT** method (Schlebusch et al., 2017; Sjödin et al., 2020) (used in Papers I and III) uses two diploid genomes, one from each of two populations. It makes the assumption that sites are independent (no linkage), and derives equations that relate observed variables (the counts of different configurations of ancestral and derived alleles in the two genomes) to parameters of interest: divergence time, ancestral population size and drift. Like G-PhoCS, it does not require phased data, but the knowledge of the ancestral state is required. It runs very fast, but how is it affected by migration is not trivial.

2.4.2 Migration

Migration is a process in which (sub-)populations exchange migrants; it is common in humans (Busby et al., 2016; Schlebusch and Jakobsson, 2018). Two important aspects of migration are its frequency (*e.g.* whether it happened once or at every generation) and its rate (which is proportional to the proportion of migrants in the population at a given generation). **Admixture** is a common consequence of migration where some offspring have parents from different sub-populations. **Gene flow** often designates the combination of migration and admixture. Gene flow can be the result of a single admixture event, at a given generation; or of a recurring process at every generation; it can be symmetrical or asymmetrical; it can be sex-biased (*e.g.* the migrants are only females). Depending on the relative size of populations and the migration and admixture rates, gene flow can even be similar to population replacement (for example if a population of small size receives a lot of migrants from a large population).

Hypotheses of gene flow can be made based on some of the descriptive approaches discussed previously, such as PCA or clustering analyses. The f-tests address different questions related to admixture and population structure (Patterson et al., 2012; Peter, 2016; Reich et al., 2009). They are based on allele frequencies, which are expected to differ between populations. The f_4 test, *e.g.*, is a formal test of admixture while the f_4 ratio (Green et al., 2010) estimates admixture fractions under a given population topology (*i.e.* the user has to make a hypothesis as to how the populations are related). More complex

methods can be applied to estimate parameters such as the time of the admixture event; whether there was one or several admixture event(s); and the admixture rates. **MOSAIC** (Salter-Townshend and Myers, 2019), used in Paper IV, is such an approach; it uses haplotype information to identify and characterise admixture events in a "target" population, by estimating the time of admixture, the fraction of the genome coming from each of the ancestries, and the populations that are the closest to the admixing populations. Compared to other methods, it does not require the user to specify sources for the admixture, sometimes called "parental populations". Moreover, it can estimate admixture coming from more than two sources. MOSAIC uses an approach called "chromosome painting", which assign local ancestries on chromosomes based on haplotype information and in particular patterns of LD. In dating admixture, the method uses how fast coancestry decays across the genome, building on the fact that recombination breaks down LD over time.

Another application of chromosome painting is "admixture masking", for example with RFMix (Maples et al., 2013). By selecting only the regions of the genome coming from a specific parental source and performing analyses on these regions, it is possible to recover patterns such as isolation by distance which are otherwise masked by recent admixture from immigrant groups (Vicente et al., 2019). We used this approach in Paper I.

2.4.3 Effective population size

In conservation genetics, the effective population size (defined in Section 2.1) of endangered species is closely monitored, as populations (or species) with small N_e may be more at risk of extinction when facing ecological, climatic, or anthropological threats. For humans, estimating past population sizes contributes to obtain a better picture of the past. Population size can be studied over time; some typical population size changes are contractions and expansions. A bottleneck is a population size trajectory in which the population size substantially decreased for a number of generations before it increased again. The "founder effect" is a consequence of an abrupt decrease of population size; only a fraction of the ancestral population – and thus of the ancestral genetic diversity - contributes to found the new population at some point in time. This is likely what happened for the humans expanding out of Africa throughout the world: a series of founding events (Ramachandran et al., 2005). Today, most human populations are growing exponentially in census size; however, since population sizes have been small for many generations, the N_e of human populations is still quite small (Jay et al., 2019).

A series of methods estimating N_e over time were published in the last decade; these methods are based on the coalescent, and in particular in estimating coalescent trees along the genome, and relating these patterns to population size. They estimate N_e in windows of time, and the succession of the

windows gives an indication of how the effective population size has varied over time. The first of these methods, **PSMC** (Li and Durbin, 2011), takes two haploid genomes as input. Due to the properties of the coalescent of two sequences, PSMC is not appropriate for studying recent times; a generalization of the method to more than two haploid genomes, **MSMC** (Schiffels and Durbin, 2014), solved this issue, with the drawback that it requires the data to be phased prior to the analysis, and that suboptimal phasing biases the results (Bergström et al., 2020). PSMC and MSMC are also used to estimate divergence times (Section 4.5). Other methods efficiently allow for incorporation of more genomes (Terhorst et al., 2017), or use SNP array instead of genomes (Browning and Browning, 2015). Nevertheless these methods suffer from not considering migration *a priori*, despite the fact that effective population size, divergence time and migration can be deeply intertwined and difficult to isolate (Section 2.1).

2.4.4 Approximate Bayesian Computation – a versatile framework

Approximate Bayesian Computation (ABC) is increasingly used in genetics over the last 20 years (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002). It allows the researchers to compare the fit of complex models to observed data, and to estimate parameters of the model. Taking a simple example, imagine that we have sequences from a population and that we are interested in reconstructing the past changes in population size of that population. We can imagine several scenarios that produced the observed genetic variation, for example a population expansion, contraction, or a constant population size. We need to define the parameters underlying the models (for example, the population size at various times) and decide on prior distribution for the parameters (*e.g.* we may think that the current effective population size is between 100 and 1000 individuals, and that each size in that range is equally likely – uniform distribution). We then generate (a lot of) simulated datasets for each model, drawing parameter values from the prior distributions. It is important to use prior information to ensure that the simulated datasets are similar to the observed dataset (for example, use a realistic mutation rate). We then calculate summary statistics on the observed and simulated datasets – we approximate the data. Based on these summary statistics, we can first select the model that has the highest posterior probability – it is more likely than the other models to have generated the observed genetic patterns. This step is called model selection. Under that model, we can calculate posterior distributions for the parameters (parameter estimation), by selecting the simulations that produce the closest summary statistics to the observed values.

ABC can be used to address any question for which we can formulate models and simulate data resembling the observations; however, generating simulated

datasets and calculating summary statistics represent a computational burden, in particular if one wants to simulate entire genomes. A recent methodological advance based on machine learning, **Random-Forest ABC** (Pudlo et al., 2016), alleviates part of this issue for the model selection step, as it drastically diminishes the number of simulations one needs to perform equally well if not better.

3. A study species: *Homo sapiens*

The tremendous advances in genomics in the last half century have demonstrated that all living beings share the same material of genetic information in the form of DNA (and in some cases, RNA) and that some genes are found in all life forms (in particular, genes involved in the copy and transfer of genetic information, as discussed in (Weiss et al., 2018)). Moreover, the population genetic models, often defined in ideal, haploid populations, can be adapted to account for other cases – for example, diploid species like mammals, or polyploid species like many plants. In short, human genetic diversity can be studied by applying the same principles that would be used to study other organisms. Nevertheless, *Homo sapiens* is a special case for several reasons; first, culture affects genetic diversity, and evidence from research fields such as linguistics and ethnology informs studies. Second, as a consequence of the central position of everything "human" in our societies, there is an abundance of resources for studies relating to humans, and consequently human geneticists have access to a wealth of data, reference datasets and analysis tools. Finally, studying human populations implies specific ethical considerations.

3.1 An interdisciplinary approach

3.1.1 Diversity of processes impacting the genome

Selection, migration, drift, mutation and recombination affect the human genome like the genome of any other species. However, culture is embedded in some of these processes – it is part of the environment of human populations. I use "culture" in a broad sense here, including language, techniques, structure of the society and diet. Selective events can be directly related to cultural factors; and cultural processes may also affect the entire genome, similar to demography. I will highlight some examples, generally relevant in the context of my thesis or particularly telling, to illustrate the interaction of culture and genetics. These examples concern relatively recent times (a few thousand years), as interpolating cultural processes in the more distant past is difficult.

I mentioned two examples of how genetic diversity can be impacted by culture in Section 2.1: the lactase persistence trait (the ability to digest the milk sugar, lactose, into adulthood) and the number of repeats of the amylase gene (related to the digestion of starch by salivary enzymes). The lactase persistence trait has been measured in many populations across the globe; it has a patchy

distribution, with particularly high frequencies in (northern) Europe (contrary to what the abundance of lactose-free products in Sweden suggests!) and in pastoralist populations in, for instance, eastern Africa. Several genetic variants associated with the lactase persistence trait were identified, and different hypotheses have been proposed to explain their distribution. Concerning Europe, strong positive selection for one specific variant, likely explained by a selective advantage associated with the adulthood consumption of milk and the cultural transmission of milk drinking practice, is the accepted model (though demographic processes might have intensified the effect). Lactase persistence is the textbook example of gene-culture coevolution, where genetic and cultural processes reinforce each other (Ségurel and Bon, 2017). Indeed, widespread adulthood consumption of fresh milk can only be achieved via herding in the first place. In Africa, there are several variants associated with lactase persistence; I explain later how this can inform us about the relationships between populations and the diffusion of cultural practices. The amylase gene is also related to diet; the number of copies of the gene varies in populations, and higher numbers of copies were found in populations with a starch-rich diet, such as Chinese and Japanese (Perry et al., 2007). Being able to efficiently break down starch must have been advantageous in these populations. The study of amylase CNVs has been refined in further studies, for example see (Carpenter et al., 2015).

Culture can also impact genetic diversity in less direct ways, for example by determining or regulating who reproduces with whom. This can result in barriers to reproduction; or increase the variance of reproductive success, if some individuals have more offspring than others. One example is the effect of patri- and matrilocality (Heyer et al., 2012). In matrilocal societies, it is the husband who moves to his wife's place of origin after marriage, and the children are raised there. Matrilocality is the exception rather than the rule in human populations today. It has been shown that patri- and matrilocality influence genetic diversity; in particular, it affects the diversity of the maternal lineage (the mtDNA) and of the paternal lineage (the Y chromosome, and in particular its non-recombining parts) (Seielstad et al., 1998; Oota et al., 2001; Marchi et al., 2017). In matrilocal societies, the within population mtDNA diversity is lower, because there is no income of new mtDNA lineages; the effective population size of the mtDNA is small and more prone to genetic drift; and the differentiation between populations is high for this genetic compartment. The reverse is true for the Y chromosome diversity (high diversity within population, low differentiation between populations).

A related phenomenon is **sex-biased admixture**. At a local scale, sex-biased admixture has been reported repeatedly in pairs of populations from central Africa, between rainforest hunter-gatherers (RHG) and neighbouring populations (Verdu et al., 2009, 2013; Patin et al., 2009). Ethnographic observations of such population pairs had shown that marriages occurred between hunter-gatherer females and neighbouring males, but rarely or never between

hunter-gatherer males and neighbouring females. Such patterns should result in hunter-gatherer female gene flow into neighbouring populations. Genetic evidence however was compatible with a different pattern: male gene flow from neighbouring populations into the hunter-gatherers. The two observations were reconciled in the following scenario: the neighbours' social discrimination against the hunter-gatherers often (if not always) results in hunter-gatherer females married to neighbouring males to divorce and eventually return to their birth family, with their children, who are then raised in the hunter-gatherer population, thus creating the observed pattern of gene flow. Another reason for the females to move back to the hunter-gatherer population is when their husbands from the neighbouring populations die. Sex-biased gene flow has also been suggested for larger scale migrations. For example during the Bantu expansion, discussed at more length in Section 4.3, local females were incorporated into incoming Bantu-speaking populations (and males to a lesser extent). This had consequences on the diversity of maternal and paternal lineages in sub-equatorial Africa (Wood et al., 2005; Schlebusch et al., 2011; Barbieri et al., 2014; Bajić et al., 2018).

Another illustration of the interplay between genetics and culture is provided by the study of agriculture-related expansions. Agriculture emerged in several regions of the world starting $\sim$10 kya and eventually became the subsistence pattern of most human populations today. How agricultural practices (and associated crop and animals) diffused is an important question. Two major hypotheses have been proposed: according to the demic diffusion hypothesis, the practices came along with a group of people who admixed or replaced local inhabitants; the cultural diffusion hypothesis (popular among archaeologists), on the other hand, implies that agricultural practices were learned by local inhabitants who changed their subsistence pattern. Provided that the incoming population (in the case of demic diffusion) is sufficiently different from local inhabitants, genetics can help to distinguish between the two hypothesis. In the case of demic diffusion, we expect genetic differences between the migrating agricultural populations and the local forager populations. In fact, it appears that in most cases, agriculture spread via demic diffusion (Stoneking, 2016). This brings us back to lactase persistence. A lactase persistence genetic variant typical from eastern Africa was found in a pastoralist Khoekhoe population from southern Africa, the Nama; furthermore, it was shown that the Nama have up to 20% ancestry from an eastern-African like population (including in the region surrounding the lactase persistence variant). This, associated with evidence from other fields, such as the date of the first evidence of sheep in southern Africa, or the type of sheep found in southern Africa, suggests that the spread of pastoralism from eastern to southern Africa was accompanied with a diffusion of genes (Breton et al., 2014; Macholdt et al., 2014; Schlebusch et al., 2017).

To conclude this section, I would like to note that I do not suggest that "culture" is unique to humans; other species use tools, or have complex social

structures impacting their genetic diversity, and several species of ants show remarkable examples of complex agricultural systems. Moreover, human culture impacts many other organisms; an obvious illustration is the domestication of plant and animal species. But I hope to have conveyed how applying population genetics to human populations can be exciting on various levels!

3.1.2 Evidence from other fields

Human population genetics is one of many sciences interested in understanding the human past and present. There is much to be gained from an interdisciplinary approach. The emergence and diffusion of agriculture, for example, had repercussions on the material culture, on languages and on genetic diversity. These different lines of evidence can be studied and *e.g.* used to date the arrival of a material culture in an area, or the linguistic divergence time. Combining these different views results in a more nuanced picture of population history.

Different lines of evidence have different time depth (Jobling et al., 2013). **Anthropology**, the study of humans, human behavior, and societies, is the science concerned with the most recent times. Like we have seen in the previous section, anthropological research is informative about patterns such as marriage rules, and can shed light on complex patterns such as the sex-biased admixture between RHG and their neighbours. By studying oral history, for example myths of origins, anthropology contributes to formulate hypotheses concerning the relationships between populations which can later be tested. Anthropology is also informative about generation time, which is an important parameter in all population genetic analyses estimating the timing of an event. (Fenner, 2005) estimated the generation interval for nation states and hunter-gatherer societies, and found an estimate of 28.6 to 30.1 years (averaged over the two sexes, as generation time is longer in males than in females for a majority of populations). When using these estimates to scale times of interest, we assume that generation time did not change too much over time.

History and **linguistics** have intermediate time depths. The earliest historical records date to $\sim$4 kya, and some linguists argue that historical linguistics can go back to $\sim$10 kya (Jobling et al., 2013). By studying shared features between languages, it is possible to formulate hypotheses about population relatedness and past contact. For example, the presence of clicks -sounds that are characteristic of languages collectively designated "Khoisan" and common among others in Southern Africa- in Bantu languages from Zambia suggests a link with Khoisan-speaking populations (Barbieri et al., 2013). Hypotheses concerning a group of populations' past interactions and possible common origin can be made based on sharing of specific vocabulary, for example concerning food, the environment, or social practices (for example the Baka and Aka RHG, and their interactions with farmers (Bahuchet, 2012)); and the patterns of

language sharing between RHG and neighbouring populations suggest migration networks (Paper III and (Bahuchet, 2012)). Although genes and languages are not transmitted over generations in the same way, they both tell the story of the same group of individuals. Therefore, combining the two lines of evidence for the same populations allows to better understand, for each population of interest, whether reproductive isolation may precede linguistic differentiation or the opposite (Thouzeau et al., 2017). Another interesting example is given by populations whose genetics and languages do not correspond to the common pattern, *e.g.* the Damara, who have a typical Bantu-speaker farmer genetics makeup but speak a Khoisan language (Pickrell et al., 2012).

Archaeology is the study of material remains that have been modified by humans, including tools and burials, but also less direct evidence such as soils or waste deposits. It goes back $\sim$2.5 million years, which makes it a precious tool for the more ancient human history. Archaeology can inform us about the diffusion of cultural practices such as agriculture, or give indication about how long populations with different cultures have interacted (*e.g.* hunter-gatherers and agriculturalists in Zambia – Paper IV). Physical remains can be dated, either in relation to each other or directly in years. For example, the estimates for the first remains of milk proteins in potteries can be compared to genetic based estimates of the arrival of pastoralism to different places (Lander and Russell, 2020). More recently, archaeology has become the provider of samples for ancient DNA studies.

The study of fossilised remains is called **paleontology**; it goes back even further than archaeology. Human fossils contribute essential evidence about the history of *Homo sapiens* (see Section 4.1), including the modalities and timing of the diversification of modern humans; migration; and interaction with other hominin species. Fossils of animals and plants are informative about human diet and interactions with the environment.

To sum up, the evidence from the fields that I described briefly is extremely helpful both to formulate hypotheses to be tested with genetic data or to put results into perspective; even if it is sometimes difficult to reconcile the different views. This is illustrated by how our object of interest is designated in different disciplines ((Henn et al., 2018)). The term **anatomically modern humans** (AMH) stems from paleontology. Membership to this category is decided based on two cranial features that distinguish AMH and "archaic" crania (Jobling et al., 2013). When a new human cranium is discovered, researchers look for these typical features to classify it as AMH or archaic. At the moment, the earliest remain with fully modern features is the $\sim$195,000 years old cranium from Omo Kibish in Ethiopia (McDougall et al., 2005). ***Homo sapiens*** is another term, stemming from biology and the concept of species – which in itself is the subject of much discussion. In this thesis, unless specified otherwise, I am concerned with current-day humans and their ancestors (excluding Neanderthal and Denisovan, as these represent a small fraction of

Sub-Saharan African genomes); and I usually designate them with the term "modern humans".

3.2 Abundance of resources

Medicine and health sciences and natural sciences are the two higher education sectors receiving the most research and development (R&D) funding in Sweden (in 2013-2017, (Hansson et al., 2019)). It is similar in the US. Human population genetics benefits from advances in both medical sciences and natural sciences and conversely, advances in human population genetics inform medical research. For example, genome-wide association studies, which are widely used to identify the genetic basis of phenotypic traits, are based on the property of LD and use SNP array data (Visscher et al., 2017). Moreover, it has been shown that taking into account population structure is important in genome-wide association studies (Price et al., 2006), as is taking into account the genetic makeup of an individual to adapt medical treatment. In this context, human population geneticists and molecular anthropologists benefit from unique resources, such as an exceptional knowledge of the human genome; an abundance of reference datasets; the access to many published datasets; and technologies and inference tools tailored to the human genome.

3.2.1 Some aspects of the human reference genome

Identifying similarities and differences between the genomes of different individuals is an essential step for many population genetic analyses, which makes an accurate reference genome invaluable. Here I discuss some aspects of the human reference genome. The first versions of the human reference genome, mentioned in Section 2.2, were "linear": the reference is a set of strings constituted of the four nucleotides. In fact, most reference genomes are linear. Optimally, we want as few strings as there are chromosomes (the longest genetic unit). In practice, it is often difficult to assemble strings into larger units; some reasons are mentioned in the next paragraph. One of the issues about the concept of a reference genome is to decide what it should be; for example, should it be the genome of a single individual (and in that case, which allele should be chosen at polymorphic sites?); or should it be a combination of different individuals? An important thing to keep in mind is that the human reference genome is not an absolute reference; this would be achieved if one could construct a genome that is equally related to all modern humans living today. Rather, the human reference genome allows researchers to make relative comparisons between individual genomes (for example, counting how many differences they have compared to the reference). In that sense, the main ancestry of the human reference genome matters; with a main Eurasian ancestral background, Eurasian genomes will on average have less differences than

e.g. African genomes. For the work in this thesis, the chosen approach for the human reference was the mosaic human reference genome (Lander et al., 2001), a "pan-genome" including fragments from different anonymous donors; a donor of presumably African-European ancestry represents ∼70% of the genome (Schneider et al., 2017). This is the approach used in most human studies at the moment. Another aspect is that sequences not found in the reference genomes will not map and thus cannot be assembled and analysed using default processing pipelines (Eisfeldt et al., 2020).

Some of the features that are hard to access in the human genome are the telomeres and the centromeres; they are essential to the duplication of the genome but are characterised by repetitive sequences (*i.e.* "words" of varying length which are present in many copies), which are hard to sequence. Similarly, some parts of the Y chromosome are difficult to sequence. Other genomic regions are particularly hard to represent with a linear genome, for example the major histocompatibility complex region, which is essential in understanding immunity; and insertions and deletions (in particular the large ones).

The different versions of the human reference genome address these issues; each version has less gaps than the previous one. The latest version of the human reference genome, **GRChg38/hg38** (patch 13), has 3,272,116,950 bp of which 3,110,748,599 (95.1%) are not "N" (*i.e.* we know which nucleotide is at the position). (GRCh38 stands for "Genome Reference Consortium Human Reference Version 38"; hg38 is an alternative name.) It is a more complete assembly than the preceding version, GRChg37/hg19. GRChg38/hg38 includes "alternate loci", in particular in the major histocompatibility complex region; this means that instead of a single string of nucleotides for the region, there are several possible strings. However, most processing and analysis tools are not able to handle this information properly. For example in this thesis (Papers II and III) I used bwakit, an extension of bwa (Li and Durbin, 2009) to ensure that the mapping would not be adversely affected by these alternate loci; but I did not take advantage of the extra information. Although GRChg38/hg38 was published in December 2013 (and is continuously being updated with the release of new patches) (Schneider et al., 2017), it has not replaced the preceding version yet, though increasingly many recent papers use it (Bergström et al., 2020) and large datasets are re-mapped to it (1000 Genomes Project Consortium, 2015). It is thus sometimes necessary to proceed to a "lift-over" to convert the positions in a dataset from a reference to another; this is done with the help of a "lift-over chain" which contains the correspondence between the two genomes. This is not ideal as some positions are lost in the process, but it allows scientists to take advantage of data generated with different reference genomes. Some reference datasets such as the ones discussed in the next section are not as readily available for GRChg38/hg38.

An alternative to linear genomes are genome graphs, that include alternative paths in regions where a single linear sequence does not appropriately repre-

sent the diversity (Paten et al., 2017). Such approaches alleviate the issue of reference bias, in which the reference allele is more likely to be called than the alternate (Degner et al., 2009; Brandt et al., 2015), thus biasing the diversity towards the reference. Another approach also reducing reference bias is to construct "local" reference genomes, with *de novo* assemblies or other means. This was done in a study of Qatari genetic variation, for which a Qatari reference genome was built (by flipping the alleles in the human reference genome to match the major allele in the Qatari sample) (Fakhro et al., 2016).

