SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE

XXIX Ciclo
(A.A. 2016/2017)

TRANS-SAHARAN CONNECTIONS: HIGH-RESOLUTION ANALYSIS OF HUMAN Y CHROMOSOME DIVERSITY

Dottorando
Eugenia D'Atanasio

Docente guida
Prof. Fulvio Cruciani

Tutore
Prof. Andrea Novelletto

Coordinatore
Prof. Silvia Bonaccorsi

TABLE OF CONTENTS

GLOSSARY

ALT = Alternative

ASD = Average of the squared distance

CNV = Copy number variations

bp = base pair

BWA = Burrows-Wheeler aligner

DHPLC = Denaturing high performance liquid chromatography

DP = Depth

FilDP4 = Filter based on DP4

kya = Kilo years ago

Mb = Mega base pairs

MQ = Mapping quality

MSY = Male-specific region of the Y chromosome

mtDNA = Mitochondrial DNA

NGS = Next generation sequencing

PAR = Pseudoautosomal region

REF = Reference

RFLP = Restriction fragment length polymorphism

SD = Standard deviation

SINE = Short interspersed nuclear element

SNP = Single nucleotide polymorphism

SNS = Single nucleotide substitution

WGA = Whole genome amplification

YAP = Y Alu polymorphism

SUMMARY

Throughout the past millennia, the Sahara underwent strong climatic fluctuations. During the humid phases, the desert became fertile and was called the Green Sahara. During these periods, it was populated by fauna and hominins. The last humid phase occurred between 12 and 5 kya and the human occupation of the Sahara in that period is testified by a bulk of archaeological and paleoanthropological evidence. About 5 kya, an abrupt climatic change put an end to the last African humid period, leading to the desertification of the Sahara. After the onset of these arid conditions, the Sahara became a geographic barrier against the human movement, a fact testified by the strong genetic differentiation between present-day populations from northern and sub-Saharan Africa.

In spite of the large amount of paleoclimatic and archaeological data, little is known regarding the dynamics of the peopling and the depopulation of the Sahara linked to the climatic changes. In this context, today, the rare Y chromosome haplogroups with a trans-Saharan distribution could represent the genetic relic of ancient widespread populations and could provide information about past expansions and migrations across the Green Sahara.

In order to investigate the role of the last Green Sahara in the peopling of Africa, we deep sequenced ~ 3.3 Mb of 104 Y chromosomes belonging to four trans-Saharan haplogroups and identified 5966 mutations, of which 51% were novel. We obtained age estimates for mutation-defined haplogroups using Y chromosome sequences from four ancient specimens as calibration points.

We also analysed the geographic distribution of 108 informative mutations by genotyping 7690 subjects from 141 populations (including 17 populations from literature), mainly from the African continent.

We found that the coalescence age of all the trans-Saharan haplogroups date back to the last African humid period (12-5 kya), while most northern African or sub-Saharan specific sub-haplogroups expanded locally in the subsequent arid phase (< 5 kya). Our findings are consistent with the hypothesis that the Green Sahara represented a corridor for human movements and exclude recent historical events, such as the Arab slave trade, as a major determinant of the gene pool of present-day northern African populations.

INTRODUCTION

The human Y chromosome

Structure of the human Y chromosome

The human Y chromosome is about 58 Mb long and is one of the smallest chromosomes in the human genome (Harris et al. 1986; Morton 1991; Foote et al. 1992; NIH/CEPH Collaborative Mapping Group 1992, International Human Genome Sequencing Consortium 2001; Skaletsky et al. 2003).

Approximately 5% of its length is represented by two PseudoAutosomal Regions (PARs), which are the telomeric portions of the chromosome on the short (PAR1) and on the long (PAR2) arm. These regions include 29 genes and show a high recombination rate with the allelic X chromosome regions during male meiosis (Ross et al. 2005).

The Male-Specific region of the Y chromosome (MSY) accounts for the remaining 95% of the Y chromosome length and is flanked by the PARs. This portion does not recombine during meiosis and follows a male uniparental inheritance pattern.

The MSY consists of both euchromatic and heterochromatic DNA sequences. The latter are subdivided in three blocks. The largest block is approximately 40 Mb long, comprising the distal part of the long arm. A smaller block of approximately 400 kb long in the proximal Yq is composed of 3,000 tandem repeats of 125 bp. The third block is in the centromeric region (a feature of all nuclear chromosomes) and is approximately 1 Mb long (Skaletsky et al. 2003).

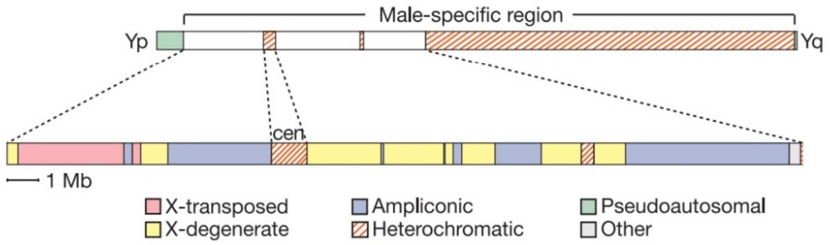


Figure 1: Structure of the MSY. a) Schematic representation of the whole chromosome. b) Enlarged view of the MSY portion, which extends from the proximal boundary of PAR1 to the proximal