3.2.2 Reference files and datasets

Depending on the type of questions addressed, different reference files and datasets are useful. In this section I present two of many examples: dbSNP and genetic maps.

dbSNP (Sherry et al., 2001) is a record of reported variants (SNPs, microsatellites and short insertions and deletions), for humans and other species. Any study identifying variants can report them to dbSNP, thereby increasing the size of the database; however, only a part of the reported variants are validated. Variants in dbSNP are given a "rs number" which can serve to identify them. The latest dbSNP release for *Homo sapiens*, build 154 (June 2020) comprises 729,491,867 variants. One common use of this database is to calculate the proportion of novel variants (*i.e.* variants not reported previously) in a new sequencing study; it is important to specify the dbSNP version this proportion refers to, as the number of variants increases with each version. The GATK Best Practices use dbSNP as a repertoire of known variation, and for variant annotation (Section 2.2 and Paper II).

Recombination is not uniform across the genome. **Genetic maps** (also called recombination, linkage or haplotype maps) record the physical location (in bp) as well as the genetic location (in centiMorgans, cM) of positions of the genome. The genetic distance between two locations is related to the frequency of recombination; the larger the distance the higher the chance that a recombination event will happen between two locations. 1 cM between two positions corresponds to a frequency of 1% of recombination per generation. By combining genetic and physical distance between two positions, it is possible to calculate the recombination rate (expressed in cM/Mb). The recombination rate varies between and within chromosomes (larger in shorter chromosomes; larger towards the telomeres than towards the centromeres); between males and females (larger in females); and between populations (Jobling et al., 2013). The most commonly used genetic maps are those of the HapMap Consortium (International HapMap Consortium and others, 2005, 2007) and the Icelandic recombination map (Kong et al., 2010). The first maps identified haplotypes in four populations and thus focused on long-term patterns of recombination; while the Icelandic map is based on a large number of families and incorpo-

rates information about short-term recombination events (happening between parent and child). HapMap maps for specific populations show similarities and differences between populations; thus, it is preferable to choose the genetic map most similar to the population of interest, or a map averaged across different populations.

3.2.3 Availability of data

In most cases, it is mandatory to make the genetic data underlying studies of human diversity available (World Medical Association and others, 2013). This can be done in specific repositories (such as the European Genome-Phenome Archive, EGA) or directly on research group websites. Accessing the data might require permission (sometimes including an application reviewed by an ethics committee) and entail specific instructions (for example, not to be used for selection scans); or it might be used without restrictions. The data might be available as raw or processed; raw data is often preferable, in particular when working with high-coverage genomes, where the sequence processing might introduce biases. The availability of datasets enables researchers to perform analyses without producing any new data, provided that they have access to sufficient storage and computing power. When assembling a dataset of SNP array data, one key aspect is the overlap of different arrays; because SNP arrays contain a limited number of variants (typically from 0.5 to 5 million), it is preferable to combine samples typed on the same array to maximise the number of sites in common; when this is not possible, approaches such as phasing and imputation (Section 2.2) help to obtain a denser set of variants. Note, though, that only known variation can be imputed (*i.e.* variants are added to a callset based on known associations of variants), which biases downstream analyses towards known variation. As a result, imputation should be avoided for some analyses.

There are different strategies when generating new genetic data; it is often a trade-off between number of variants and number of individuals, or number of populations and number of individuals. In studies of molecular anthropology, less dense datasets, such as those based on SNP arrays, tend to have larger sample sizes while studies based on denser datasets, such as high-coverage genomes or exomes, tend to sample fewer individuals. However, there are larger genome datasets generated by national consortia, generally focusing on the ancestry in a specific country (UK10K Consortium et al., 2015; Ameur et al., 2017; Nagasaki et al., 2015; Choudhury et al., 2017; Okada et al., 2018; Telenti et al., 2016).

A few historical collections of samples have a sampling scheme more representative of worldwide diversity, and were created with the specific aim to be of aid to the scientific community in general; consequently, the samples included in these datasets are frequently used in studies. The **Human Genome**

Diversity Project (HGDP) and the **1000 Genomes** project are two iconic examples. HapMap, mentioned above, is another one.

Almost thirty years ago, a group of scientists published a short article presenting the idea of the HGDP project (Cavalli-Sforza et al., 1991). At that time, the project for the sequencing of the human genome had been started, and these scientists were calling attention on the information entailed in the DNA of modern human, in particular from "[the populations] that have been isolated for some time, [that] are likely to be linguistically and culturally distinct, and [that] are often surrounded by geographic barriers" (Cavalli-Sforza et al., 1991). The authors were trying to fund their project of obtaining DNA from ∼25 individuals from ∼500 populations representing worldwide ethnic and geographic diversity. This goal was not reached due to lack of funding and ethical concerns (among other things); nevertheless, 1064 cell lines representing 51 populations were established and kept at the Centre d'Étude du Polymorphisme Humain (CEPH), allowing for virtually infinite preservation of DNA. Populations from Africa, Asia, Europe, Oceania and the Americas are represented. Numerous studies have generated data for these samples or included them (Cann, 2002; Rosenberg et al., 2002); in 2020, high-coverage genomes for 929 of the individuals were published (Bergström et al., 2020). Prior to this, high-coverage genomes had been obtained for 142 of the individuals in other studies, including the Simon's Genome Diversity Project (SGDP); these genomes represent the majority of the comparative datasets in Papers I, II and III (Meyer et al., 2012; Rasmussen et al., 2014; Mallick et al., 2016). Of particular interest in my thesis are the Ju|'hoansi (a San population from southern Africa), two Bantu-speaker populations from southern Africa, and the Mbuti and Biaka (RHGs from central Africa).

The primary aim of the 1000 Genomes project was to discover all variants present at 1% frequency or more in order to facilitate identification of disease-associated variants (Jobling et al., 2013). Investigating questions of evolutionary relevance was a secondary aim. The 1000 Genomes project started in 2008 and like the HGDP, lead to many major publications; the third and last phase of the project presented the genomes of 2,504 individuals from 26 populations, sequenced using different technologies and at various depths (1000 Genomes Project Consortium, 2015). It includes populations originally sampled for the pilot phase of the HapMap project (International HapMap Consortium and others, 2005) (CEU, CHB, JPT and YRI, see Table 3.1 for explanation of the abbreviations) and, in turn, some individuals from the 1000 Genomes collection were included in other studies (Mallick et al., 2016). Samples from the 1000 Genomes project have IDs starting with "NA" or "HG". In Papers II and III we include some of the high-coverage genomes from the 1000 Genomes project, while in Paper IV we include individuals typed on the Illumina 2.5M array (1000 Genomes Project Consortium, 2015). The three letter population codes for the 1000 Genomes populations often confuse researchers from other fields; the ones most used in my thesis are listed in the Table 3.1.

Code	Full name	In paper
CEU	Utah residents (CEPH) with Northern and Western European ancestry	I, II, III, IV
CHB	Han Chinese in Beijing, China	IV
LWK	Luhya in Webuye, Kenya	I, III, IV
YRI	Yoruba from Ibadan (Nigeria)	I, III, IV
MKK	Maasai from Kinshasa (Kenya)	IV
ESN	Esan in Nigeria	III
MSL	Mende in Sierra Leone	III

Table 3.1. *Some populations from the 1000 Genomes project.*

It is worth mentioning that although most of the larger human genetic diversity studies are based in Eurasia and North America, there is an increase of studies funded and conducted in Africa; in particular, in Paper III I use the high-coverage genomes generated by the Southern African Genome Programme (SAHGP) (Choudhury et al., 2017), and in Paper IV we generated genome-wide data for Zambian populations using the H3Africa SNP array, specifically designed to better capture African diversity and developed by the African Genome Variation Project (AGVP) and the Human Health and Heredity in Africa (H3Africa) consortium (Rotimi et al., 2014; Gurdasani et al., 2015; Ramsay et al., 2016).

3.3 Ethical aspects

It should now be apparent that humans are a special case in population genetics studies; on top of the evolutionary forces common to all living organisms, special attention should be given to cultural processes when formulating hypotheses, designing studies and contextualizing the results. Moreover, molecular anthropologists benefit from incomparable resources, including rich reference datasets and access to complex analysis tools. Studies concerning humans are also special in terms of ethical considerations, and are strictly regulated (as are studies of other species, in particular mammals, though not to the same extent nor for the same reasons). Here I give an overview of some ethical concerns raised by human evolutionary genetic studies; I then follow the different steps of a typical study and present the important ethical moments; and finally I discuss issues related to naming and categorizing.

3.3.1 Ethical concerns in evolutionary genetics

The outcome of evolutionary genetic studies has the potential, as any study of human diversity, to be used to discriminate individuals or populations. We generally study individuals in categories defined by different factors (examples are given later in this section), which is one of the basic elements of racism. I introduce some ideas about why we do this in evolutionary genetics and why I believe that, provided that care is taken to consider the implications of one's

work and to try to prevent any adverse effect, our work should be performed (even if there is a potential for misuse of the results).

Three elements are necessary to define racism: distinct categories based on observed differences (*e.g.* based on skin colors), that are heritable and are given a value (Heyer, 2017). The object of genetics is observed and heritable differences; and some concepts can easily be (mis)interpreted in terms of values, such as positive or purifying selection, genetic burden or fitness. In human population genetics, we use categories to make sense of what we observe. For example, we might observe that individuals in a village self-assign themselves to one of two groups; we can then test whether grouping the individuals into these two self-assigned groups explains more of the variance of the observed diversity than considering all individuals as a single group. If we do not explain more of the variance by using the two groups, we discard these categories (for genetic studies). We can then try different categories.

Are human "races" useful categories to make sense of the diversity observed in current-day populations? (I am not considering that races are given different values here.) We (human geneticists) rarely if ever directly address this question (which in itself suggests that the answer is no). I will present some elements of answers. First, the definition and usage of the term "race" is fluctuating. In English, historically, it was applied to humans and animals – in the latter case, it is what we today call "breed". In French among other languages, there is still a single word for race and breed. Breeds are the product of artificial selection: humans have selected animals for specific features and, over time, created different breeds. This has obviously not happened in human history. Race might also be used as a synonym of "subspecies" - another term that is not easily defined. While it is quite clear whether a domestic animal belongs to one breed or to another, it is much more difficult (and, in fact, impossible) to classify humans in a single way (see *e.g.* the classifications of Linnaeus and Galton in (Jobling et al., 2013)). As an example, we can take the four categories defined by Linneaus based on different criteria, including skin color; they correspond broadly to continents: Europeans, Asians, native Americans and Africans. We can ask whether using these categories explain a significant part of human diversity, for example by identifying mutations that differ between the four groups. Most likely, we will find some mutations, for example involved in the genetic determination of skin color, which is not surprising since skin color is one of the criteria used to define the categories. The mutations will represent a small fraction of the genome, and therefore will not be useful to explain diversity unless one is interested in the genetics of skin color (or other specific traits). Even then, the identified mutations will not explain all the variation in skin color. In fact, skin color is a complex trait with continuous variation; when trying to elucidate the genetic bases of skin color, scientists measure the degree of melanin on a continuous scale instead of using discrete categories. This was done by *e.g.* (Crawford et al., 2017), who identified loci associated with skin pigmentation in African populations; among

other things, they noted that this trait is highly variable within Africa. Because the categories are not informative about human diversity in general, they will be discarded and we will look for other categories.

Researchers have estimated (for example with F_{ST}) the amount of variation explained by differences between groups (*e.g.* continents), among the groups within a continent, and within a group. In a study of 52 populations from the HGDP project and using 377 autosomal microsatellites, (Rosenberg et al., 2002) calculated that the variance explained by variation within populations was $\sim$94.6%, and among populations $\sim$5.4%. Similar findings, *i.e.* that the majority of the variation is within groups (or populations depending on the chosen grouping), and not between continents, have been made in other studies, though the exact values differ, depending among other on the type of variants (Holsinger and Weir, 2009).

Another thing to keep in mind is that fixed polymorphisms – *i.e.*, variants where one population has allele "A" exclusively and another population has allele "a" exclusively - are very rare in humans. Using 929 high-coverage genomes, Bergström et al. (2020) identified such private variants for the continents or major regions defined as: Central & South Asia, Middle East, East Asia, America, Oceania, Europe and Africa. From the 67.3 million SNPs in the dataset, none was private to a continent or major region. Moreover, only a few tens of variants were present at a frequency of >70% in one of the regions.

Much of the debate around races happens in the US, where "race" is also the product of social history. Censuses are organised every ten years since 1790, encouraging categorization; note that the number of categories changes from census to census, which illustrates that the categories are constructs. Moreover, many regions prohibited "inter-racial" marriages, starting during the 16$^{\text{th}}$ century in Maryland. It is therefore not surprising that it is possible to find some genetic differentiation between communities: barriers to gene flow have been introduced and strongly enforced socially for centuries before the abolition of segregation. Indeed, the last law forbidding "inter-racial" marriages was definitively abrogated in the state of Alabama in 2000, 33 years after the Loving v. Virginia Supreme Court ruling repelling inter-racial marriage bans at the federal level in the US.

To sum up - it is essential to remember that "races", as other categories (based on *e.g.* language or lifestyle) are constructs; they do not exist by themselves. If such a construct is useful to answer questions of evolutionary interest, it will be used. If not, it will be discarded. This does not relieve scientists from all responsibilities on how their research can be used. Often, the key aspect is to be explicit, *e.g.* about the categories we use and why we use them, or the models we use. This is illustrated in (Lieberman and Jackson, 1995), where three models for the origins of modern humans are presented together with their implications for race (from an anthropological point of view).

Another general concern is related to the fact that much of the work of evolutionary geneticists focuses on small, often marginalised populations, some-

times living in remote areas, and with a nomadic lifestyle (Bankoff and Perry, 2016). While these populations are of great evolutionary interest to researchers, we have to be careful of our position as "outsiders" and realise that the interest of the populations might not be what we think it is, and that our research might have unforeseen effects. Some of the most obvious issues, particularly for nomadic populations in a world based on land ownership and boundaries, are related to the right to live somewhere; in fact, right to the land is sometimes based on the ability to demonstrate local ancestry, and genetic studies have been used to do that, though it is very difficult (if not impossible) to define and demonstrate "local ancestry" (Bankoff and Perry, 2016; Verdu, 2019a). In fact, DNA from current-day populations does not tell where their ancestors lived; at best, hypotheses based on signals of adaptation exclusive to a specific environment could be made (if we were able to identify such signals). Even so, the fact that genetics is one view among others defining the identity of an individual or of a population remains.

Associating specific genetic variants (in particular disease-related) to a population might also have negative consequences, such as singling out the population. On the other hand, the populations and the individuals participating in studies might benefit from research in different ways; this will be developed in the next section.

3.3.2 The different stages of a research project - focus on sampling

After painting a somewhat scary picture of ethical concerns for human evolutionary genetics, I will now introduce considerations related to the different steps of a study of modern human genetic diversity. Such studies often involve sampling, though it is also possible to work exclusively with published datasets; this raises additional questions depending on the ethical agreements associated to the data (Kowal et al., 2017; Stoneking, 2016). I was not involved in sampling myself; all samples included in my projects were collected prior to the beginning of my PhD. Thus, the paragraphs that follow are based on discussion with colleagues and supervisors and on reading rather than on first hand experience.

Research on humans is regulated by the Declaration of Helsinki (World Medical Association and others, 2013), which was written in 1964, and revised regularly (the latest version is from 2013). The Declaration is not a legally binding document, and its status is fluctuating; since 2008, a different set of rules regulates human research in the US. However, if one replaces the "physician" by the "researcher", the Declaration covers all aspects of human evolutionary genetics studies; in particular, it lists all information that has to be provided, and discussed with, potential subjects, to achieve the informed consent of the participants. In Sweden, the Declaration was used during prelimi-

nary work for the "Act concerning the Ethical Review of Research involving Humans" (SFS, 2003).

The Declaration can be used as a support to develop the sampling protocol, which has to be reviewed by ethical boards prior to the sampling. In general, the first step happens at the researcher's institution, with an institutional review board. In Sweden, this used to be the task of regional review boards, but it is now performed by a national board. Then, local ethics committees (at the place of sampling) review the project. These committees are sometimes government agencies. In some cases, there are also local councils to whom the studies are presented. In South Africa for example, the South African San Institute issued the "San code of research ethics", which presents guidelines for conducting research according to four principles: respect; honesty; justice and fairness; and care. Research proposals should be reviewed for these principles by an ethical board (South African San Institute, 2017). Similar initiatives exist in Australia and in North America. Finally, there might be more informal responsible entities to whom the project should be presented at an early stage. For example, WIMSA (the Working group of Indigenous Minorities in Southern Africa) is the umbrella body for southern African San; and Namibia, Botswana and South Africa have their own San councils; the local representative of the San council should be contacted prior to sampling.

Once this legal framework is in place, the recruitment of voluntary participants "capable of giving informed consent" (World Medical Association and others, 2013) can begin, according to the sampling protocol. While the Declaration mentions clearly the aspects to consider (and I cite the Declaration directly in the following paragraphs), each and every aspect can be problematic in the field (Verdu, 2019b). On the other hand, valuable experience and information can be gained from these difficulties.

First, the "aims, methods, sources of funding, any possibly conflict of interest, institutional affiliations of the researcher" have to be presented to the community. (Bankoff and Perry, 2016; Verdu, 2019b) raise difficulties related to explaining the aims and methods of evolutionary genetics studies to the general public and in particular to populations who have a low rate of formal education. In this context, it is essential to work with local researchers who know the community well, with translators, with local guides, or with researchers from other fields such as cultural anthropology. Time is also a key aspect: by repeatedly presenting the research project to potential participants, and giving them the opportunity to ask questions, it is possible for the researchers to know whether they managed to make themselves understood and thus provided a good setting for the informed consent (Verdu, 2019b). These extensive discussions can also result in new directions for the research.

The potential participants must also be informed of "the anticipated benefits". This is a complex question due to the unbalanced relationship between researchers and potential study participants. Researchers and participants benefit from the research in very different ways, and generally the researchers

benefit the most. This is particularly true for the populations that I studied in my thesis, who generally have access to less resources than people who engage in academic human evolutionary genetic studies. The benefits for the participants – *i.e.* added knowledge of the population's history - are immaterial and long term. In this context, it is essential that the terms of the asymmetry are clear and agreed upon between the researchers and the participants. No financial retribution of biological samples are allowed and thus material compensation for participation should be limited (World Medical Association and others, 2013). One option is to give utilitarian items to the community; but this also can create unforeseen issues (Verdu, 2019b).

The next point is about the "potential risks of the study and the discomfort it may entail". The sampling itself does not represent much risk or inconvenience, especially since saliva samples, less invasive than blood samples, have become more common. The greater risks are less foreseeable and therefore more serious. Researchers are in most cases outsiders and their work might intertwine with local political and social issues, for example when deciding to work in one village rather than another, or with one particular family rather than another. Another risk lies in the information contained in the DNA, and in the fact that we cannot say today which information might be obtained from DNA samples in the future.

The potential subject must also be informed of "post-study provisions" and of his/her right to refuse to participate or "to withdraw consent to participate at any time without reprisal". This is called "opting-out". After discussion of the aspects described above– the researcher must "seek the potential subject's freely-given informed consent". The consent can be written or oral (in that case it might be recorded); consent forms should be translated to a language understood in the community. Only once this is done can the sampling itself happen.

The samples are then usually brought to the researcher's home institution. This usually requires a material transfer agreement between the academic institution and the relevant body at the location of sampling. The next step is to make the identity of the donor confidential. In general, this means that samples are given an ID (such as "KSP001") and the correspondence between the ID and the individual is kept in a single copy in a locked archive; this information is kept solely in case an individual wants to opt-out of the study later on. This does not prevent cross-referencing with other datasets but is sufficient for most purposes, in particular in the absence of phenotypic measurements.

The last step is communicating the results of the study. In general, this will be done via research articles, but presentation of the results to the community is important (Bankoff and Perry, 2016). This might take different forms, for example a popular science lecture explaining the findings to the community, or individual personal reports with ancestry information. Results are often also commented in the press; discussing with the journalists can help to get the message across in a "right" fashion. A common misconception is that Khoe-San

people have “the oldest DNA”; when noticing such statements, it might be appropriate to contact the authors and explain what the scientific interpretation is (*i.e.* in this case, that current-day Khoe-San people have evolved as much as all other current-day people, but that some of their ancestors diverged early from the ancestors of the rest of current-day humans). Similar misinterpretations are made about central African populations. Some projects (see examples in (Jobling et al., 2013; Stoneking, 2016)) go further in involving the local communities, by consulting them about questions they would like to address and including them in the continuous development of the project.

3.3.3 Issues of naming and categorization

The question of naming populations in the framework of scientific communications should not be neglected. Concerning the Khoe-San, representatives of the San communities (through WIMSA and the South African San Institute) have explicitly stated how they would like to be named, *i.e.* preferably by population name, *e.g.* !Xun or Ju|’hoansi; if not possible, collectively referred to as San or if including the Khoekhoe pastoralists: Khoe-San. Terms like “Hottentot” or “Bushmen”, which were used by European colonialists, should be avoided (Schlebusch, 2010).

There is no comparable organization and no such general guidelines for the RHGs of the Congo Basin. Some scientists advocate the blanket term “rainforest hunter-gatherers” instead of the historically common “Pygmy” to designate these populations. The RHG are marginalised populations and the name “Pygmy” can have a derogatory connotation; it is a name given by foreigners and not how the people call themselves (Bahuchet, 1993). In this thesis, I use the expression “rainforest hunter-gatherers”, abbreviated as RHG. This expression is not ideal either, because none of the populations under consideration is strictly hunter-gatherer and some are fishermen or mostly farmers. Moreover, it might conceal population specificities, and the fact that the communities we are referring to are evolving. When possible, we thus refer to the individual, self-determined populations names.

When dealing with specific populations there might also be confusion about the names. In particular, one characteristics of Bantu languages is to incorporate the plural mark into the noun, sometimes as a prefix. Thus population names start with “Ba” (a different article is used for languages for example), but only denotes the plural of what follows. For instance, Bakola (or Ba.Kola) refers to the Kola(s). When using Bantu names, these articles are sometimes included and sometimes not, which can be confusing. Moreover in some cases the name chosen to refer to a specific population is not accurate in terms of ethnology. For example, the term “Mbuti” is used to designate famous RHG samples from the Ituri forest in the Democratic Republic of the Congo (DRC), who are often used as a reference for the RHG. However, “Mbuti” is a blanket

term for three different populations: the Efe, the Asua and the Sua. It is not clear to which of these populations "Mbuti" refers to, albeit these populations do not even speak the same language and even speak languages from different linguistic families (namely Sudanic and Bantu). Similarly, the blanket term "San" was used to designate the samples from the Ju|'hoansi population in the HGDP collection, while population structure among San populations goes back hundred of thousand years back ((Schlebusch et al., 2012), Paper I). Finally, the same name might be used to designate different populations; for example, "Batwa" and derivative thereof is a name repeatedly given by Bantu-speakers to surrounding populations experienced as different (Bahuchet, 1993, 2012). In fact, I study three different "Batwa" populations in Papers III and IV.

Another issue is when there is a lack of consistency in how contemporaneous populations are named or categorised. Categories can be defined by language (Bantu-speakers, Khoisan) or lifestyle (hunter-gatherer, farmers); other categories have a meaning in a specific historical or political context (Pygmy (Bahuchet, 1993), "Coloured" in South Africa). Because these categories are simplified, it is difficult to use them consistently. Different categories are often combined, with datasets containing populations labelled as "rainforest hunter-gatherers" and "Bantu-speakers", obscuring the facts that RHG groups also commonly practice agriculture, and that most of them are Bantu-speakers. While we need labels to refer to populations or groups of individuals, it is difficult for labels to reflect the complexity of individuals and populations. It is thus important to be explicit in what we mean with the labels and to recognise their limitations.

4. What we know about *Homo sapiens* history with a focus on Africa

In my thesis I address questions about human history in Africa, with the methods and tools described in the previous chapters. Here I present a summary of the current knowledge about the evolutionary history of *Homo sapiens*, focusing mostly on the demographic aspects and on Sub-Saharan Africa. I start by an overview of modern human origins in Africa and their expansion to the rest of the world. Then I describe pre-farming population structure in Africa and the populations that are the focus of this thesis. This is followed by a description of some of the major gene flow events that modified the pre-existing genetic landscape (including putative archaic gene flow). I describe briefly how a comparative dataset is assembled, and finally, I focus on challenges related to estimating the timing of different events, based on a comparison of divergence times from the literature.

4.1 Human origins in Africa

It is now widely accepted that Africa is the "cradle of humans" (Jobling et al., 2013; Stoneking, 2016). After the split between the chimpanzee and the human lineage, dated to ∼6-7 million years ago (Jobling et al., 2013), hominins (*i.e.* species with whom we share a common ancestor after the split with the chimpanzee lineage) mostly developed in Africa. The emergence of the genus *Homo* (our genus) is dated to ∼2-2.5 million years ago in Africa, and it is the first hominin genus for which we have evidence of expansions into the rest of the world (Stoneking, 2016). The emergence of our species, *Homo sapiens*, happened in Africa as well, presumably from *Homo erectus* (Jobling et al., 2013). It is difficult to pinpoint exactly when anatomically modern humans emerged (Stoneking, 2016; Stringer, 2016) and whether it is justified that such a moment can be defined. The split between the ancestors of modern humans and the ancestors of Neanderthals and Denisovans is dated to ∼750-550 kya (Prüfer et al., 2014). The period 300-200 kya seems to be an important period in our species evolution (Schlebusch et al., 2017; Hublin et al., 2017; Grün et al., 1996). The accepted model for the origin of modern humans is that of an origin in Africa followed by an expansion to the rest of the world giving rise to a serial founder effect (Nielsen et al., 2007; Ramachandran et al., 2005). This expansion, the "out-of-Africa" (OOA) event, essentially replaced local

hominin populations; it is dated to ∼100-50 kya (Nielsen et al., 2007). The populations outside of Africa are all descendants of a small number of founding individuals – a "bottleneck". As a consequence, genetic diversity is highest in African populations and decreases with distance from Africa – a serial bottleneck. Most phylogenies for different genetic markers root in African populations, supporting the OOA model (Jobling et al., 2013). The genetic diversity observed in non-African populations is a subset of the diversity in African populations – with added complexity due to recurring migrations, back-migrations to Africa and admixture with archaic humans (Jakobsson et al., 2008; Green et al., 2010; Reich et al., 2010; Meyer et al., 2012). Two main routes are suggested for modern humans expanding out of Africa, a northern and a southern route (both from northeastern Africa into the Middle East). The number of OOA events is still debated. Besides the main event, some argue that there was an early migration with small contribution to modern genetic diversity. A whole genome sequencing study of worldwide populations argues that they found signatures of an early OOA event in Papuans (that contributed about 2% to the genome) (Pagani et al., 2016). However more work is needed, as another study with a similar design did not find that pattern (Mallick et al., 2016). Confounding factors could be the use of SNPs versus haplotypes, archaic admixture and/or the choice of threshold to define a genetic component due to an earlier dispersion.

While the history of modern humans outside Africa is relatively well known (Nielsen et al., 2017), their history within Africa has been less studied in terms of paleontology, archaeology and genetics (Schlebusch and Jakobsson, 2018; Vicente and Schlebusch, 2020).

4.2 Pre-farming population structure in Sub-Saharan Africa

It is difficult to pinpoint where in Africa modern humans originated, and how many populations lived in Africa at the time of and before the OOA event. Based on fossil evidence classified as AMH from Ethiopia and dated to 150-195 kya, eastern Africa has long been considered as both the location from which AMH expanded into the rest of the world, and the place of origin of AMH (McDougall et al., 2005; White et al., 2003). The high genetic diversity of some current-day southern African populations was used to support a southern African origin of AMH (Henn et al., 2011). However, evidence from the fossil and archaeological record from northern, eastern and southern Africa and from population genetics studies support a model of population structure in Africa well before the OOA event (Hublin et al., 2017; Grün et al., 1996; Schlebusch and Jakobsson, 2018; Schlebusch et al., 2012, 2017). Details of this model and how different populations contributed to current-day genetic diversity are debated and I elaborate more on it in Section 4.5 and in Paper III.

Genetically, the ancient population structure was largely modified by migrations associated to agriculture and is mostly investigated through studies on current-day hunter-gatherer populations (who usually represent local populations with long histories in the region – while farmer populations usually are recent immigrants to the region) or in aDNA when such samples exist (Schlebusch and Jakobsson, 2018; Vicente and Schlebusch, 2020). In fact, a few current-day populations, the Khoe-San from southern Africa, the RHG from central Africa, and eastern African hunter-gatherers, contribute a lot of information to our understanding of the history of modern humans in Africa (as shown in Papers I and III). Other current-day populations, to date less (or not at all) represented in genetic studies, can also be informative (see Paper IV for an example). Finally, it is possible that some populations do not have any descendants today or became incorporated into other populations to the extent that we cannot distinguish them anymore. In such cases, aDNA studies provide invaluable information. I will shortly present the current knowledge about the pre-farming population structure in Africa; estimates of the times of divergence between different groups are further discussed in Section 4.5.

The current-day Khoekhoe and San populations from southern Africa have been shown to harbor the most divergent genetic lineages among current-day humans (Schlebusch et al., 2012; Tishkoff et al., 2009; Gronau et al., 2011; Veeramah et al., 2012). The divergence event separating the lineages ancestral to Khoe-San populations from the lineages ancestral to the rest of modern humans is dated to ∼300-200 kya (Section 4.5). It is the deepest divergence event in the tree of modern humans. Note that I use "lineage" not only in the strict sense of "genealogical lineage" but also as a synonym to "the branch leading to a population". Khoe-San populations today live in Angola, Botswana, Namibia and South Africa. They are (or were until recently) hunter-gatherers (San) or pastoralists (Khoekhoe) and speak Khoisan languages, characterised by clicks. Khoisan languages form a loosely knit group of five language families that are not related to each other; but are distinct from the other four major language families spoken in Africa. These families are: Niger-Congo (common in western and sub-equatorial Africa, and to which the "Bantu languages" already mentioned a couple of times belong); Afro-Asiatic (common in northern Africa); Nilo-Saharan (common in northeast and eastern Africa); and Indo-European (not native to Africa, but found throughout the continent due to the colonial history). Three groups of Khoe-San populations can be distinguished genetically: southern, central and northern Khoe-San (Schlebusch et al., 2012). These groups loosely correspond to the Khoisan language families from southern Africa. In Paper I, I present high-coverage genomes from each of these groups.

A few years ago, aDNA data from southern and eastern African individuals added new elements to the understanding of the history of the Khoe-San and their relationship with other populations (Schlebusch et al., 2017; Skoglund et al., 2017). One of the main results is that the Ju|'hoansi, a San population that

was considered unadmixed, has ∼9% genetic ancestry from a mixed eastern African-European-like group. This genetic component was absent from a 2000 years old San individual (Schlebusch et al., 2017). This has implications in inferences of divergence times: divergence times are pushed back if gene flow is taken into account (Sections 2.1 and 4.5). Another main result is that the genetic component present in southern African Khoe-San populations today had a wider range in the past (Skoglund et al., 2017). The ancient DNA results from (Skoglund et al., 2017) suggested that there was a genetic cline or gradient between eastern and southern Africa, with *i)* a southern African (Khoe-San like) component gradually declining toward the northeast and *ii)* an eastern African (eastern African HG like component, defined by an ancient sample from Mota, Ethiopia) declining towards the south. This pre-farming southern to northeastern genetic gradient is completely absent from certain African countries today, *e.g.* Mozambique and Malawi (Semo et al., 2020; Skoglund et al., 2017). However certain eastern African HG populations, the Hadza, Sandawe and Sabue, still retain significant genetic components related to the northeastern end of the Khoe-San-like genetic gradient (Skoglund et al., 2017; Scheinfeldt et al., 2019; Vicente and Schlebusch, 2020). Studies on haploid markers also suggest ancient lineage sharing between southern, eastern and central African HGs – suggesting pre-farming gene flow (Naidoo et al., 2020).

Following the first population divergence event between the ancestors of the Khoe-San and the ancestors of the rest of modern humans, a second divergence event, dated to ∼50-150 kya, separated the lineage ancestral to the rainforest hunter-gatherers from the lineage ancestral to the rest of modern humans (except the Khoe-San). The RHG are many populations living in the Congo Basin. Contrary to the Khoe-San, they are not characterised by different languages than those spoken by their neighbours, but rather by complex relationships with their neighbouring populations (Paper III). Several studies have focused on stature in RHG populations, for example by searching underlying genetic variants or by testing hypothesis explaining the adaptive advantage of short stature. However in this thesis, I focus on the demography of RHG (Paper III). Based on geographic factors, they are grouped as eastern and western RHG. The common origin of eastern and western RHG was demonstrated (Patin et al., 2009; Batini et al., 2011) and the divergence of the two lineages estimated to ∼40-80 kya (Table 4.4). In recent times (last few thousand years) there was a fast diversification in the western RHG, which might be correlated with the expansion of Bantu-speaking populations, fragmenting their territories (Verdu et al., 2009). It is worth noting that only a subset of RHG populations are represented in genetic studies, and that several populations have never been sampled.

4.3 Gene flow events modified the population structure

The early modern humans populations in Africa did not remain isolated throughout their history. Rather, there probably was frequent gene flow between populations, though it is difficult to characterise it for much of human history (Section 4.5 and Paper III). Recent gene flow events are easier to characterise, not least because such events can also be studied by archaeology or linguistics. I mentioned one of these events previously: the diffusion of pastoralist practices from eastern to southern Africa ∼1500 years ago, including gene flow that introduced the "Pastoral Neolithic" component into the ancestors of current-day Khoe-San populations (Schlebusch et al., 2017). Recent aDNA studies are providing information about the details of this migration event and about the gene flow that gave rise to the Pastoral Neolithic genetic component in eastern Africa (Vicente and Schlebusch, 2020; Wang et al., 2020; Prendergast et al., 2019).

Another event, the "Bantu expansion", modified the genetic landscape in sub-equatorial Africa. This expansion was not only a spread of genes but a spread of technology (agricultural practices and later iron working) and of languages (Bantu languages, one sub-group of the Niger-Congo language family). The expansion started 3-5 kya in western Africa, and reached southern and eastern Africa ∼1,300 years ago (Schlebusch and Jakobsson, 2018). The routes of the Bantu expansion have been investigated using archaeological, linguistic and genetic evidence (Patin et al., 2017; Bostoen et al., 2015; de Filippo et al., 2012; Li et al., 2014; Busby et al., 2016; Alves et al., 2011; Currie et al., 2013). Recent genetic studies have focused on the genetic diversity of Bantu-speaking populations, that have often been considered as a single homogeneous group (Patin et al., 2017; Choudhury et al., 2017; Semo et al., 2020). In my thesis, particularly in Papers III and IV, I focus on the interactions between "local" groups (descendants of local pre-farming populations) and incoming Bantu-speaking agriculturalists. Different patterns have been described for these interactions: in some cases, the local groups have been replaced by the incoming agriculturalists (Skoglund et al., 2017; Semo et al., 2020); in other cases, local populations have admixed with the incoming groups (Bajić et al., 2018). The modalities of admixture vary in frequency and intensity, and examples are given in Papers I, III and IV. The RHG are a special case, with long-term association with neighbouring populations (some, but not all, Bantu-speaking agriculturalists) (Joiris, 2003; Verdu et al., 2013; Bahuchet, 2012; Hewlett, 2014). Sex-biased admixture patterns are commonly observed (Section 3.1). Local groups might also have been displaced by incoming populations, as demonstrated for Malawi with aDNA studies (Skoglund et al., 2017), though this is difficult to see without a dense aDNA record.

Signatures of Eurasian gene flow during colonial time, or related to back-migration to eastern Africa followed by gene flow throughout the continent, can also be detected in most current-day Sub-Saharan African populations

(de Wit et al., 2010; Choudhury et al., 2017; Schlebusch et al., 2012; Chen et al., 2020). With the exception of studies focusing on these events (for example on the transatlantic slave trade or on the Afrikaner or Coloured populations from South Africa) (Fortes-Lima et al., 2017; Hollfelder et al., 2020), most studies of human evolutionary history in Sub-Saharan Africa focus on genetic component descending from Sub-Saharan African populations, and it is common to remove individuals (or genetic components within individuals) with recent Eurasian gene flow from analyses after the initial screening of the dataset.

Gene flow from archaic humans is more difficult to characterise but there is increasing evidence that it contributed to current-day genetic diversity. It corresponds to admixture where one of the admixing populations is, in general, not sampled today, and has separated earlier from the rest of the species than the common ancestor of the modern-day populations (excluding the archaic genetic component in current-day population that is due to the admixture). I use the term "archaic" for events involving a population that diverged earlier than the diversification within current-day modern humans (for example, Neanderthal or Denisovan). Archaic admixture was suggested by evidence in the fossil record (*e.g.* (Trinkaus et al., 2003)). Advances in genomics in the last decade have contributed evidence of such events, in particular in Eurasia. In fact, DNA was successfully extracted and sequenced from several remains of Neanderthal (Green et al., 2010); the same year, DNA from a hominin that was not known through the fossil record (unlike Neanderthal) was extracted from a phalanx of a "Denisovan" human (Reich et al., 2010). By comparing these archaic genomes and modern human genomes, researchers demonstrated that admixture happened between modern humans, Neanderthal and Denisovan, and that non-Africans had a few percentages of their genome deriving from Neanderthal (Denisovan ancestry is more geographically restricted, in southeast Asia and Oceania). The absence of such a signal in African populations can be explained by the fact that Neanderthal populations were not found in Africa (but see the next paragraph).

There are no archaic genomes from Africa at the moment, but archaic introgression has been suggested via *in silico* approaches that identify fragments of the genome with certain properties. S^* was the first statistical tool used to detect archaic admixture in African populations; it looks for highly diverged fragments with extensive linkage disequilibrium; application of this method to various African populations (Yoruba, San, Biaka, Mbuti, Baka) suggested archaic admixture (Plagnol and Wall, 2006; Hammer et al., 2011; Lachance et al., 2012; Hsieh et al., 2016b). Another method based on the conditional site frequency spectrum suggested archaic admixture in several western African populations (Durvasula and Sankararaman, 2020). IBDmix is a recent method (Chen et al., 2020) which has the advantage of not requiring a current-day reference population with no archaic component (*e.g.* African populations are typically considered unadmixed reference populations for detecting Neanderthal

ancestry in Eurasian populations). Applications of IBDmix suggested that at least some African populations (from western and eastern Africa) have a signal of Neanderthal ancestry; this could be explained by back-to-Africa migrations. Finally, ABC approaches demonstrated that genetic data from present day central African populations is better explained when archaic admixture is incorporated into the demographic models (Lorente-Galdos et al., 2019). Similarly (though with a very different approach), Lipson et al. (2020) found that an "admixture graph" (Patterson et al., 2012) of sequences obtained from modern and ancient DNA of modern humans was better fitted with inclusion of two "ghost" populations, an archaic and a modern one, with differential contributions to current-day human populations (the "ghost modern" population is one of four populations separating at the base of the tree of modern humans). Related to this is the discussion of a "basal African" lineage to explain patterns of genetic diversity in some western African populations (Skoglund et al., 2017; Vicente and Schlebusch, 2020; Lipson et al., 2020).

4.4 Summary: assembling a comparative dataset for studies of Sub-Saharan African demographic history

Before further detailing the current genetic understanding of the history of modern humans, I will summarise the major genetic components found in current-day modern humans and explain how I assembled comparative datasets in the different papers of my thesis. As presented in Sections 4.2 and 4.3, we can simplify the history of modern humans in Sub-Saharan Africa by two layers of populations: an ancient, pre-farming layer, represented today mostly by hunter-gatherer populations; and a layer related to the large scale migrations associated with agriculture. In my thesis, I focused on representatives of the first layer (Khoe-San, RHG and, as we show in Paper IV, BaTwa populations with a fishermen subsistence from Zambia), though we include representatives of the second layer, since they represent some of the genetic makeup of the hunter-gatherers today. This is particularly important for the RHG, as relationships with RHG neighbours are at the heart of inter-group dynamics, social organizations, and population categorization in the region. In this case, when possible, pairs of populations were sampled (a RHG population and a neighbouring population). In the case of the BaTwa from Zambia, which had not been part of any genetic study previously, including Zambian "typical" populations (Bantu-speaking agropastoralists) was important in order to investigate whether the new samples were genetically different from other Zambian populations.

Besides representatives of the Bantu expansion and of the Pastoral Neolithic migration, we included western African populations (representing other ancestries than the Bantu-speakers), and eastern and northern African populations. Note that northern African populations today are genetically close to Euro-

peans and Middle Eastern populations: aDNA studies have shown that this is due to several back-to-Africa events, and that there had been intermittent gene flow between northern and Sub-Saharan Africa (Vicente and Schlebusch, 2020; Fregel et al., 2018). The populations are sometimes grouped according to language instead of geography – we then speak of Niger-Congo, Nilo-Saharan and Afro-Asiatic speakers. The geographical and linguistic grouping overlap partially (Tishkoff et al., 2009; Schlebusch and Jakobsson, 2018). Additionally, non-African populations are included to characterise possible gene flow between African and non-African populations, and to contextualise the genetic patterns in African populations.

The exact panel of populations depends on the availability of published datasets, and it is often necessary to make compromises between the quality of the data, the overlap of the variants, and the choice of populations. The datasets of Papers I, III and IV illustrate this.

Finally, one is confronted with more questions if one wishes to include ancient samples in the analysis. On the one hand, including ancient samples can shed a new light on old questions (see for example (Schlebusch et al., 2017)); on the other hand, it is difficult to obtain genetic data from ancient samples that is of the same quality compared to modern data, and it is not always clear how combining ancient and modern samples in a dataset impacts analyses (Günther and Jakobsson, 2019). Moreover, ancient populations, particularly from Africa, are often represented by single individuals, which is known to affect analyses. In Paper I, we included an ancient individual (for which an entire genome of relatively high-coverage is available) in descriptive analyses. This was not done in Papers III and IV, but it is an interesting future task. It would be particularly interesting to include the ancient RHGs from Shum Laka for Paper II (Lipson et al., 2020), and the ancient Malawi hunter-gatherers for Paper IV (Skoglund et al., 2017)).

4.5 Estimating divergence times and models of modern human evolution

Time estimates for the Khoe-San ancestors divergence event, the RHG ancestors divergence event, as well as divergence within Khoe-San groups and within RHG groups are presented in Tables 4.1, 4.2, 4.3, 4.4 (at the end of this section). Figure 4.1 gives an overview of the order of the different events. A quick glance at these tables reveals that estimates for these events have a wide range. There are several reasons for that. One reason is that estimates based on genetic data are typically either scaled with mutation rate and generation time or with population size and generation time. And, in turn, these parameters need to be estimated. In particular, the mutation rate for SNPs in human autosomal chromosomes was historically based on the number of substitutions between the human and the chimpanzee genomes and on the estimated diver-

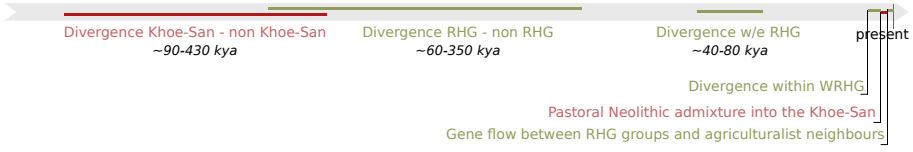

Figure 4.1. Timeline of human evolutionary history in Sub-Saharan Africa. This is a schematic representation indicating the relative order of the events based on a wide range of estimates. See Tables 4.1, 4.3 and 4.4 for detailed information.

gence time between the two species (based on the fossil record). More recently, estimates of the mutation rate based on pedigrees (*i.e.* comparison of genomes of parents and child) have resulted in a mutation rate about half of the human-chimpanzee divergence estimate (Scally and Durbin, 2012). This means that a difference of a factor two between two time estimates can be solely due to different choices of mutation rate. We cannot tell which scaling parameter values are the right ones; moreover, we often assume that the scaling parameters are the same in different populations, and that they did not change over time, which is particularly unlikely for population size. These simplifications are necessary. To ease the discussion, I rescaled the different time estimates in Tables 4.1, 4.2, 4.3, 4.4 with the mutation rate and generation time used throughout this thesis: 1.25×10^{-8} mutation per site per generation (pedigree based estimate) and generations of 30 years.

Another reason for these wide ranges of estimates is that they are based on vastly different methods (Section 2.4): MSMC (and a more recent installment, MSMC2) (Schiffels and Durbin, 2014; Malaspinas et al., 2016); G-PhoCS (Gronau et al., 2011); two genealogical concordance methods, the "Two plus one" and the TT method (Schlebusch et al., 2012; Sjödin et al., 2020); fastsimcoal2 (Excoffier and Foll, 2011; Excoffier et al., 2013); $\delta a \delta i$ (Gutenkunst et al., 2009); and ABC. These methods are based on different aspects of the data, make different assumptions, and measure different things. Importantly, some do not allow for gene flow (MSMC, TT, one version of G-PhoCS), which can impact estimates of divergence time and genetic diversity. MSMC and apparented methods are special in the sense that they do not model directly a divergence event; rather, they give the proportions of lineages from the populations of interest that have coalesced at different time epochs. Generally, the time at which 50% of lineages have coalesced is taken as the divergence time estimate. However, one could use a different definition to say that two populations are separated. It is also known that a divergence event with an instant population split looks gradual using MSMC cross-coalescence rate.

An additional confounding factor is the genetic data itself, and the choice of populations. I included exclusively estimates based on the autosomes in Tables 4.1, 4.2, 4.3, 4.4, and mostly estimates based on high-coverage genomes or exomes, though some estimates are based on SNP array or on Sanger sequencing of short regions of the genome. The effect of the choice of populations can be

seen for example when estimating the Khoe-San divergence time between a San individual and either a Han Chinese, a Yoruba or RHG individual (Table 4.1, (Bergström et al., 2020) row).

Overall, estimates based on MSMC and MSMC2 are more recent than estimates based on other methods (after rescaling). This is particularly evident for the Khoe-San divergence event and the internal divergence of Khoe-San populations (Tables 4.1, 4.2). The Khoe-San divergence event is estimated at ∼90-160 kya with MSMC/MSMC2 while all other estimates are older than 200 kya, and up to 300 kya (Table 4.1). (The (Excoffier et al., 2013) result of 432 kya has to be taken with care, as it is a midpoint between two ascertained SNP panels, and it is unclear which panel is more appropriate.) Note that the ABC approach (Song et al., 2017) (Table 4.1), which resulted in a recent divergence time (130 kya) is based on PSMC as a summary statistics, so it is perhaps not surprising that it results in an estimate similar to the MSMC based ones. Estimates for the divergence between northern and southern San vary widely, from ∼30 to ∼170kya (Table 4.2). The range of estimates for the RHG divergence is also important, though overall, the times are more recent that those for the Khoe-San divergence (Table 4.3). The estimates for the divergence between western and eastern RHG are perhaps the most consistent across studies, from ∼40 to ∼80 kya (note that dates still vary by a factor two) (Table 4.4).

So far I have focused on divergence between two populations, though some of the models (*e.g.* the ones based on ABC) include more than two populations (or more than two groups of populations). Modeling *Homo sapiens* evolutionary history as a bifurcating tree is a convenient tool; however, it is a simplification, as all models are. It is necessary to include gene flow in the tree; while we have a relatively good idea of the timing and intensity of gene flow between the most recent branches, gene flow in the internal branches is more difficult to estimate (see Paper III). This is reflected in the four models suggested for modern human origins in Africa (Henn et al., 2018). The model that could encompass a bifurcating tree is the "Single origin range expansion with regional persistence", in which a main population contributed most to current-day diversity, with contributions from other populations. The other models are: the "African multiregionalism", in which several populations, connected by gene flow, can be distinguished already 300 kya; the "Single origin range expansion with local extinctions", in which a main population contributed most of the diversity and replaced all other populations; and the "Archaic hominin admixture in Africa", in which the contribution of archaic humans is highlighted. Even though the "extreme" models are very different, they can be seen as the extremities of gradients by modulating the intensity of gene flow. This raises important questions and implications, about *e.g.* the time to the most recent common ancestor of modern humans (excluding the genetic components due to archaic admixture). Gene flow from lineages that diverged before the diversification of current-day *Homo sapiens* lineages also complicates the picture – not least because of the hypotheses we need to make (mutation rate, genera-

tion time, population size, genetic diversity) in the absence of any genetic data from such lineages.

To conclude, the uncertainties related to estimating divergence times underline the significance of combining approaches. In particular, archaeological and paleontological remains can be dated with complementary approaches, and ancient DNA sequences are informative about the diversity before the agriculture-related migration events in the last few thousand years. Finally, reconstructions of past climates (paleoclimatology) can corroborate and provide explanations for events observed in the genetic record, such as changes in population size (Paper I) or changes in connection between populations (Verdu et al., 2009; Patin et al., 2009, 2014; Scerri et al., 2018).

Study	Data type	n	Populations[5]	Method (details)	Original (kya)[7]	Rescaled (kya)[8]
(Bergström et al., 2020)	Genomes[1]	4	**KS**: San; **nonAf**: Han	MSMC2 (no gene flow)	162	168
			KS: San; **otherAf**: Yoruba		126	130
			KS: San; **RHG**: Mbuti, Biaka		110	114
(Fan et al., 2019)	Genomes	2	**KS**: #Khomani, Ju|'hoansi; **eHG**: Hadza, Sandawe	MSMC (no gene flow)	68-85 (44-100)	70-88 (46-103)
			KS: #Khomani, Ju|'hoansi; **otherAf**: Yoruba		100 (59-160)	103 (61-166)
			KS: #Khomani, Ju|'hoansi; **otherAf**: Sengwer (Nilo-Saharan)		100-120 (44-160)	103-124 (46-166)
			KS: #Khomani, Ju|'hoansi; **otherAf**: Rendille (Afro-Asiatic)		100-120 (52-160)	103-124 (54-166)
			KS: #Khomani, Ju|'hoansi; **RHG**: Baka, Biaka, Bakola, Bedzan, Mbuti		78-85 (52-120)	81-88 (54-124)
(Mallick et al., 2016)	Genomes	2	**KS**: #Khomani; **nonAf**: French	MSMC (no gene flow)	131 (82-173)	136 (85-179)
			KS: #Khomani; **otherAf**: Yoruba		87 (58-120)	90 (60-124)
(Gronau et al., 2011)	Genome	2	**KS**: San; **otherAf**: Yoruba	G-PhoCS (gene flow)	131 (127-135)	210 (203-216)
			KS: San; **otherAf**: Bantu		129 (126-133)	206 (202-213)
			KS: San; **otherAf**: Yoruba	G-PhoCS (no gene flow)	130 (108-157)	208 (173-251)
			KS: San; **otherAf**: Bantu		121 (117-124)	194 (187-198)
(Lorente-Galdos et al., 2019)	Genomes	7	**Archaic humans**: Altai, Denisovan; **KS**: Ju|'hoansi; **RHG**: Mbuti; otherAf: Mandenka; **nonAf**: Han, French	ABC (archaic gene flow)	191 (161-245)	254 (214-327)

Continued on next page

(Excoffier et al., 2013)	SNP array[2]	56	**KS**: San; **otherAf**: Yoruba	fastsimcoal2 (gene flow)	Model B: 180 Model A: 113	432 270
(Schlebusch et al., 2012)	SNP array	3 haploid	**KS**: Ju|'hoansi; **RHG**: Mbuti	Genealogical concordance (no gene flow)[6]	98 (88-107)	209 (189-229)
(Schlebusch et al., 2017)	Genomes[3]	2	**KS**: ancient San; **otherAf**: Dinka	G-PhoCS (no gene flow) TT method (no gene flow)	336 (SD: 7) Dinka branch: 265 (SD: 5) Ancient San branch: 301 (SD: 5)	-
	Genomes		**KS**: San; **otherAf**: Dinka	G-PhoCS (no gene flow) TT method (no gene flow)	282 (SD: 7) 255 (SD: 5)	- -
(Schlebusch et al., 2020)	Genomes	3	**KS**: Ju|'hoansi, !Xun, Nama, Karretjie people, |Gui and ‖Gana, San; **RHG**: Mbuti; **otherAf**: Dinka, Mandenka, Yoruba; **nonAf**: Dai, French, Han, Karitiana, Papuan, Sardinian	G-PhoCS (no gene flow)	241 (SD: 29)	-
		2		TT method (no gene flow)	241 (SD: 25)	-
(Song et al., 2017)	Genomes[1]	1[4]	**KS**: San; **nonAf**: CEU	ABC (gene flow)	130 (121-141)	-
		2	**KS**: San; **nonAf**: CEU **KS**: San; **RHG**: Mbuti	MSMC (no gene flow)	166 (90-290) 130 (80-205)	- -
(Veeramah et al., 2012)	Autosomal fragments	85	**KS**: San; **RHG**: Bakola, Biaka, Mbuti; **otherAf**: Ngumba, Luhya, Shona	ABC (gene flow)	111 (53-187)	266 (127-449)

Table 4.1: *Literature overview: estimates of the Khoe-San divergence time. Genomes means high-coverage genome unless specified. Autosomal fragments corresponds to Sanger sequencing of autosomal DNA. Additional information on the next page.*

Continued on next page

[1]Physically phased.

[2]The analysis was performed with the Yoruba and the San ascertained panels separately. Here I give midpoint of the 95% CI for the two panels.

[3]One of the genomes is an ancient San sample with diploid called sites.

[4]Pseudo-diploid.

[5]**KS**: Khoe-San; **nonAf**: non African; **otherAf**: other African (non KS, non RHG, non eHG); **RHG**: rainforest hunter-gatherer; **eHG**: eastern African hunter-gatherer; **sKS**: southern KS; **nKS**: northern KS; **wRHG**: western RHG; **eRHG**: eastern RHG.

[6]"Two plus one" method, genealogical concordance assuming a population divergence model (Schlebusch et al., 2012).

[7]In parenthesis, indication of: 95% confidence or credibility interval; or for MSMC analyses: time when 25% and 75% respectively of lineages have coalesced; or standard deviation (SD).

[8]Rescaled with mutation rate 1.25×10^{-8} mutation per base pair per generation, and generations of 30 years.

Study	Data type	n	Populations	Method details	Original (kya)[7]	Rescaled (kya)[8]
Northern/southern Khoe-San divergence						
(Fan et al., 2019)	Genomes	2	**nKS**: Ju|'hoansi; **sKS**: #Khomani	MSMC (no gene flow)	30 (24-30)	31 (25-31)
(Mallick et al., 2016)	Genomes	2	**nKS**: Ju|'hoansi; **sKS**: #Khomani	MSMC (no gene flow)	21 (21-26)	22 (22-27)
(Schlebusch et al., 2017)	Genomes[3]	2	**nKS**: San; **sKS**: ancient San	G-PhoCS (no gene flow) TT method (no gene flow)	185 (SD: 6) San branch: 156 (SD: 5) Ancient San branch: 183 (SD: 5)	-
(Schlebusch et al., 2012)	SNP array	3 haploid	**nKS**: Ju|'hoansi; **sKS**: #Khomani, Karretjie people	Genealogical concordance (no gene flow)[6]	35 (26-42)	76 (55-91)
Internal Khoe-San divergence						
(Schlebusch et al., 2020)	Genomes	2	**KS**: Ju|'hoansi, !Xun, Nama, Karretjie people, |Gui and ||Gana, San	G-PhoCS (no gene flow) TT method (no gene flow)	157 (SD: 23) 194 (SD: 18)	- -

Table 4.2. *Literature overview: estimates of internal Khoe-San divergence times. Genomes means high-coverage genome unless specified. For additional information, see the legend of Table 4.1.*

Study	Data type	n	Populations	Method (details)	Original (kya)[7]	Rescaled (kya)[8]
(Bergström et al., 2020)	Genomes[1]	4	**RHG**: Mbuti; **nonAf**: Han	MSMC2 (no gene flow)	123	127
			RHG: Mbuti; **otherAf**: Yoruba		69	71
			RHG: Biaka; **nonAf**: Han		96	99
			RHG: Biaka, Mbuti; **KS**: San		110	114

Continued on next page

(Fan et al., 2019)	Genomes	2	**RHG**: Baka, Bakola, Bedzan, Biaka, Mbuti; **KS**: Ju|'hoansi, #Khomani	MSMC (no gene flow)	78-85 (52-120)	81-88 (54-124)
(Hsieh et al., 2016a)	Genomes	16	**RHG**: Baka, Biaka; **otherAf**: Yoruba	δaδi (continuous asymmetric gene flow)	156 (140-164)	351 (315-371)
				δaδi (one admixture event)	90 (86-92)	202 (193-207)
(Mallick et al., 2016)	Genomes	2	**RHG**: Mbuti; **nonAf**: Eurasians	MSMC (no gene flow)	112 (66-171)	116 (68-177)
			RHG: Mbuti; **otherAf**: Yoruba		56 (32-84)	58 (33-87)
(Lopez et al., 2018)	Exomes	400	**RHG**: Baka, Batwa; **otherAf**: Nzebi, Bapunu, BaKiga; **nonAf**: Belgians	fastsimcoal2 (gene flow)	135 (58-259)	152 (65-292)
(Lorente-Galdos et al., 2019)	Genomes	7	**Archaic humans**: Altai, Denisovan; **KS**: Ju|'hoansi; **RHG**: Mbuti; otherAf: Mandenka; **nonAf**: Han, French	ABC (archaic gene flow)	118 (86-192)	157 (114-256)
(Patin et al., 2009)	Autosomal fragments	124	**RHG**: Baka (Gabon, Cameroon), Biaka, Mbuti, Twa; **otherAf**: Yoruba, Ngumba, Akele, Chagga, Mozambicans	ABC (gene flow)	56 (26-131)	135 (62-313)
(Schlebusch et al., 2012)	SNP array	3 haploid	**RHG**: Mbuti; **eHG**: Hadza	Genealogical concordance (no gene flow)[6]	Mbuti branch: 64 (55-72)	136 (118-154)

Continued on next page

(Schlebusch et al., 2017)	Genomes	2	**RHG**: Mbuti; **otherAf**: Dinka, Mandenka, Yoruba; **nonAf**: Dai, French, Han, Karitiana, Papuan, Sardinian	G-PhoCS (no gene flow)	221	-
(Schlebusch et al., 2020)	Genomes	2	**RHG**: Mbuti; **otherAf**: Dinka, Mandenka, Yoruba; **nonAf**: Dai, French, Han, Karitiana, Papuan, Sardinian	G-PhoCS (no gene flow)	218 (SD: 113)	-
				TT method (no gene flow)	215 (SD: 8.7)	-
(Song et al., 2017)	Genomes[1]	1[4]	**RHG**: Mbuti; **nonAf**: CEU	ABC (gene flow)	118 (103-139)	-
		2		MSMC (no gene flow)	120 (73-188)	-
(Veeramah et al., 2012)	Autosomal frag-ments	85	**KS**: San; **RHG**: Bakola, Biaka, Mbuti; **otherAf**: Ngumba, Luhya, Shona	ABC (gene flow)	49 (10-106)	117 (24-126)
(Verdu et al., 2009)	Microsat.	604	**wRHG**: Bezan, Baka, Kola, Koya, Bongo; **otherAf**: Tikar, Nzime,	ABC (gene flow)	With Bongo: 54 (21-121)	65 (25-146)
		544	Bangando, Fang, Akele, Teke, Nzebi, Kota, Tsogho, Ewondo		Without Bongo: 90 (23-123)	108 (28-148)

Table 4.3: *Literature overview: estimates of the rainforest hunter-gatherer divergence time. Genomes means high-coverage genome unless specified. Autosomal fragments corresponds to Sanger sequencing of autosomal DNA. Microsat.: microsatellites. For additional information, see the legend of Table 4.1.*

Continued on next page

Study	Type of data	n	Populations	Method details	Original (kya)[7]	Rescaled (kya)[8]
Western/eastern rainforest hunter-gatherers divergence						
(Bergström et al., 2020)	Genomes[1]	4	**wRHG**: Baka; **eRHG**: Mbuti	MSMC2 (evidence of gene flow)	62	64
(Fan et al., 2019)	Genomes	2	**wRHG**: Baka, Biaka, Bakola, Bedzan; **eRHG**: Mbuti	MSMC (no gene flow)	44 (31-50)	46 (32-52)
(Mallick et al., 2016)	Genomes	2	**wRHG**: Biaka ; **eRHG**: Mbuti	MSMC (no gene flow)	38 (27-44)	40 (28-46)
(Patin et al., 2009)	Autosomal frag-ments	124	**RHG**: Baka (Gabon, Cameroon), Biaka, Mbuti, Twa; **otherAf**: Yoruba, Ngumba, Akele, Chagga, Mozambicans	ABC (gene flow)	22 (14-66)	53 (34-160)
(Veeramah et al., 2012)	Autosomal frag-ments	85	**KS**: San; **RHG**: Bakola, Biaka, Mbuti; **otherAf**: Ngumba, Luhya, Shona	ABC (gene flow)	32 (5-79)	77 (12-190)
Internal western rainforest hunter-gatherers divergence						
(Verdu et al., 2009)	Microsat.	604 544	**wRHG**: Bezan, Baka, Kola, Koya, Bongo; **otherAf**: Tikar, Nzime, Bangando, Fang, Akele, Teke, Nzebi, Kota, Tsogho, Ewondo	ABC (gene flow)	With Bongo: 2.9 (0.9-30.1) Without Bongo: 2.6 (0.7-34.3)	3.5 (1.0-36.1) 3.2 (0.9-41.1)

Table 4.4. *Literature overview: estimates of internal rainforest hunter-gatherer divergence times. Genomes means high-coverage genome unless specified. Autosomal fragments corresponds to Sanger sequencing of autosomal DNA. Microsat.: microsatellites. For additional information, see the legend of Table 4.1.*

5. Summary of the papers

This summary follows mostly the order of the papers in my thesis, except for the second section, on processing of high-coverage genomes, as this was done in Papers II and III. Figure 5.1 gives the sampling locations of the populations for which I generated genetic data in Papers I, III and IV. If not specified otherwise, I performed the analyses that I describe and wrote the corresponding articles with input of co-authors.

5.1 25 Khoe-San genomes: unique variation, deep divergence, changes in population size and adaptation (Paper I)

The Khoekhoe pastoralists and San hunter-gatherers from Southern Africa (collectively referred to as Khoe-San) are characterised by their languages sharing click consonants (Khoisan languages) and their subsistence patterns, as well as high genetic diversity. Moreover, they have been shown to harbor deep diverging genetic lineages, particularly on the autosomes and mitochondrial DNA. In fact, the population ancestral to the Khoe-San autosomal genetic component is estimated to have diverged from the population ancestral to the rest of modern humans ∼200-300 kya.

In this paper, we generated 25 high-coverage genomes, five genomes from each five populations: !Xun, Ju|'hoansi, Nama, Khutse San and Karretjie people (Figure 5.1). These populations represent the diversity of Khoe-San populations: northern, central and southern Khoe-San, pastoralists and hunter-gatherers (these categories are not mutually exclusive). This substantially increases the number of published Khoe-San genomes to date (14) (Schuster et al., 2010; Kim et al., 2014; Mallick et al., 2016; Meyer et al., 2012; Bergström et al., 2020; Lorente-Galdos et al., 2019) and in particular the number of populations for which several genomes are available.

We assembled several comparative datasets, either by processing raw sequencing data with the same pipeline as for the new Khoe-San genomes (11 individuals representing 11 worldwide populations) (Meyer et al., 2012); or by generating all-sites VCFs from published studies based on the Complete Genomics platform and merging with these VCFs (∼70 comparative samples) (Drmanac et al., 2010; 1000 Genomes Project Consortium, 2015; Lachance et al., 2012). Some analyses even included a ∼2000 years old San individual

Figure 5.1. Sampling locations of some of the populations in Papers I, III and IV (and explanation of key landmarks on the thesis' cover).

and two archaic genomes (from Neanderthal and Denisova) (Schlebusch et al., 2017; Meyer et al., 2012; Prüfer et al., 2014).

We described the autosomal diversity in this dataset. In particular, we identified novel variants and confirmed that the Khoe-San have the greatest genetic diversity of all sampled current-day populations. However, we masked recent admixture from a mixed Eastern African-Eurasian group and showed that the genetic diversity of the Khoe-San specific component is then comparable to the genetic diversity of other African groups. This underlines the role of admixture in genetic diversity.

We estimated divergence time between Khoe-San and other populations using two complementary methods, TT and G-PhoCS (neither allowing for gene flow) (Sjödin et al., 2020; Gronau et al., 2011). This confirmed that the earliest divergence among the populations represented in our comparative datasets is between the Khoe-San branch and the branch ancestral to the rest of modern humans. Recent admixture impacts estimates of divergence times; in fact, the event is estimated to $\sim$210 kya for the Nama, the population with the greatest level of recent admixture; and $\sim$50 kya older for the Ju|'hoansi, with the lowest level of admixture. This confirms the results from (Schlebusch et al., 2017) who obtained even older estimates for the same event using the genome of an ancient San individual not affected by recent admixture. The divergence among the different Khoe-San groups is estimated to $\sim$160 kya.

We applied two methods to estimate the trajectories of effective population sizes (N_e) in the Khoe-San and comparative populations (Li and Durbin, 2011; Schiffels and Durbin, 2014). We observed a decrease in N_e in all populations starting around 150-100 kya, more intense in non-African populations (in which a bottleneck associated with the out-of-Africa event is well documented). Back-migration to Africa and gene flow from populations that have

a genomic signature of the out-of-Africa bottleneck could explain the reduced N_e in African populations; however, a similar signal was observed in the ancient San individual who predates most documented back-migration events (Schlebusch et al., 2017).

For the Khoe-San populations, we estimated changes in population size with MSMC (Schiffels and Durbin, 2014) for a varying number of individuals and observed a reduction in N_e that was most visible with two individuals. To confirm that different number of haplotypes are more accurate for different times (due to the properties of the coalescent), I simulated genomic data under a bottleneck model and ran MSMC for different numbers of individuals. Bottlenecks starting at different times, of various duration and intensity were tested. The obtained patterns were qualitatively similar to our observed MSMC curves, with different patterns depending on the number of genomes; in particular, estimates based on a single individual (which is relatively common in the literature) did not reflect the decrease in N_e for bottlenecks of moderate intensity starting $\sim$50 kya. Other models incorporating gene flow or population structure could be tested, as it is known that different demographic histories can create the same MSMC curves (Mazet et al., 2016; Mather et al., 2020).

Because of the early divergence of the Khoe-San branch, we could investigate signs of adaptation at different time periods of human evolution, including among early humans and between Khoe-San groups. Selection scans based on PBS-derived statistics suggest selection in early modern humans for genes involved in brain development and immunity, as well as in sperm / flagellum motility. While the exact candidate genes do not necessarily overlap with previous published results, there are similarities in the type of functions under selection. Signals of selection within and between groups frequently included genes related to immunity, diet, and muscle development.

My contribution to this article includes the simulation study of effective population size described above, as well as the assembly of some comparative datasets and the corresponding descriptive analyses.

5.2 Processing high-coverage genomes to study human evolution (Papers II and III)

The raw data from high-coverage genome studies has to be processed to transform a set of reads – in the case of Paper II, strings of 100 or 250 bp – into a callset containing all positions of the genome and their genotype. This processing is not a trivial task; it is time and computationally demanding. Because of the large size of the initial and intermediate files (in the order of $\sim$100 GiB for a $\sim$40 X human genome), it requires important storage. Moreover, because of the many decisions invoked during the processing, it is beneficial that all samples in a study are processed in the same way, which often requires reprocessing of comparative samples (when the raw data is accessible). While

there are recommended processing pipelines, they are generally developed and tested on genomes of Eurasian origin.

The Genome Analysis ToolKit proposes several "Best Practices workflows" (DePristo et al., 2011; Van der Auwera et al., 2013). In Paper II, I examined one of these workflows, the "Germline short variant discovery (SNPs + Indels)" workflow. In particular, I *i)* evaluated the effect of the "triple mask BQSR" step for the processing of autosomal data; *ii)* tested the effect of removing the "Indel realignment" step, which used to be recommended in the GATK workflow but was removed due to a change in the variant calling step; and *iii)* compared the effect of the larger number of individuals at the joint genotyping step. The "triple mask BQSR" is a modification of the "Base Quality Score Recalibration" (BQSR) step that we first introduced in Paper I and used in Paper III as well. The BQSR step improves the quality scores generated by the sequencing machine for each position in the read, and uses dbSNP as a repertoire of known variants. The modification, "triple mask BQSR", consists in adding a variant calling step to perform BQSR with dbSNP and with variants called directly on the individual (Pipelines 3 and 4 in Figure 5.2). The goal of the "triple mask BQSR" is to avoid to disproportionately penalise variants absent from dbSNP because they are mostly or only present in Sub-Saharan African populations (such variants are relatively underrepresented in dbSNP). The ultimate goal was to limit loss of true genetic diversity.

The pipeline comparison was performed mostly on a set of 28 individuals from five groups: European background, Yoruba, Dinka, Khoe-San and rainforest hunter-gatherers (Mallick et al., 2016; Meyer et al., 2012; 1000 Genomes Project Consortium, 2015). The pipelines are represented in Figure 5.2; I compared the GATK Best Practices workflow at the time of writing the article (pipeline 1) to the 2015 Best Practices workflow (pipeline 2, addition of the "Indel realignment" step) and to the pipeline used in Paper III with the "triple mask BQSR" step (pipelines 3 and 4, which are identical but for the number of individuals at the joint genotyping step). I compared the pipelines in terms of number of variants in the final callset (before and after filtering) and overlap of the callsets. I found that the standard pipeline and the standard pipeline plus "Indel realignment" step resulted in almost identical callsets. The replacement of the standard BQSR step by the "triple mask BQSR" impacted the callset slightly, particularly before callset filtering. No significant population-specific effect was detected. We found that including more individuals at the joint genotyping step resulted in different variant counts, likely as a balance between bi- and multiallelic SNPs and/or simple and complex indels, and that including more individuals resulted in slightly more variants even after filtering. Finally, we observed a correlation between average genomic coverage and number of called variants.

Paper II also includes a literature review of processing pipelines of high-coverage genomes in humans and other species (29 studies in total). This revealed a wide diversity of usages; despite the GATK "Best Practices" workflow

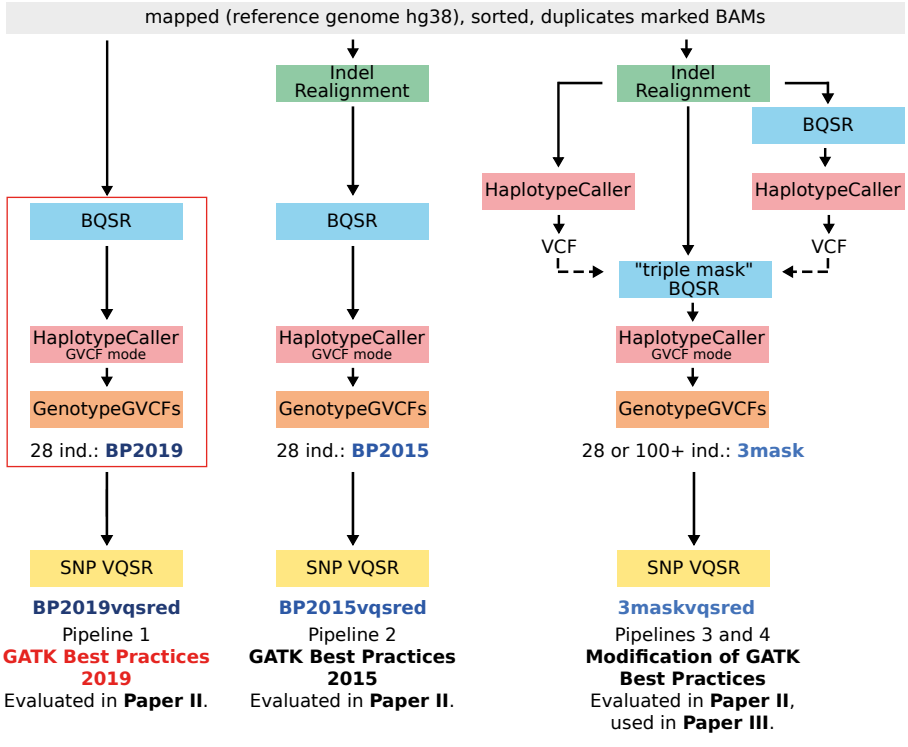

Figure 5.2. Processing pipelines used in Papers II and III. Figure adapted from Figure 1 in Paper II.

being considered a standard pipeline for processing human data, it is rarely followed entirely. An explanation could be that the "Best Practices" change relatively regularly – possibly more often than it takes for a full genome study to be completed.

In Paper III, we tested different options for the processing of the sex chromosomes and the mitochondrial DNA, which is less documented than the processing of the autosomes in GATK. In particular, we compared whether including autosomal data at the BQSR and VQSR (callset filtering) steps made a difference. For the Y chromosome and mitochondrial DNA, we did not detect any significant effect of including the autosomes for the BQSR step and thus decided to leave out the autosomes, in order to save computing resources. For the X chromosome, we decided to include the autosomes at the VQSR step, as the tool benefits from large amount of data and we did not observe a drastic increase of filtered out variants by doing so.

My contribution to this part of Paper III includes planning and discussion about the processing of the sex chromosomes and the mitochondrial DNA; the processing was mostly done by a master student that I co-supervised.

5.3 49 genomes from Central Africa and evolutionary history of Sub-Saharan Africa (Paper III)

The second divergence event in the history of modern humans is between the ancestors of modern-day Central African rainforest hunter-gatherers and the lineage ancestral to the rest of modern human (except the Khoe-San). In Paper III, we expanded the sampling scheme of Paper I and generated high-coverage genomes for 49 Central African individuals: five rainforest hunter-gatherer populations and four neighbouring populations, represented by five or seven individuals (Figures 5.1 and 5.3). The five rainforest hunter-gather populations - Ba.Kola, Baka, Bi.Aka Mbati, Nsua and Ba.Twa - include representatives of the geographically defined western and eastern groups. For four of the five populations, an immediate neighbouring population was included, resulting in four population pairs. This represents a substantial increase in the number of high-coverage full genomes from Central Africa; so far, 51 genomes from rainforest hunter-gatherer have been published, half from the "Biaka" (or Bi.Aka) population (Lachance et al., 2012; Meyer et al., 2012; Hsieh et al., 2016a; Mallick et al., 2016; Fan et al., 2019; Lorente-Galdos et al., 2019; Bergström et al., 2020). Moreover, it is the first time, to our knowledge, that whole genomes from rainforest hunter-gatherers are analysed jointly with whole genomes from their immediate neighbouring populations (but see (Lopez et al., 2018) for such design with high-coverage exomes). The 25 Khoe-San individuals from Paper I are included in this paper as well, with additional sequences. Figure 5.3 provides some information about the different populations and proposes testable hypotheses (note that we did not test directly all of them).

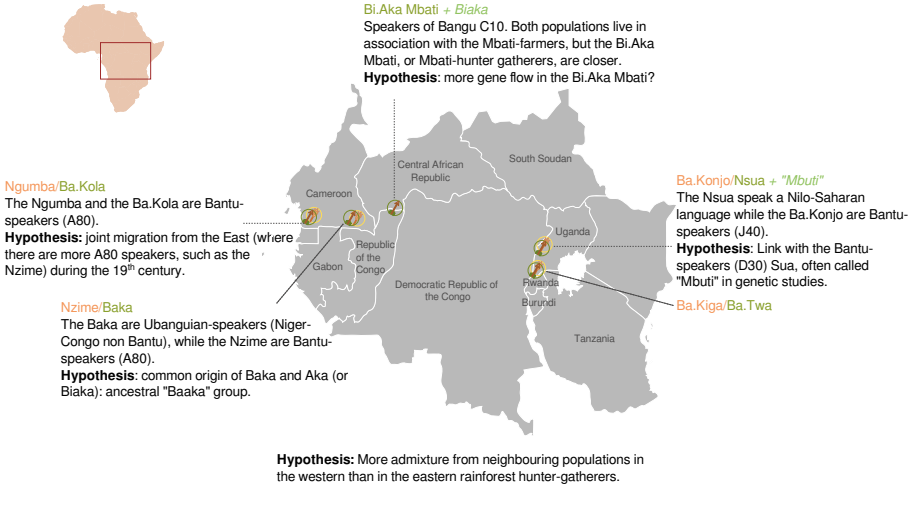

Figure 5.3. Detail of the populations from the Congo Basin and associated hypotheses.

A comparative dataset of 104 full genomes for which raw data was accessible was assembled; 93 had a coverage of 30 X or higher. It includes several African populations as well as five non-African populations. The sample size for comparative populations varies from one to eight individuals, with a median of two individuals per population.

We described the genetic diversity in this dataset. In particular, we demonstrated that in a set of non-ascertained variants, the rainforest hunter-gatherers represent a major axis of genetic variation. The eastern and the western populations can be distinguished. At a more local scale, we contributed new elements in understanding the relationships between different rainforest hunter-gatherer populations, in particular the Baka, Bi.Aka Mbati and Biaka to the west of the Congo Basin; and the Nusa, Ba.Twa and Mbuti to the east. We also confirmed a small population size in the Ba.Twa. We calculated the X-to-autosome heterozygosity ratio, which is informative about past events differentially affecting the autosomes and the X chromosome, such as sex-biased admixture; among Sub-Saharan African populations, we observed a lower ratio in two eastern rainforest hunter-gatherer populations, similar to (Mallick et al., 2016).

This dataset is ideal to investigate questions related to the evolutionary history of modern humans. We estimated the divergence times between pairs of populations under a simple split model with no gene flow (TT method). Similarly to Paper I and previous results, the oldest estimates are for population pairs consisting of one Khoe-San population and one non-Khoe-San population. More surprisingly, the next oldest estimates are for the divergence event between eastern and western rainforest hunter-gatherer populations; the estimates for the divergence event between rainforest hunter-gatherer and other populations (excluding Khoe-San) tend to be younger (there is some overlap). This is at odds with previously reported results and with the other inference method that we explored; one explanation is that the TT method does not allow for gene flow.

The importance of gene flow is further highlighted in the ABC analysis that we performed to compare 24 models of evolutionary history in Sub-Saharan Africa. These models consist of four divergence events starting with a single ancestral population and resulting in five current-day populations (northern and southern Khoe-San, western and eastern rainforest hunter-gatherers, and a rainforest hunter-gatherer neighbour population); two examples are given in Figure 5.4. The models differ in *i)* the possibility and intensity of gene flow, *ii)* in the populations involved in the oldest divergence event, and *iii)* in the relative order of the split between southern and northern Khoe-San on the one hand, and between western and eastern rainforest hunter-gatherers on the other hand. We performed model selection with Random-Forest ABC (Pudlo et al., 2016) on groups of models. In particular, the group of models incorporating the possibility of high levels of gene flow was selected over the two other groups with no possibility of gene flow or possibility of intermediate level of gene flow. The group in which the Khoe-San branch separates first had stronger

support than the two other groups. Finally, there is some evidence that the western/eastern rainforest hunter-gatherer divergence is more recent than the southern/northern Khoe-San divergence.

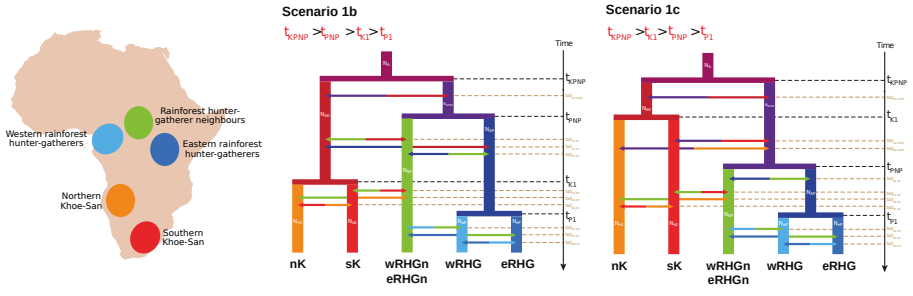

Figure 5.4. Two models of human evolutionary history and schematic map of Africa with the populations included in the models. Adapted from Figure 2 in Paper III.

By combining these results, we selected two models to perform parameter estimation with Neural-Network ABC on 100,000 simulated datasets for each model (the models are shown in Figure 5.4). From the 47 parameters in the models, we were able to estimate well the four divergence times; we had some power to estimate population sizes; admixture times could be estimated, albeit posterior results may only result from constraints on split time events; and the admixture rates could not be estimated. The divergence between eastern and western rainforest hunter-gatherers is estimated to ∼16 kya with one model and ∼43 kya with the other; both estimates are within the range of dates found in the literature (Table 4.4). The oldest divergence event is estimated to ∼330-370 kya, which is on the old end of previous estimates (Table 4.4).

The ABC framework used could be further developed and built upon, for example by incorporating additional summary statistics, which could contribute to diminishing the credibility intervals of the parameter estimates. Additional models, for example a trifurcation, models with the possibility of gene flow from an archaic source, or isolation by distance models, will be explored in future studies. Finally, models focusing on more recent events, such as complex gene flow processes between rainforest hunter-gatherers and their neighbours will also need to be investigated to further unravel the evolutionary history of these populations.

In this article, I performed the majority of the processing and of the analyses, except the divergence time estimates with the TT method. Co-authors contributed to several steps of the ABC analysis, such as defining the demographic models and generating vectors of parameters values for the simulations.

5.4 A new piece to the puzzle: autosomal diversity from two BaTwa Zambian aboriginal populations (Paper IV)

While Southern and Central African populations are relatively well-known, genetically and otherwise, populations from South-Central Africa, in particular Zambia and the DRC, are seldom represented in genetic studies. Today, Zambia is populated mainly by Bantu-speaking farmers, and the knowledge of local populations before agriculture-related expansions is based on archaeological, linguistic and anthropological records. The Bantu expansion is thought to have reached the Luangwa valley in northern Zambia ∼100 AD. Is it not known whether the Bantu expansion resulted in a total replacement of local genetic diversity (such as in neighbouring Mozambique (Semo et al., 2020) and Malawi (Skoglund et al., 2017)) or whether local, pre-farming components are retained in some populations (similarly to the populations studied in Papers I and III). Y chromosome and mitochondrial DNA diversity have been investigated in Zambian populations (Barbieri et al., 2013; de Filippo et al., 2009; de Filippo et al., 2010). Two populations from South-Western Zambia, the Fwe and the Shanjo, harbor mitochondrial lineages common in Khoe-San groups (Barbieri et al., 2013). In the absence of autosomal data, it is difficult to reconstruct a more precise history.

In this paper, we generated genome-wide SNP data for two "BaTwa" populations from South-Western (Kafue flats) and north-eastern (Bangweulu lake) Zambia (Figure 5.1). The BaTwa are fishermen living in swamp areas. They are Bantu-speakers but are considered "different" by their neighbours and marginalised. Oral history suggests a different origin for the BaTwa than for their neighbours. We also included data from three additional Zambian populations, the Bemba, the Lozi and the Tonga (Fortes Lima et al. *in prep.*); these populations are Bantu-speaker agropastoralists and it is not suggested that they have retained a local genetic component.

Two comparative datasets were prepared by merging SNP array data from various African and non-African populations. The less dense dataset, used for a majority of analyses, consists of ∼345,000 autosomal markers in 973 individuals (39 populations). We included current-day representatives of the deep diverging lineages in Africa, *i.e.* the Khoe-San, the rainforest hunter-gatherers and the Eastern African hunter-gatherers. The denser dataset has ∼1.2 million variants.

Genetic diversity analyses revealed that the two BaTwa populations had a hunter-gatherer-like component that is very low or absent in the three other Zambian populations. That component is larger in the BaTwa from Kafue (∼31%) than in the BaTwa from Bangweulu population (∼19%). An analysis of the distribution of runs of homozygosity suggested recent isolation and inbreeding in some of the Zambian populations, in particular in the BaTwa from Bangweulu. The Y chromosome diversity is similar to previous reports from

Bantu-speaking agriculturalist populations, though two haplogroups frequent in the Khoe-San were found in the BaTwa.

We investigated admixture models for the Zambian populations with the software MOSAIC (Salter-Townshend and Myers, 2019). A two-way admixture scenario is the best fit for the five Zambian populations, though the timing of the admixture event, as well as the populations closest to the admixing sources, differ between populations. The major source for the five Zambian populations were the Bantu-speaking Zambian agropastoralists or other Central African agriculturalist Bantu-speaker populations (interestingly, rainforest hunter-gatherer neighbours were often selected). The minor source for the BaTwa from Kafue was Khoe-San-like, while for the BaTwa from Bangweulu it is was less clear (though Khoe-San-like populations are good proxies). However, the results showed that current-day Khoe-San populations, despite being the best proxies in our dataset, are not a good match for the minor source in the BaTwa. This could be related to the observation that BaTwa populations also seem to have some RHG-like genetic component. The fact that neither the current-day Khoe-San nor the current-day RHGs are good proxies for the hunter-gatherer ancestry in the BaTwa is in line with morphological evidence: cranial remains from Zambia (∼4,000 years ago) and Malawi (∼5,000 to 500 years ago) do not show an exclusive Khoe-San ancestry (De Villiers and Fatti, 1982; Morris and Ribot, 2006).

The estimated date for the admixture event in the BaTwa from Bangweulu is ∼1200 years ago and ∼480 years ago in the Kafue group (assuming a generation time of 30 years). This is consistent with archaeological evidence which suggest an earlier settlement of farmers in the region around Bangweulu lake, than in the region of modern-day's Kafue population. It is also coherent with estimated dates for admixture with Bantu-speaker farmers in populations from southern Mozambique and South Africa (Semo et al., 2020; Schlebusch et al., 2016).

This paper demonstrates the importance of studying neglected populations, for example to better understand the pre-farming population structure in South-Central Africa.

In this article, I prepared the datasets and performed the genetic analyses; I wrote a version of the manuscript which was complemented by co-authors, in particular for the archaeological aspects.

6. Conclusions and future prospects

In this thesis, I studied diverse Sub-Saharan African human populations with newly generated genome-wide data from high-coverage genomes (Papers I, III) and SNP arrays (Paper IV). Because of the relative novelty of high-coverage genomes, particularly applied to the study of human diversity, there are uncertainties about how to process and analyze them. In Papers II and III, I investigated aspects of the processing pipelines and showed that a standard workflow can be applied to genomes from Sub-Saharan African populations without resulting in a loss of genetic diversity. In Paper I, I investigated the behavior of a popular method, MSMC, with a varying number of haplotypes in a bottleneck model, and showed that at a time relevant for the history of modern humans, different number of haplotypes had different power to detect a bottleneck. I studied populations that are important for understanding pre-farming population structure in Africa: the Khoe-San (Papers I, III) and the Central African rainforest hunter-gatherers (Paper III). I confirmed their high genetic diversity, which is partially due to recent gene flow, and I investigated divergence times with different approaches. In Paper III, I directly compared different models for the early human population history with an ABC approach, and found that gene flow has been perennial throughout the entire population history of modern humans. The ABC analysis also supported that the first divergence in modern humans was between the branch ancestral to the Khoe-San and the branch ancestral to the rest of modern humans. The selected models were refined with parameter estimation. In Paper IV, I studied two "BaTwa" populations from Zambia, and I showed that they harbor lineages that likely trace back to local pre-farming populations. The admixture between the pre-farming populations and a presumed farmer population occurred $\sim$20 generations ago in one population and $\sim$40 generations ago in the other, consistent with the archaeological record.

When I started my thesis in 2015, the number of high-coverage genomes from Sub-Saharan Africa, and in particular hunter-gatherers, was very limited; it has since increased (Mallick et al., 2016; Fan et al., 2019; Choudhury et al., 2017; Bergström et al., 2020), but some challenges associated with this type of data remain. I mentioned them earlier: small sample size, computational burden, dataset biases when using processed data. In theory, high-coverage genomes should enable us to calculate the true heterozygosity and thus the famous population mutation rate parameter (Θ). We now have much better estimates of Θ, but there is still room for improvement. For instance, sequencing

errors and processing pipelines tend to focus on variable sites and usually non-variable sites are less well characterised, which complicates filtering (among other issues) and can impact estimates of Θ.

Nevertheless, we can learn a lot from genomes. In this thesis I gave examples of the questions that we can address with genomes from Sub-Saharan African populations, for example estimation of divergence times. Study after study, new aspects of human evolutionary history are uncovered. I also gave examples of how a few populations with relatively small census size are essential to our understanding of the evolutionary history of our species. The study of the BaTwa populations from Zambia is a good illustration. Some regions, such as the DRC, remain un- or under-sampled, and therefore we can expect to get a more complete image of pre-farming population structure in Sub-Saharan Africa in the future. aDNA will also be of great help, though it is in its infancy in Sub-Saharan Africa (Vicente and Schlebusch, 2020) and the preservation conditions for human DNA in much of Sub-Saharan Africa are bad. However, it has already provided important information. Finally, we should not forget that studies of modern and ancient genomes are definitely not a priority in several countries of Sub-Saharan Africa that experience political unrest and overall difficult living conditions.

Besides access to relevant populations and assembly of datasets that appropriately reflect human diversity, there is much left to do in terms of inferences of human history. As mentioned previously, we are often restricted by methods that cannot handle too complex scenarios. I utilised an approach that allows for complex models, ABC. The drawback of ABC analysis is that it has many steps involving many decisions. On the other hand, once these decisions have been taken and results have been obtained, the framework (demographic and genetic model for the simulations, calculation of summary statistics) can be improved, for example by including more summary statistics capturing other aspects of the data; or it can be applied to different models. In the case of the study presented here, one natural extension would be to use ABC to investigate archaic or ancient admixture; or to try to translate to concrete models the different scenario proposed for African history by Henn et al. (2018). There are relatively few studies using ABC approaches to study human evolutionary history, particularly based on genomes; but this is changing, and in the future we might expect more advances on this front.

Of course, new inference methods will also help to answer some of the questions I introduced. Some of the particularly intense areas of methodological development are methods evaluating past population sizes; and methods detecting archaic admixture. Methods taking explicitly into account geography – spatial analyses – are also promising as it is evident that gene flow, which is often directly impacted by geographical distance, is an important factor in human evolution. Finally, another development would be to include further the evidence from other fields. Besides developing new methods, comparing existing methods is also valuable and often overlooked as illustrated in my work;

it should be apparent from the tables comparing divergence times (Tables 4.1, 4.2, 4.3 and 4.4) that we are far from agreeing on an exact time for, say, the divergence between the ancestors of the Khoe-San and the ancestors of the rest of modern humans. Estimates are based on different methods, types of data, and populations; and importantly, inference methods might be estimating different things. Given these technical aspects and the complexity of human history, we should not be surprised that we obtain such varied results; studies such as (Zhou and Teo, 2016), that apply different methods on simulated datasets – *i.e.*, datasets for which we know the true history – are valuable to benchmark methods and put their respective results in perspective.

Even if that might be said for any project, I would like to highlight that there are many more questions that can be addressed with datasets like the ones in Papers I and III – because of the wealth of information contained in sequencing data. In fact, the vast majority of what I presented is based on SNPs, which is the simplest type of genetic marker. We do call indels, and sometimes structural variation, but we do so mostly to localise regions to avoid for analyses – because they are too complex. In fact, SNPs are easier to call and to compare between studies, and many tools are designed for SNPs. When I mention the difficulties of finding good reference datasets, or of understanding what exactly was done in a study, it would likely be much more complicated for indels and structural variants. Nevertheless, these markers are bound to contain a lot of information, and we will likely focus increasingly on them in the years to come (possibly with different types of sequencing technologies as well). Another interesting and new perspective is to work with graph assembly (as opposed to mapping to a linear genome).

All in all, there are still many things to explore with human genomes, both in terms of methodological developments and in terms of human evolutionary history. And to finally conclude – after working five years with genomes (starting with zero experience), I despaired that the processing was tedious and never ending, but I remain extremely interested in the questions that we can address, and I would gladly continue to test and evaluate methods (particularly if given infinite time and number of samples).

7. Svensk sammanfattning

Min avhandling handlar om att förstå människans evolutionära historia med hjälp av genetik. Det finns många vetenskaper som studerar människans historia, till exempel antropologi, arkeologi, språkvetenskap och paleontologi. Olika områden bidrar med olika delar av information och är användbara för olika tidsperioder. Till exempel är språkvetenskap informativ för de senaste 10 000 åren medan arkeologi kan bidra med information som sträcker sig flera miljoner år tillbaka.

Jag har använt en annan vetenskap för att studera människans historia, nämligen populationsgenetik eller molekylär antropologi. Populationsgenetik studerar hur mutationer sprider sig i populationer. Vi ärver DNA – arvsmassa, det kemiska ämnet som innehåller den genetiska informationen – från våra biologiska föräldrar, vilka ärvde deras DNA från deras föräldrar, osv. På det sättet innehåller DNA information om alla våra förfäder, och vi kan använda populationsgenetiska principer för att förstå hur nuvarande genetisk variation uppstått. Vi kan studera hur populationer rört sig och förflyttat sig över världen men också studera variationer i populationers storlek. Till exempel, om antalet individer i en population har minskat snabbt och sedan ökat igen, lämnas ett särskilt mönster i populations DNA variation, en så kallad "flaskhalseffekt". DNA från nu levande populationer är informativ för händelser som uppstod för hundratusen år sedan, men också för händelser som uppstod för några decennier sedan. Man kan också undersöka DNA från människor som levde för flera tusen år sedan (ancient DNA), för att studera människans historia.

Det är särskilt spännande att studera evolutionshistoria i afrikanska populationer därför att människan *Homo sapiens* utvecklades i Afrika. Idag vet vi att den moderna människan spred sig via en enskild migrationshändelse från Afrika till resten av världen. Den moderna människans historia i Afrika, före och efter migrationen från Afrika, är mindre studerad, bland annat därför att det finns mindre pengar till forskningen i Afrika än i Eurasien och Nordamerika. Vi vet däremot att *Homo sapiens* utvecklades för cirka 300 000-200 000 år sedan, förmodligen i olika delar av Afrika. Hur många populationer det fanns då, hur de var relaterade till varandra och till populationen (eller populationer) som vandrade ut ur Afrika samt till populationer som lever idag, är viktiga frågor som vi vill besvara.

Tidigare studier har visat att några afrikanska populationer är särskilt intressanta för att förstå människans historia före jordbruket, som förändrade det genetiska landskapet för cirka 5 000 år sedan (söder om Sahara). Dessa populationerna är Khoe-San från södra Afrika; jägare-samlare från regnskogen i

central Afrika; jägare-samlare från öst Afrika; och några populationer som är mindre väl kända.

Jag har i min avhandling använt mig av genetiska analysmetoder för att studera några av dessa populationer. Tekniker för att skaffa genetisk information har utvecklats mycket de senaste decennierna. Ett sätt är att analysera förutbestämda positioner i genomet, positioner där vi redan vet att det finns genetiska varianter, så kallade SNPar, hos människan. Det är samma teknik som används i kommersiella genetiska tester som visar om man är mer släkt med människor från ett eller ett annat land. Det är en relativt billig teknik som används när man vill få en första genetisk överblick av en population eller när man vill titta på många människor. Ett annat sätt är att analysera (nästan) alla positioner i genomet, dvs att sekvensera DNAt. Det kostar mer än SNP-analys, och det tar längre tid och mer resurser för att förbereda data och genomföra analyser. Men man får mycket mer information från hela genomsekvenser.

I det andra projektet i min avhandling visade jag hur man förbereder hela genomsekvenser för analyser. När man studerar genetisk variation i afrikanska populationer behöver man vara extra försiktig då de har större genetisk variation än icke-afrikanska populationer och dessutom är afrikansk variation underrepresenterade i genetiska databaser. Jag visade att vi kan förbereda sekvenser från afrikanska populationer på samma sätt som sekvenser från icke-afrikanska populationer, men också att kvaliten på datat beror på hur många gånger varje position i genomet har blivit sekvenserad.

Moderna människans historia brukar representeras som ett träd, och Khoe-San och sina förfäder representerar första grenen i trädetn. I det första projektet i avhandlingen genererade och analyserade jag genomsekvenser för 25 Khoe-San individer. Tidigare studier har uppskattat att Khoe-San förfäder skiljer sig från förfäder av den moderna människan för cirka 300 000-200 000 åren sedan. I denna studie uppskattade vi tiden av denna händelse igen, och fick liknande resultat. Vi visade också att den höga genetiska variationen hos Khoe-San individer beror, delvis, på geneflöde från populationer som kom till södra Afrika under de senaste 2 000 åren och förde med sig husdjursskötsel och växtodling. Vi använde metoder som beräknar hur storleken i en befolkning förändrar sig med tiden, och visar att storleken hos Khoe-San populationer börjar minska för cirka 100 000 år sedan. Det ser man i alla befolkningar i världen och en förklaring till det kan vara att klimatet förändrades vid denna tidpunkt. Jag genomförde simulationer av den så kallade flaskhalseffekten för att se hur metoder som beräknar befolkningsstorlek förhåller sig till sådana förändringar. Resultaten förstärkte våra tolkningar av orsaken till de observerade mönstrena i datat. Vi letade också efter gener som har påverkats av selektion och visade att sådana gener ofta är involverade i immunförsvaret samt spermier, hjärnan, metabolism och musklerna.

I det tredje projektet studerades hela genomsekvenser från jägare-samlare från regnskogen i centrala Afrika som jämfördes med sekvenser från andra populationer. Jag byggde 24 modeller som beskriver människans historia i

Afrika, från en enda ursprunglig population till fem nutida populationer, och använde simulerat och observerat data för att identifiera den mest sannolika modellen. Vi ser att det är viktigt att ha modeller med möjlighet för genflöde. Resultaten från modellerna visar att det är mer sannolikt att Khoe-San-grenen som först knoppas av från den ursprungliga populationen, samt att den interna uppdelningen av jägare-samlare från regnskogen har inträffat mer nyligen än den interna uppdelningen av Khoe-San linjen. Den bästa modellen valdes ut och andra parametrar uppskattades som t.ex. tiden för när olika populationer splittrats/separerats.

Det fjärde projektet handlar om populationer från Zambia, som kallas BaTwa och som idag lever som fiskare. Flera spår, till exempel från arkeologi, tyder på att BaTwas förfäder kunde vara jägare-samlare som levde i regionen förre jordbrukare kom till dagens Zambia. Vi genotypade mer än en miljon SNPar för två BaTwa populationer och jag jämförde detta dataset med liknande data från andra populationer. Resultatet visar att BaTwa har en genetisk komponent som liknar (men är ej identisk med) nuvarande jägare-samlare, särskilt Khoe-San. De har också en (större) genetisk komponent som liknar populationer som medförde jordbruket till söder-central Afrika.

De olika projekten i min avhandling bidrar med ny information till tidigare studier av *Homo sapiens* utvecklingshistoria i Afrika, särskilt för händelser före spridningen av jordbruket. De bekräftar att det är värdefullt att studera även mindre, och okända populationer. Jag använde olika metoder och utvärderade hur de fungerar. Resultaten bekräftar att *Homo sapiens* historia är invecklad och att vi behöver testa flera modeller för att förstå den ännu bättre.

8. Résumé en français

Dans ma thèse, j'essaye de comprendre certains aspects de l'Histoire de l'Homme grâce à la génétique. De nombreuses disciplines s'intéressent à l'histoire humaine, comme l'anthropologie, l'archéologie, la linguistique, ou la paléontologie. Elles livrent des informations différentes qui nous informent sur des périodes différentes de cette Histoire. Par exemple, la linguistique comparative classique, reposant principalement sur l'étude de textes écrits, considère souvent n'informer que les dix derniers millénaires environ, alors que l'archéologie nous permet de remonter sur plus de deux millions d'années.

J'ai utilisé une autre discipline pour étudier l'histoire humaine, à savoir la génétique des populations (humaines) ou anthropologie génétique. La génétique des populations étudie la façon dont les mutations se propagent dans les populations. Nous héritons notre ADN – une molécule chimique qui contient de l'information génétique – de nos parents biologiques, qui ont hérité leur ADN de leurs parents, etc. Le « génome » représente l'ensemble de l'information codée par l'ADN d'un individu. Ainsi, notre génome contient de l'information sur tous nos ancêtres, et nous pouvons utiliser les principes de la génétique des populations pour comprendre ce qui a formé la diversité génétique que l'on observe aujourd'hui entre les êtres humains à travers le monde. Nous pouvons étudier de quelles façons les populations se sont déplacées ou se sont rencontrées. Nous pouvons aussi étudier les changements de taille de population : le phénomène de « goulot d'étranglement génétique », dans lequel la taille d'une population diminue rapidement puis augmente de nouveau, laisse, par exemple, une signature particulière dans le génome. De même, les métissages génétiques entre population laissent des traces reconnaissables dans les génomes de leurs descendants.

Le génome des populations contemporaines nous informe donc sur des processus qui ont eu lieu il y a des centaines de milliers d'années, comme des évènements beaucoup plus récents il y a quelques dizaines ou centaines d'années. Il est aussi possible d'étudier l'ADN obtenu à partir d'échantillons vieux de plusieurs milliers d'années – c'est l'étude de l'ADN ancien ou « ADN fossile », ce qui nous permet notamment de découvrir à quoi ressemblait réellement la diversité génétique des populations dans un lointain passé.

Comprendre l'histoire évolutive des populations africaines est particulièrement intéressant puisque l'Homme moderne, *Homo sapiens*, est originaire et a longtemps évolué en Afrique avant de conquérir le reste du monde. Grâce en partie à la génétique des populations humaines actuelles, nous savons aujourd'hui que l'Homme moderne s'est répandu progressivement sur tous les

continents à partir de ce continent originel, et connaissons parfois les différentes routes et les modalités de ces expansions à des échelles géographiques très locales. L'histoire de l'Homme moderne en Afrique, avant et après la « sortie d'Afrique » (il y a ∼80 000 ans), n'est pas aussi bien connue, notamment car il y a moins de moyens pour la recherche en Afrique qu'en Eurasie ou en Amérique du nord. Nous savons toutefois que *Homo sapiens* s'est développé il y a environ 300 000-200 000 ans, vraisemblablement dans plusieurs régions d'Afrique. Le nombre de populations à cette époque, la façon dont elles étaient apparentées entre elles et avec la (ou les) population(s) qui a (ont) quitté l'Afrique ainsi qu'avec les populations contemporaines, sont autant de questions importantes auxquelles nous souhaitons répondre.

Des études ont montré que certaines populations sont cruciales pour mieux comprendre l'histoire de l'Afrique sub-saharienne avant la diffusion de l'agriculture, qui a profondément modifié le paysage génétique il y a environ 5 000 ans. Ces populations sont les Khoe-khoe et les San d'Afrique australe (collectivement appelés les Khoe-San) ; les chasseurs-cueilleurs de la forêt tropicale d'Afrique centrale (souvent appelés « Pygmées ») ; et les chasseurs-cueilleurs de l'est de l'Afrique.

Ma thèse porte sur certaines de ces populations, et j'ai utilisé des analyses de données génétiques à l'échelle du génome entier pour les étudier. Les techniques d'obtention de données génétiques ont beaucoup progressé cette dernière décennie. Une approche consiste en l'analyse d'un jeu de positions prédéterminées dans le génome grâce aux « puces à ADN ». C'est souvent cette technique, aujourd'hui très fiable et peu coûteuse, qui est utilisée pour générer les données génétiques des tests commerciaux abusivement dits « récréatifs ». Les puces à ADN sont employées par les chercheurs désireux d'étudier un grand nombre d'individus en même temps. Une autre approche se base sur l'observation de (presque) toutes les positions dans le génome ; c'est le séquençage du génome entier. Cett technique, plus onéreuse en terme de chimie moléculaire impliquée, nécessite aussi beaucoup plus de temps de travail et de ressources humaines pour préparer et analyser les données mais fournit massivement plus d'informations de très haute qualité.

Ainsi, dans le deuxième projet de ma thèse, j'étudie la façon de préparer les données de génomes entiers avant leur analyse. Il convient d'être particulièrement prudent en étudiant les populations africaines car elles sont plus diverses génétiquement que celles du reste du monde ; de surcroît, elles sont sous-représentées dans les bases de données génétiques ce qui rend leur analyse, faute de base de données de références fiable, plus ardue. Je montre que l'on peut préparer des séquences de populations africaines de la même façon que des séquences d'autres populations, mais qu'il vaut mieux reprendre l'analyse du début, plutôt que d'intégrer brutalement ces nouvelles données aux bases existantes. Un autre paramètre, le nombre de fois que chaque position du génome a été « lue » pendant le séquençage, se révèle plus important que l'origine géographique pour expliquer les différences entre échantillons.

En d'autres termes, la qualité du séquençage est encore un critère décisif pour découvrir la diversité génétique des populations africaines.

Il est commun de représenter l'histoire évolutive humaine par un arbre, dans lequel les ancêtres des Khoe-San représentent la première branche. Dans le premier projet de ma thèse, nous avons généré et analysé des génomes entiers pour 25 individus Khoe-San. Le moment où les ancêtres des Khoe-San et les ancêtres du reste des Hommes modernes se sont séparés a été estimé précédemment à 300 000-200 000 ans. Dans cette étude, nous ré-estimons cette date et obtenons des résultats similaires. Nous montrons également que la diversité génétique importante des Khoe-San est partiellement due à l'afflux de gènes de populations qui sont arrivées en Afrique australe durant les deux derniers millénaires et y ont introduit l'élevage et l'agriculture. Nous employons des méthodes qui estiment les changements de taille de population au cours du temps, et nous montrons que la taille de population des Khoe-San a commencé à diminuer il y a environ 100 000 ans. La même tendance est observée dans toutes les populations du monde ; nous proposons une explication possible de ce phénomène en émettant l'hypothèse qu'elle serait due aux conséquences d'un changement climatique ayant largement perturbé l'écologie des populations humaines de l'époque. J'ai simulé des données génétiques sous plusieurs modèles de goulots d'étranglement génétique, afin d'étudier comment la méthode d'estimation des changements de taille de population réagit dans ces différents cas. Les résultats confirment que le phénomène génétique observé sur nos données réelles est probablement authentique, et pas le fruit de biais méthodologiques. Nous avons également identifié des gènes qui ont été potentiellement sélectionnés, et nous montrons que ces gènes sont impliqués dans la réponse immunitaire ainsi que dans le motilité du sperme, le développement et le fonctionnement du cerveau, le développement des muscles et le métabolisme.

Dans le troisième projet, nous avons généré des génomes entiers pour des populations d'Afrique centrale – dont des chasseurs-cueilleurs – et nous les avons comparés à des séquences d'autres populations. Nous avons construit 24 modèles qui décrivent l'évolution de l'Homme en Afrique, d'une population ancestrale unique à cinq populations contemporaines, tout en incluant la possibilité de métissages entre lignées tout au long de l'histoire. En comparant nos données réelles et des données obtenues par simulations nous avons sélectionné le modèle le plus probable. Cette analyse nous livre plusieurs informations. D'abord, nous montrons qu'il est important d'inclure la possibilité de flux de gènes importants entre populations pour expliquer la diversité génétique actuelle, soit l'occurrence de métissages génétiques substantiels tout au long de l'histoire évolutive des populations africaines. Ensuite, nous confirmons, à l'échelle du génome et grâce à des méthodes statistiques puissantes reposant sur plusieurs techniques d'apprentissage machine profond (« deep learning » en anglais), que les modèles les plus probables sont ceux où la première séparation dans l'histoire des populations modernes est entre les ancêtres

des Khoe-San et les ancêtres du reste des Hommes modernes (représentés dans nos modèles par des populations d'Afrique centrale). Finalement, nous démontrons que la séparation des Khoe-San en deux groupes est plus ancienne que la séparation des chasseurs-cueilleurs de la forêt tropicale en deux groupes à l'Est et l'Ouest du bassin Congolais. Après avoir identifié les meilleurs modèles, nous estimons les paramètres des modèles (par exemple les temps de séparation entre populations).

Dans mon dernier projet, je me suis intéressée à des populations de pêcheurs de Zambie que l'on appelle les BaTwa. Différentes données, d'archéologie par exemple, suggèrent que les ancêtres de BaTwa pourraient avoir été une population locale de chasseurs-cueilleurs. Nous avons généré des données de puces à ADN pour deux populations de BaTwa et les avons comparées à des données similaires pour d'autres populations. Nous montrons que les BaTwa ont un composant génétique qui ressemble à celui de chasseurs-cueilleurs contemporains, en particulier les Khoe-San. Ils ont aussi un composant (plus important) qui ressemble à celui des populations qui ont introduit l'agriculture dans la région correspondant actuellement à la Zambie.

Les différents projets de ma thèse apportent de nouveaux éléments aux études déjà publiées sur l'histoire évolutive d'*Homo sapiens* en Afrique, en particulier sur la période précédant la diffusion de l'agriculture. Ils confirment d'une part, la nécessité d'étudier des populations peu connues ou représentées par un nombre restreint d'individus, d'autre part la complexité de l'histoire d'*Homo sapiens* et enfin, l'utilité d'évaluer toujours plus de modèles pour la comprendre encore mieux.

9. Acknowledgments

Five years is a long journey, even if September 2015 feels rather close at times! Many people contributed one way or another to my PhD and helped me along the way, and I would like to thank all of them.

First, of course, I am grateful to all subjects who contributed genetic data, and to the researchers who collected samples and shared their knowledge of the field, as none of this would have been possible otherwise.

Three of my supervisors, Mattias Jakobsson, Carina Schlebusch and Paul Verdu (and some more people), planned the "African whole genome project" (translated here in Paper III), a couple of years before I started my PhD. I would change many things if I were to redo my PhD (for example, I would write a shorter thesis!), but I would definitely not change project - thank you for this great idea, and for trusting me (and my initial lack of knowledge) with this amazing dataset!

I would like to thank Mattias, my main supervisor, for his enthusiasm for all new projects, conferences, various commitments that I became involved in during these five years. You gave me a lot of independence - it felt too much at times - but I am grateful that you always let me judge how, when and where it was best for me to work. It is also amazing how you can take any text and make it so much clearer!

Carina introduced me to human population genetic studies (and to the command line!) in 2013, during an internship - back then I did not recognise how lucky I had been to be involved in such a nice, straightforward project! This for sure contributed to my willingness to continue. Thank you for listening to me when I needed it most, and for sharing your knowledge about Southern Africa.

Paul gave a conference at my university (back in 2012...) that sparked my interest about Central African populations. I visited the lab in Paris when I was looking for a master internship, but since I wanted to go abroad, I looked for co-authors with exotic sounding names, and found the Jakobsson's lab. But I made it back somehow! Paul, you are an example of scientific rigor to me, and I have piles of notes from our conversations about Central African populations or human evolutionary genetics in general. Your projects are so inspiring! Thank you for re-motivating me and stepping up when I was despairing of making concrete progresses. Thank you to everyone who welcomed me in the lab in Paris during my visits, in particular Evelyne Heyer for her optimism!

Per Sjödin accepted to become my fourth supervisor a couple of years ago. Thank you for running analyses on a short notice, for your questions, for replacing most of my "which" by "that", and generally for commenting rapidly

my manuscripts and pointing out all of the convoluted or incomprehensible sentences! Thank you as well for checking on me and listening.

I like rhythms, and I very grateful to Anna Johansson and Hanna Edlund, whom I met regularly at two different periods of my PhD. Anna, as my "bioinformatics advisor", discussed with me of the states of my projects; thank you for that! Preparing for our meetings motivated me and contributed a lot to Paper II. Hanna, our weekly meetings helped me tremendously during the last four months of my thesis. Thank you so much.

The Human Evolution program has changed a lot in five years - in fact, it has been its own program only for half of that time. People have come and gone, but it has always be nice to share a work space, fika, lunch, meetings, outings and more. Thank you to everyone for that. I hope we meet again in a more regular environment than that of spring 2020! I am particulary grateful to Luciana - I am glad that we are office mate since we moved and that I got to know you well! Thank you for the cakes, for our daily motivation messages when we worked from home, for taking over the organization of Mário's spex, for the holidays in Portugal, for your support in the end of my PhD, and so much more. Mário, my "research twin" who took the leap six months earlier, thank you for showing me the way (I missed smarties cookies though, it is a pity that you left!). TJ, I see less of you since you changed office, but I always appreciate your wise comments when I am annoyed at anything and everything; thank you for answering all of my Y chromosome questions too! Thank you to all other PhD students from the program - now that I went through writing my thesis, I appreciate much better what it means and admire anyone who went through it! Thank you Lucie, Agnès, Nina, TJ, Alex and Mário for sharing your experiences, and good luck to Luciana, Rickard, James, Pedro and Imke.

Even if my thesis felt like a solitary effort at times, it was just an impression, and I would like to thank my co-authors for their various contributions to the papers in this thesis: Lawrence Barham, Luis Barreiro, Michael de Jongh, Lucie Gattepaille, Torsten Günther, Barry Hewlett, Evelyne Heyer, Nina Hollfelder, Mattias Jakobsson, Anna Johansson, Romain Laurent, Marlize Lombard, Helena Malmström, George Mudenda, Thijessen Naidoo, George Perry, Carina Schlebusch, Douglas Scofield, Per Sjödin, Agnès Sjöstrand, Himla Soodyall, Paul Verdu, Mário Vicente, Jingzi Xu and Panagiotis Zervakis. Special thanks to Lawrence Barham who contributed enormously to the project about the BaTwa populations from Zambia; to Barry Hewlett who shared his knowledge about the Bi.Aka, Bi.Aka Mbati, and Mbati; and to Panagiotis Zervakis, my first master student, who did a great job at investigating the uniparental markers that I was not so keen on exploring! Thank you as well to people who answered my questions, in particular Cesar Fortes Lima and Flora Jay.

July in Sweden is all about summer and holidays, so it was perhaps not the best timing to wrap-up a thesis. Therefore, I am extra grateful to everyone

who read, commented, and otherwise contributed to the revision of my thesis; besides my supervisors, a big thank you to Hanna, Carro, Luciana, TJ, Nina, Karin, Sonja, and Katell.

Over the years, I organised a number of projects, from collective gifts to a game explaining DNA degradation based on Duplo, as well as more "official" teaching projects. Thank you to everyone who joined, even when the deadlines were very short! It would not be half the fun (if possible at all) to conduct such projects alone.

The Evolutionary Biology Center was not only a place of work, and I met many friends there. Thank you for the knitting club (and the derived drawing club), the Friday "beverages", the occasional hikes, swims, and travels. In particular thank you Nina, Arielle, Linnéa, Frida, Martin, Rhiannon, Tuuli, Karin, Berrit, Imke, Anna, Ayça, Luciana, João, Rickard, Alex, James, Homa, Cécile, Lore, Sarina, and Manolis, for many occasions to leave the computer earlier and enjoy nice company! Somehow the number of babies increased over the years, and babies happen to be a good way to relax and focus on something else for me, so thank you to Ylves & Oskar, Julie & Lili, Astrid and Yannik, and their parents for letting me play with them!

I also met nice people from a bit further away at the faculty during the time I sat on committees - thank you for interesting meetings on many topics, and for letting me learn from you, especially to Anna Frost, Katharina Rudisch and Frauke Augstein.

March through July 2020 were very special: not only did I write my thesis, I did it at my "home office" next to the dinner table because of a global pandemic. The end of my PhD was much harder than I expected, and every mark of support was much appreciated. I had the pleasure to have a flatmate for a few months during that period - thank you, Lucie, it was great to have someone to talk to constantly (and in French!) and to cook together. Luciana, thank you for taking care of me. Berrit and Karin, I am very glad for our "Nachbarn" group, for planned and unplanned meetings, and I look forward to watch the rest of the Harry Potter movies with you soon! I also took advantage of free online content such as yoga courses and inspiring podcasts, and this definitely contributed to my well-being - thank you to the people who share their skills! Merci pour les échanges de mails Iris, ça m'a fait du bien de vous savoir toi et Mattias en proie plus ou moins aux mêmes préoccupations au même moment! On y est presque!

I planned to write an entire page about my thesis' cover (which is a bag in reality), but I think my thesis is already long enough as it is. I will only say that the design was inspired by the kjolsäckar from Dalarna, small bags used as pockets by women, and the beautiful book about them by Helena Bengtsson, "Sy väskor vackra som smycken". The mask is based on an insignia of the Bwami society (Lega people, DRC), sketched at a museum. The motives on the back represent prior and posterior distributions, and the heavy rain falling on Uppsala on July 5, 2020! (Which seems appropriate for a rainforest-focused

thesis.) Thank you to Martin Lind for photographing and thereby giving me a deadline to finish it.

Sonja, I am glad we started talking after spelling our personnumren to one another in Swedish class! It is always a pleasure to meet you and Stephan and to watch Kilian grow. Vielen Dank für die Hilfe!

In the past few years, I became involved in a gardening association and attended a lot of yoga classes. Going to the garden or to the yoga school were always nice opportunities to take a break, clear my thoughts from work, and occasionally talk Swedish! I am glad to have found "Matparken" and "Uppsala Yogaskola".

Thank you for everyone who visited me in Uppsala - my close family, but also Mathilde, Aude, Pablo, Noémie and her dad. It is always nice to have a reason to go to the cafés, to explore new places, or to rediscover the old ones!

Merci Lucile (et Robert), Mathilde (et Vincent), Noémie, Aude, Pauline, Myriam, Lucille, Pablo, Iris, Matthias, Ann (et Mattias), Pierrick et Marie, pour votre présence, vos messages, lettres ou appels, les moments passés ensemble souvent sur la route des vacances ou d'une conférence, les échanges autour de nos thèses et aussi, heureusement, d'autres choses! Il me tarde de vous revoir en ayant un peu plus de temps devant nous! En particulier, merci à Mathilde & Vincent - les photos d'Amandine et de vous ont été des bouffées de bonheur depuis avril, ainsi qu'une motivation certaine à boucler tout ça au plus vite pour aller vous voir et rencontrer Amandine avant qu'elle ne soit géante!

Merci, enfin, à mes parents, à Bleuenn et Tudi, à ma mamie et à ma tante. Merci pour votre soutien, à distance et sur place, pour les messages quotidiens de Maman, les appels plus irréguliers, les courriers de Mamie et de Gaëla, pour vos visites ici, y compris pour me déménager! Votre soutien m'est indispensable.

10. References

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

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–64, sep 2009.

I. Alves, M. Coelho, C. Gignoux, A. Damasceno, A. Prista, and J. Rocha. 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, feb 2011.

A. Ameur, J. Dahlberg, P. Olason, F. Vezzi, R. Karlsson, M. Martin, J. Viklund, A. K. Kähäri, P. Lundin, H. Che, J. Thutkawkorapin, J. Eisfeldt, S. Lampa, M. Dahlberg, J. Hagberg, N. Jareborg, U. Liljedahl, I. Jonasson, Å. Johansson, L. Feuk, J. Lundeberg, A.-C. Syvänen, S. Lundin, D. Nilsson, B. Nystedt, P. K. Magnusson, and U. Gyllensten. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*, 25(11):1253–1260, 2017.

M. L. Antonio, Z. Gao, H. M. Moots, M. Lucci, F. Candilio, S. Sawyer, V. Oberreiter, D. Calderon, K. Devitofranceschi, R. C. Aikens, et al. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*, 366(6466):708–714, 2019.

O. T. Avery, C. M. Macleod, and M. McCarty. 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. *The Journal of experimental medicine*, 79(2):137–58, feb 1944.

S. Bahuchet. L'invention des Pygmées (Inventing Pygmies). *Cahiers d'Études Africaines*, 33:153–181, 1993.

S. Bahuchet. Changing language, remaining pygmy. *Human biology*, 84(1):11–43, feb 2012.

S. Baichoo, Y. Souilmi, S. Panji, G. Botha, A. Meintjes, S. Hazelhurst, H. Bendou, E. de Beste, P. T. Mpangase, O. Souiai, M. Alghali, L. Yi, B. D. O'Connor, M. Crusoe, D. Armstrong, S. Aron, F. Joubert, A. E. Ahmed, M. Mbiyavanga, P. van Heusden, L. E. Magosi, J. Zermeno, L. S. Mainzer, F. M. Fadlelmola, C. V. Jongeneel, and N. Mulder. Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. *BMC Bioinformatics*, 19(1):457, dec 2018.

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

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

C. Barbieri, A. Butthof, K. Bostoen, and B. Pakendorf. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *European Journal of Human Genetics*, 21(4):430–436, apr 2013.

C. Barbieri, T. Güldemann, C. Naumann, L. Gerlach, F. Berthold, H. Nakagawa, S. W. Mpoloka, M. Stoneking, and B. Pakendorf. Unraveling the complex maternal history of Southern African Khoisan populations. *American Journal of Physical Anthropology*, 153(3):435–48, mar 2014.

C. Batini, J. Lopes, D. M. Behar, F. Calafell, L. B. Jorde, L. van der Veen, L. Quintana-Murci, G. Spedini, G. Destro-Bisol, and D. Comas. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Molecular Biology and Evolution*, 28(2):1099–110, feb 2011.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–35, dec 2002.

A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J. F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, and C. Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), mar 2020.

K. Bostoen, B. Clist, C. Doumenge, R. Grollemund, J.-M. Hombert, J. K. Muluwa, J. Maley, R. Blench, P. Di Carlo, J. Good, et al. Middle to late Holocene Paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa. *Current Anthropology*, 56(3):367–368, 2015.

D. Y. Brandt, V. R. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, and D. Meyer. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes project phase I data. *G3: Genes, Genomes, Genetics*, 5(5): 931–941, 2015.

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

K. W. Broman and J. L. Weber. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. *American Journal of Human Genetics*, 65(6):1493–1500, dec 1999.

S. R. Browning and B. L. Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2015.

G. B. Busby, G. Band, Q. Si Le, M. Jallow, E. Bougama, V. D. Mangano, L. N. Amenga-Etego, A. Enimil, T. Apinjoh, C. M. Ndila, A. Manjurano, V. Nyirongo, O. Doumba, K. A. Rockett, D. P. Kwiatkowski, C. C. Spencer, and Malaria Genomic Epidemiology Network. Admixture into and within sub-Saharan Africa. *eLife*, 5, jun 2016.

H. M. Cann. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566): 261b–262, apr 2002.

D. Carpenter, S. Dhar, L. M. Mitchell, B. Fu, J. Tyson, N. A. Shwan, F. Yang, M. G. Thomas, and J. A. Armour. Obesity, starch digestion and amylase: association between copy number variants at human salivary (amy1) and pancreatic (amy2) amylase genes. *Human molecular genetics*, 24(12):3472–3480, 2015.

L. L. Cavalli-Sforza, A. C. Wilson, C. R. Cantor, R. M. Cook-Deegan, and M.-C. King. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. *Genomics (San Diego, Calif.)*, 11(2): 490–491, 1991.

L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton University Press, 1994.

L. Chen, A. B. Wolf, W. Fu, L. Li, and J. M. Akey. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell*, 180(4):677–687.e16, feb 2020.

A. Choudhury, M. Ramsay, S. Hazelhurst, S. Aron, S. Bardien, G. Botha, E. R. Chimusa, A. Christoffels, J. Gamieldien, M. J. Sefid-Dashti, F. Joubert, A. Meintjes, N. Mulder, R. Ramesar, J. Rees, K. Scholtz, D. Sengupta, H. Soodyall, P. Venter, L. Warnich, and M. S. Pepper. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, 8(1):2062, dec 2017.

P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71, apr 2010.

N. G. Crawford, D. E. Kelly, M. E. Hansen, M. H. Beltrame, S. Fan, S. L. Bowman, E. Jewett, A. Ranciaro, S. Thompson, Y. Lo, et al. Loci associated with skin pigmentation identified in African populations. *Science*, 358(6365), 2017.

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

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, aug 2011.

C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, Albemarle Street, London, 1st edition edition, 1859.

C. de Filippo, P. Heyn, L. Barham, M. Stoneking, and B. Pakendorf. Genetic perspectives on forager-farmer interaction in the Luangwa Valley of Zambia. *American Journal of Physical Anthropology*, 141(3):NA–NA, mar 2009.

C. de Filippo, C. Barbieri, M. Whitten, S. W. Mpoloka, E. Drofn Gunnarsdóttir, K. Bostoen, T. Nyambe, K. Beyer, H. Schreiber, P. De Knijff, D. Luiselli, M. Stoneking, and B. Pakendorf. Y-Chromosomal Variation in Sub-Saharan Africa: Insights Into the History of Niger-Congo Groups. *Molecular Biology and Evolution*, 28(3):1255–1269, 2010.

C. de Filippo, K. Bostoen, M. Stoneking, and B. Pakendorf. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3256–3263, aug 2012.

H. De Villiers and L. Fatti. The antiquity of the Negro. *South African Journal of Science*, 78:321–333, 1982.

E. de Wit, W. Delport, C. E. Rugamika, A. Meintjes, M. Möller, P. D. van Helden, C. Seoighe, and E. G. Hoal. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, 128(2): 145–153, 2010.

J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, 2009.

M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, may 2011.

R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, and C. A. Reid. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, N.Y.)*, 327(5961):78–81, jan 2010.

B. L. Dumont and B. A. Payseur. Evolution of the genomic rate of recombination in mammals. *Evolution*, 62(2):276–294, feb 2008.

A. Durvasula and S. Sankararaman. Recovering signals of ghost archaic introgression in African populations. *Science Advances*, 6(7):eaax5097, feb 2020.

J. Eisfeldt, G. Mårtensson, A. Ameur, D. Nilsson, and A. Lindstrand. Discovery of Novel Sequences in 1,000 Swedish Genomes. *Molecular Biology and Evolution*, 37(1):18–30, 2020.

L. Excoffier and M. Foll. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, may 2011.

L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. Robust demographic inference from genomic and SNP data. *PLoS genetics*, 9(10): e1003905, oct 2013.

K. A. Fakhro, M. R. Staudt, M. D. Ramstetter, A. Robay, J. A. Malek, R. Badii, A. A.-N. Al-Marri, C. Abi Khalil, A. Al-Shakaki, O. Chidiac, D. Stadler, M. Zirie, A. Jayyousi, J. Salit, J. G. Mezey, R. G. Crystal, and J. L. Rodriguez-Flores. The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Human genome variation*, 3:16016, 2016.

S. Fan, D. E. Kelly, M. H. Beltrame, M. E. B. Hansen, S. Mallick, A. Ranciaro, J. Hirbo, S. Thompson, W. Beggs, T. Nyambo, S. A. Omar, D. W. Meskel, G. Belay, A. Froment, N. Patterson, D. Reich, and S. A. Tishkoff. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, 20(1):82, dec 2019.

Z. Fan, P. Silva, I. Gronau, S. Wang, A. S. Armero, R. M. Schweizer, O. Ramirez, J. Pollinger, M. Galaverni, D. Ortega Del-Vecchyo, L. Du, W. Zhang, Z. Zhang,

J. Xing, C. Vilà, T. Marques-Bonet, R. Godinho, B. Yue, and R. K. Wayne. Worldwide patterns of genomic variation and admixture in gray wolves. *Genome research*, 26(2):163–73, feb 2016.

J. N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128(2):415–423, oct 2005.

R. A. Fisher. XXI.—On the Dominance Ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341, sep 1923.

C. Fortes-Lima, A. Gessain, A. Ruiz-Linares, M.-C. Bortolini, F. Migot-Nabias, G. Bellis, J. V. Moreno-Mayar, B. N. Restrepo, W. Rojas, E. Avendaño-Tamayo, et al. Genome-wide ancestry and demographic history of African-descendant maroon communities from French Guiana and Suriname. *The American Journal of Human Genetics*, 101(5):725–736, 2017.

R. Fregel, F. L. Méndez, Y. Bokbot, D. Martín-Socas, M. D. Camalich-Massieu, J. Santana, J. Morales, M. C. Avila-Arcos, P. A. Underhill, B. Shapiro, G. Wojcik, M. Rasmussen, A. E. Soares, J. Kapp, A. Sockell, F. J. Rodríguez-Santos, A. Mikdad, A. Trujillo-Mederos, and C. D. Bustamante. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 115(26):6774–6779, jun 2018.

E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al. The international HapMap project. *Nature*, 426, dec 2003.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

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

I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10):1031–1034, oct 2011.

R. Grün, J. S. Brink, N. A. Spooner, L. Taylor, C. B. Stringer, R. G. Franciscus, and A. S. Murray. Direct dating of Florisbad hominid. *Nature*, 382(6591):500–501, 1996.

T. Günther and M. Jakobsson. Population Genomic Analyses of DNA from Ancient Remains. In *Handbook of Statistical Genomics: Two Volume Set*, pages 295–40.

Wiley Online Library, 2019.
D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M. O. Pollard, A. Choudhury, G. R. S. Ritchie, Y. Xue, J. Asimit, R. N. Nsubuga, E. H. Young, C. Pomilla, K. Kivinen, K. Rockett, A. Kamali, A. P. Doumatey, G. Asiki, J. Seeley, F. Sisay-Joof, M. Jallow, S. Tollman, E. Mekonnen, R. Ekong, T. Oljira, N. Bradman, K. Bojang, M. Ramsay, A. Adeyemo, E. Bekele, A. Motala, S. A. Norris, F. Pirie, P. Kaleebu, D. Kwiatkowski, C. Tyler-Smith, C. Rotimi, E. Zeggini, and M. S. Sandhu. The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517 (7534):327–332, jan 2015.
R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, oct 2009.
M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, and J. D. Wall. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37):15123–8, sep 2011.
G. Hansson, S. G. Barriere, J. Gurell, M. Lindholm, P. Lundin, and M. Wikgren. The Swedish Research Barometer 2019 - The Swedish Research System in International Comparison. Technical Report 3.5-2019-102, Swedish Research Council, 2019.
B. M. Henn, C. R. Gignoux, M. Jobin, J. M. Granka, J. M. Macpherson, J. M. Kidd, L. Rodríguez-Botigué, S. Ramachandran, L. Hon, A. Brisbin, A. A. Lin, P. A. Underhill, D. Comas, K. K. Kidd, P. J. Norman, P. Parham, C. D. Bustamante, J. L. Mountain, and M. W. Feldman. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, 108(13):5154–62, mar 2011.
B. M. Henn, T. E. Steele, and T. D. Weaver. Clarifying distinct models of modern human origins in Africa. *Current Opinion in Genetics & Development*, 53: 148–156, dec 2018.
R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, M. Przeworski, et al. Classic selective sweeps were rare in recent human evolution. *science*, 331(6019):920–924, 2011.
B. S. Hewlett. *Hunter-gatherers of the Congo Basin : cultures, histories and biology of African Pygmies*. Routledge, 2014.
E. Heyer. Race and racism in France. *Journal of Anthropological Sciences*, 95: 307–310, 2017.
E. Heyer, R. Chaix, S. Pavard, and F. Austerlitz. Sex-specific demographic behaviours that shape human genomic variation. *Molecular ecology*, 21(3): 597–612, 2012.
N. Hollfelder, J. C. Erasmus, R. Hammaren, M. Vicente, M. Jakobsson, J. M. Greeff, and C. M. Schlebusch. Patterns of African and Asian admixture in the Afrikaner population of South Africa. *BMC biology*, 18(1):1–13, 2020.
K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, 10(9): 639–650, 2009.
P. Hsieh, K. R. Veeramah, J. Lachance, S. A. Tishkoff, J. D. Wall, M. F. Hammer, and R. N. Gutenkunst. Whole-genome sequence analyses of Western Central African

Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome research*, 26(3):279–90, mar 2016a.

P. Hsieh, A. E. Woerner, J. D. Wall, J. Lachance, S. A. Tishkoff, R. N. Gutenkunst, and M. F. Hammer. 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, mar 2016b.

M. Huang, J. Tu, and Z. Lu. Recent Advances in Experimental Whole Genome Haplotyping Methods. *International Journal of Molecular Sciences*, 18(12):1944, sep 2017.

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

International HapMap Consortium and others. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

International HapMap Consortium and others. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851, 2007.

M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, feb 2008.

F. Jay, S. Boitard, and F. Austerlitz. An ABC Method for Whole-Genome Sequence Data: Inferring Paleolithic and Neolithic Human Expansions. *Molecular Biology and Evolution*, 36(7):1565–1579, 2019.

M. A. Jobling, E. Hollox, M. Hurles, T. Kivisild, and C. Tyler-Smith. *Human Evolutionary Genetics (2nd edition)*. Garland Science, New York, 2013.

D. V. Joiris. The framework of Central African hunter-gatherers and neighbouring societies. *African Study Monographs*, 2003.

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

M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3): 235–248, 1982.

M. Kirin, R. McQuillan, C. S. Franklin, H. Campbell, P. M. Mckeigue, and J. F. Wilson. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE*, 5(11), 2010.

A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

E. Kowal, B. Llamas, and S. Tishkoff. Data-sharing for indigenous peoples. *Nature*, 546(7659):474–474, 2017.

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

J. Lachance, B. Vernot, C. Elbers, B. Ferwerda, A. Froment, J.-M. Bodo, G. Lema, W. Fu, T. Nyambo, T. Rebbeck, K. Zhang, J. Akey, and S. Tishkoff. Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. *Cell*, 150(3):457–469, aug 2012.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409, feb 2001.

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

K. Landsteiner. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Centralblatt fuer Bakteriologie, Parasitenkunde und Infektionskrankheiten*, 27:357–62, 1900.

D. J. Lawson, L. van Dorp, and D. Falush. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1):1–11, dec 2018.

R. C. Lewontin. The apportionment of human diversity. In *Evolutionary biology*, pages 381–398. Springer, 1972.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009.

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

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, aug 2009.

J. Li, H. Li, M. Jakobsson, S. Li, P. Sjödin, and M. Lascoux. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, 21(1):28–44, jan 2012.

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

L. Lieberman and F. L. C. Jackson. Race and Three Models of Human Origin. *American Anthropologist*, 97(2):231–242, jun 1995.

G. Lightbody, V. Haberland, F. Browne, L. Taggart, H. Zheng, E. Parkes, and J. K. Blayney. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings in bioinformatics*, 20(5):1795–1811, 2019.

M. Lipson, I. Ribot, S. Mallick, N. Rohland, I. Olalde, N. Adamski, N. Broomandkhoshbacht, A. M. Lawson, S. López, J. Oppenheimer, K. Stewardson, R. N. Asombang, H. Bocherens, N. Bradman, B. J. Culleton, E. Cornelissen, I. Crevecoeur, P. de Maret, F. L. M. Fomine, P. Lavachery, C. M. Mindzie, R. Orban, E. Sawchuk, P. Semal, M. G. Thomas, W. Van Neer, K. R. Veeramah, D. J. Kennett, N. Patterson, G. Hellenthal, C. Lalueza-Fox,

S. MacEachern, M. E. Prendergast, and D. Reich. Ancient West African foragers in the context of African population history. *Nature*, 577(7792):665–670, jan 2020.

M. Lopez, A. Kousathanas, H. Quach, C. Harmant, P. Mouguiama-Daouda, J.-M. Hombert, A. Froment, G. H. Perry, L. B. Barreiro, P. Verdu, E. Patin, and L. Quintana-Murci. The demographic history and mutational load of African hunter-gatherers and farmers. *Nature Ecology & Evolution*, 2(4):721–730, apr 2018.

B. Lorente-Galdos, O. Lao, G. Serra-Vidal, G. Santpere, L. F. K. Kuderna, L. R. Arauna, K. Fadhlaoui-Zid, V. N. Pimenoff, H. Soodyall, P. Zalloua, T. Marques-Bonet, and D. Comas. 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 Biology*, 20(1):77, dec 2019.

E. Macholdt, V. Lede, C. Barbieri, S. Mpoloka, H. Chen, M. Slatkin, B. Pakendorf, and M. Stoneking. Tracing Pastoralist Migrations to Southern Africa with Lactase Persistence Alleles. *Current Biology*, 24(8):875–879, apr 2014.

A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvåg, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F.-X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207–214, oct 2016.

S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, and D. Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, oct 2016.

B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American journal of human genetics*, 93(2):278–88, aug 2013.

N. Marchi, T. Hegay, P. Mennecier, M. Georges, R. Laurent, M. Whitten, P. Endicott, A. Aldashev, C. Dorzhu, F. Nasyrova, et al. Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *American Journal of Physical Anthropology*, 162(4):627–640, 2017.

N. Mather, S. M. Traves, and S. Y. Ho. A practical introduction to sequentially markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, 10(1):579–589, 2020.

O. Mazet, W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi. On the importance of being structured: Instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity*, 116(4):362–371, apr 2016.

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

G. Mendel. *Versuche über Plflanzenhybriden.* Verhand- lungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, 1866.

P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792, sep 1978.

M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andres, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Paabo. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338(6104):222–226, oct 2012.

Y. Miar, M. Sargolzaei, and F. S. Schenkel. A comparison of different algorithms for phasing haplotypes using Holstein cattle genotypes and pedigree data. *Journal of Dairy Science*, 100(4):2837–2849, apr 2017.

M. Mielczarek and J. Szyda. Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics*, 57(1):71–79, feb 2016.

M. Morey, A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, and J. A. Cocho. A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110(1-2):3–24, sep 2013.

A. G. Morris and I. Ribot. Morphometric cranial identity of prehistoric Malawians in the light of sub-Saharan African diversity. *American Journal of Physical Anthropology*, 130(1):10–25, may 2006.

K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1:263–73, 1986.

M. Nagasaki, J. Yasuda, F. Katsuoka, N. Nariai, K. Kojima, Y. Kawai, Y. Yamaguchi-Kabata, J. Yokozawa, I. Danjoh, S. Saito, Y. Sato, T. Mimori, K. Tsuda, R. Saito, X. Pan, S. Nishikawa, S. Ito, Y. Kuroki, O. Tanabe, N. Fuse, S. Kuriyama, H. Kiyomoto, A. Hozawa, N. Minegishi, J. Douglas Engel, K. Kinoshita, S. Kure, N. Yaegashi, Y. ToMMo Japanese Reference Panel Project, and M. Yamamoto. Rare variant discovery by deep whole-genome sequencing of

1,070 Japanese individuals. *Nature communications*, 6:8018, aug 2015.

T. Naidoo, J. Xu, M. Vicente, H. Malmström, H. Soodyall, M. Jakobsson, and C. M. Schlebusch. Y-Chromosome Variation in Southern African Khoe-San Populations Based on Whole-Genome Sequences. *Genome Biology and Evolution*, 12(7): 1031–1039, 07 2020.

M. Nei and A. K. Roychoudhury. Sampling variances of heterozygosity and genetic distance. *Genetics*, 76(2), 1974.

R. Nielsen and M. Slatkin. *An introduction to population genetics*. Sunderland, MA: Sinauer Associates., 2013.

R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark. Recent and ongoing selection in the human genome. *Nature reviews. Genetics*, 8(11):857–68, nov 2007.

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

J. Novembre. Pritchard, Stephens, and Donnelly on Population Structure. *Genetics*, 204(2):391–393, 2016.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, nov 2008.

Y. Okada, Y. Momozawa, S. Sakaue, M. Kanai, K. Ishigaki, M. Akiyama, T. Kishikawa, Y. Arai, T. Sasaki, K. Kosaki, M. Suematsu, K. Matsuda, K. Yamamoto, M. Kubo, N. Hirose, and Y. Kamatani. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature Communications*, 9(1):1631, dec 2018.

H. Oota, W. Settheetham-Ishida, D. Tiwawech, T. Ishida, and M. Stoneking. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature genetics*, 29(1):20–21, 2001.

L. Pagani, D. J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, J. D. Wall, A. Cardona, R. Mägi, M. A. W. Sayres, S. Kaewert, C. Inchley, C. L. Scheib, M. Järve, M. Karmin, G. S. Jacobs, T. Antao, F. M. Iliescu, A. Kushniarevich, Q. Ayub, C. Tyler-Smith, Y. Xue, B. Yunusbayev, K. Tambets, C. B. Mallick, L. Saag, E. Pocheshkhova, G. Andriadze, C. Muller, M. C. Westaway, D. M. Lambert, G. Zoraqi, S. Turdikulova, D. Dalimova, Z. Sabitov, G. N. N. Sultana, J. Lachance, S. Tishkoff, K. Momynaliev, J. Isakova, L. D. Damba, M. Gubina, P. Nymadawa, I. Evseeva, L. Atramentova, O. Utevska, F.-X. Ricaut, N. Brucato, H. Sudoyo, T. Letellier, M. P. Cox, N. A. Barashkov, V. Škaro, L. Mulahasanovic´, D. Primorac, H. Sahakyan, M. Mormina, C. A. Eichstaedt, D. V. Lichman, S. Abdullah, G. Chaubey, J. T. S. Wee, E. Mihailov, A. Karunas, S. Litvinov, R. Khusainova, N. Ekomasova, V. Akhmetova, I. Khidiyatova, D. Marjanović, L. Yepiskoposyan, D. M. Behar, E. Balanovska, A. Metspalu, M. Derenko, B. Malyarchuk, M. Voevoda, S. A. Fedorova, L. P. Osipova, M. M. Lahr, P. Gerbault, M. Leavesley, A. B. Migliano, M. Petraglia, O. Balanovsky, E. K. Khusnutdinova, E. Metspalu, M. G. Thomas, A. Manica, R. Nielsen, R. Villems, E. Willerslev, T. Kivisild, and M. Metspalu. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538(7624):238–242, oct 2016.

B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.

E. Patin, G. Laval, L. B. Barreiro, A. Salas, O. Semino, S. Santachiara-Benerecetti, K. K. Kidd, J. R. Kidd, L. Van der Veen, J.-M. Hombert, A. Gessain, A. Froment, S. Bahuchet, E. Heyer, and L. Quintana-Murci. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics*, 5(4):e1000448, apr 2009.

E. Patin, K. J. Siddle, G. Laval, H. Quach, C. Harmant, N. Becker, A. Froment, B. Régnault, L. Lemée, S. Gravel, J.-M. Hombert, L. Van der Veen, N. J. Dominy, G. H. Perry, L. B. Barreiro, P. Verdu, E. Heyer, and L. Quintana-Murci. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications*, 5:3163, feb 2014.

E. Patin, M. Lopez, R. Grollemund, P. Verdu, C. Harmant, H. Quach, G. Laval, G. H. Perry, L. B. Barreiro, A. Froment, E. Heyer, A. Massougbodji, C. Fortes-Lima, F. Migot-Nabias, G. Bellis, J.-M. Dugoujon, J. B. Pereira, V. Fernandes, L. Pereira, L. Van der Veen, P. Mouguiama-Daouda, C. D. Bustamante, J.-M. Hombert, and L. Quintana-Murci. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, 356(6337):543–546, may 2017.

N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3): 1065–93, nov 2012.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, nov 1901.

G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256–1260, 2007.

B. M. Peter. Admixture, population structure, and f-statistics. *Genetics*, 202(4): 1485–1501, 2016.

J. K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Güldemann, B. Kure, S. W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P.-R. Loh, J. Lachance, J. Mountain, C. D. Bustamante, B. Berger, S. A. Tishkoff, B. M. Henn, M. Stoneking, D. Reich, and B. Pakendorf. The genetic prehistory of southern Africa. *Nature Communications*, 3:1143, oct 2012.

V. Plagnol and J. D. Wall. Possible Ancestral Structure in Human Populations. *PLoS Genetics*, 2(7):e105, jul 2006.

E. Porcu, S. Sanna, C. Fuchsberger, and L. G. Fritsche. Genotype Imputation in Genome-Wide Association Studies. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, jul 2013.

M. E. Prendergast, M. Lipson, E. A. Sawchuk, I. Olalde, C. A. Ogola, N. Rohland, K. A. Sirak, N. Adamski, R. Bernardos, N. Broomandkhoshbacht, et al. Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science*, 365(6448):eaaw6275, 2019.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, aug 2006.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, jan 2014.

P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2016.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75, sep 2007.

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

M. Ramsay, N. Crowther, E. Tambo, G. Agongo, V. Baloyi, S. Dikotope, X. Gómez-Olivé, N. Jaff, H. Sorgho, R. Wagner, C. Khayeka-Wandabwa, A. Choudhury, S. Hazelhurst, K. Kahn, Z. Lombard, F. Mukomana, C. Soo, H. Soodyall, A. Wade, S. Afolabi, I. Agorinya, L. Amenga-Etego, S. A. Ali, J. D. Bognini, R. P. Boua, C. Debpuur, S. Diallo, E. Fato, A. Kazienga, S. Z. Konkobo, P. M. Kouraogo, F. Mashinya, L. Micklesfield, S. Nakanabo-Diallo, B. Njamwea, E. Nonterah, S. Ouedraogo, V. Pillay, A. M. Somande, P. Tindana, R. Twine, M. Alberts, C. Kyobutungi, S. A. Norris, A. R. Oduro, H. Tinto, S. Tollman, and O. Sankoh. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global health, epidemiology and genomics*, 1:e20, nov 2016.

M. Rasmussen, S. L. Anzick, M. R. Waters, P. Skoglund, M. DeGiorgio, T. W. Stafford, S. Rasmussen, I. Moltke, A. Albrechtsen, S. M. Doyle, G. D. Poznik, V. Gudmundsdottir, R. Yadav, A.-S. Malaspinas, S. S. White, M. E. Allentoft, O. E. Cornejo, K. Tambets, A. Eriksson, P. D. Heintzman, M. Karmin, T. S. Korneliussen, D. J. Meltzer, T. L. Pierre, J. Stenderup, L. Saag, V. M. Warmuth, M. C. Lopes, R. S. Malhi, S. Brunak, T. Sicheritz-Ponten, I. Barnes, M. Collins, L. Orlando, F. Balloux, A. Manica, R. Gupta, M. Metspalu, C. D. Bustamante, M. Jakobsson, R. Nielsen, and E. Willerslev. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506(7487):225–9, feb 2014.

A. A. Regier, Y. Farjoun, D. E. Larson, O. Krasheninina, H. M. Kang, D. P. Howrigan, B.-J. Chen, M. Kher, E. Banks, D. C. Ames, A. C. English, H. Li, J. Xing, Y. Zhang, T. Matise, G. R. Abecasis, W. Salerno, M. C. Zody, B. M. Neale, and I. M. Hall. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications*, 9(1):4038, dec 2018.

D. Reich, K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, 2009.

D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053, 2010.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–97, may 2015.

A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, oct 2015.

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

C. Rotimi, A. Abayomi, A. Abimiku, V. M. Adabayeri, C. Adebamowo, E. Adebiyi, A. D. Ademola, A. Adeyemo, D. Adu, D. Affolabi, et al. Research capacity. Enabling the genomic revolution in Africa. *Science (New York, NY)*, 344(6190): 1346–1348, 2014.

M. Salter-Townshend and S. Myers. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, 212(3):869–889, jul 2019.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, dec 1977.

A. Scally and R. Durbin. Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753, sep 2012.

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

L. B. Scheinfeldt, S. Soi, C. Lambert, W. Y. Ko, A. Coulibaly, A. Ranciaro, S. Thompson, J. Hirbo, W. Beggs, M. Ibrahim, T. Nyambo, S. Omar, D. Woldemeskel, G. Belay, A. Froment, J. Kim, and S. A. Tishkoff. Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4166–4175, 2019.

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

C. M. Schlebusch. Issues raised by use of ethnic-group names in genome study. *Nature*, 464(7288):487–487, mar 2010.

C. M. Schlebusch and M. Jakobsson. Tales of human migration, admixture, and selection in Africa. *Annual Review of Genomics and Human Genetics*, 19: 405–428, 2018.

C. M. Schlebusch, M. De Jongh, and H. Soodyall. Different contributions of ancient mitochondrial and Y-chromosomal lineages in 'Karretjie people' of the Great Karoo in South Africa. *Journal of Human Genetics*, 56(9):623–630, sep 2011.

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

C. M. Schlebusch, F. Prins, M. Lombard, M. Jakobsson, and H. Soodyall. The disappearing San of southeastern Africa and their genetic affinities. *Human genetics*, 135(12):1365–1373, 2016.

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

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

V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H. C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C. S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, and D. M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, may 2017.

S. C. Schuster, W. Miller, A. Ratan, L. P. Tomsho, B. Giardine, L. R. Kasson, R. S. Harris, D. C. Petersen, F. Zhao, J. Qi, C. Alkan, J. M. Kidd, Y. Sun, D. I. Drautz, P. Bouffard, D. M. Muzny, J. G. Reid, L. V. Nazareth, Q. Wang, R. Burhans, C. Riemer, N. E. Wittekindt, P. Moorjani, E. A. Tindall, C. G. Danko, W. S. Teo, A. M. Buboltz, Z. Zhang, Q. Ma, A. Oosthuysen, A. W. Steenkamp, H. Oostuisen, P. Venter, J. Gajewski, Y. Zhang, B. F. Pugh, K. D. Makova, A. Nekrutenko, E. R. Mardis, N. Patterson, T. H. Pringle, F. Chiaromonte, J. C. Mullikin, E. E. Eichler, R. C. Hardison, R. A. Gibbs, T. T. Harkins, and V. M. Hayes. Complete Khoisan and Bantu genomes from southern Africa. *Nature*, 463(7283):943–947, feb 2010.

L. Ségurel and C. Bon. On the evolution of lactase persistence in humans. *Annual Review of Genomics and Human Genetics*, 18, 2017.

M. T. Seielstad, E. Minch, and L. L. Cavalli-Sforza. Genetic evidence for a higher female migration rate in humans. *Nature genetics*, 20(3):278–280, 1998.

A. Semo, M. Gay A-Vidal, C. Fortes-Lima, B. Er Enice Alard, S. Oliveira, J. Ao Almeida, A. Onio Prista, A. Damasceno, A.-M. Fehn, C. M. Schlebusch, and

J. Rocha. Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. *Molecular Biology and Evolution*, 32(2):406–416, 2020.
SFS. The act concerning the ethical review of research involving humans, 2003.
S. Sherry, M. Ward, M. Kholodov, J. Baker, L. Phan, E. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1): 308–311, jan 2001.
P. Sjödin, J. McKenna, and M. Jakobsson. Estimating divergence times from DNA sequences. *under revision*, 2020.
P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomani-Chindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, and D. Reich. Reconstructing Prehistoric African Population Structure. *Cell*, 171(1):59–71.e21, sep 2017.
S. Song, E. Sliwerska, S. Emery, and J. M. Kidd. Modeling Human Population Separation History Using Physically Phased Genomes. *Genetics*, 205(1):385–395, jan 2017.
South African San Institute. San code of research ethics, 2017.
M. Stoneking. *An introduction to molecular anthropology*. John Wiley & Sons, 2016.
C. Stringer. The origin and evolution of Homo sapiens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698), jul 2016.
S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–18, feb 1997.
A. Telenti, L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, and J. C. Venter. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11901–11906, 2016.
J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2): 303–309, feb 2017.
V. Thouzeau, P. Mennecier, P. Verdu, and F. Austerlitz. Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations. *Proceedings of the Royal Society B: Biological Sciences*, 284(1861):20170706, 2017.
S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. The Genetic Structure and History of Africans and African Americans. *Science*, 324(5930):1035–1044, may 2009.
R. Torres, Z. A. Szpiech, and R. D. Hernandez. Human demographic history has amplified the effects of background selection across the genome. *PLoS genetics*,

14(6):e1007387, 2018.

E. Trinkaus, O. Moldovan, A. Bîlgăr, L. Sarcina, S. Athreya, S. E. Bailey, R. Rodrigo, G. Mircea, T. Higham, C. B. Ramsey, et al. An early modern human from the Peştera cu Oase, Romania. *Proceedings of the National Academy of Sciences*, 100(20):11231–11236, 2003.

W. UK10K Consortium, K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J. R. B. Perry, C. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A. E. Hendricks, P. Danecek, R. Li, J. Floyd, L. V. Wain, I. Barroso, S. E. Humphries, M. E. Hurles, E. Zeggini, J. C. Barrett, V. Plagnol, J. B. Richards, C. M. T. Greenwood, N. J. Timpson, R. Durbin, and N. Soranzo. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, oct 2015.

G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, M. A. DePristo, G. A. Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy□Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 43(1):1–33, oct 2013.

K. R. Veeramah, D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins, G. Destro-Bisol, H. Soodyall, L. Louie, and M. F. Hammer. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Molecular Biology and Evolution*, 29(2):617–30, feb 2012.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

P. Verdu. Pourquoi la génétique des populations ne dit-elle rien sur l'autochtonie des populations Pygmées d'Afrique Centrale? In *Quel avenir pour les Pygmées à l'orée du XXIè siècle ?* L'Harmattan, 2019a.

P. Verdu. Do you consent to participate in the research study? *The Secret Lives of Anthropologists: Lessons from the Field*, 2019b.

P. Verdu, F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Théry, A. Froment, S. Le Bomin, A. Gessain, J.-M. Hombert, L. Van der Veen, L. Quintana-Murci, S. Bahuchet, and E. Heyer. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology : CB*, 19(4):312–8, feb 2009.

P. Verdu, N. S. A. Becker, A. Froment, M. Georges, V. Grugni, L. Quintana-Murci, J.-M. Hombert, L. Van der Veen, S. Le Bomin, S. Bahuchet, E. Heyer, and F. Austerlitz. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution*, 30(4):918–37, apr 2013.

M. Vicente and C. M. Schlebusch. African population history: an ancient DNA perspective. *Current opinion in genetics & development*, 62:8–15, 2020.

M. Vicente, M. Jakobsson, P. Ebbesen, and C. M. Schlebusch. Genetic affinities among southern Africa hunter-gatherers and the impact of admixing farming and

pastoralist populations. *Molecular Biology and Evolution*, 36(9):1849–1861, 2019.
P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22, jul 2017.
K. Wang, S. Goldstein, M. Bleasdale, B. Clist, K. Bostoen, P. Bakwa-Lufu, L. T. Buck, A. Crowther, A. Dème, R. J. McIntosh, J. Mercader, C. Ogola, R. C. Power, E. Sawchuk, P. Robertshaw, E. N. Wilmsen, M. Petraglia, E. Ndiema, F. K. Manthi, J. Krause, P. Roberts, N. Boivin, and S. Schiffels. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. *Science Advances*, 6(24):eaaz0183, jun 2020.
J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, apr 1953.
G. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.
B. S. Weir and C. C. Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.
M. C. Weiss, M. Preiner, J. C. Xavier, V. Zimorski, and W. F. Martin. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, 14(8):e1007518, aug 2018.
T. D. White, B. Asfaw, D. DeGusta, H. Gilbert, G. D. Richards, G. Suwa, and F. C. Howell. Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature*, 423 (6941):742–747, jun 2003.
A. Wollstein and W. Stephan. Inferring positive selection in humans from genomic data. *Investigative Genetics*, 6(1):5, apr 2015.
E. T. Wood, D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B. I. Strassmann, H. Soodyall, et al. 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, 2005.
World Medical Association and others. Declaration of Helsinki: ethical principles for medical research involving human subjects. Adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964, and last amended by the 64th WMA General Assembly, Fortaleza, Brasil, October 2013, 2013.
S. Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, mar 1931.
S. Wright. Genetical Structure of Populations. *Nature*, 166(4215):247–249, aug 1950.
L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, and P. M. Visscher. Meta-analysis of genome-wide association studies for height and body mass index in ∼700 000 individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649, 2018.
J. Zhou and Y.-Y. Teo. Estimating time to the most recent common ancestor (TMRCA): comparison and application of eight methods. *European Journal of Human Genetics*, 24(8):1195–1201, 2016.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1949

Editor: The Dean of the Faculty of Science and Technology

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

Distribution: publications.uu.se
urn:nbn:se:uu:diva-416653

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2020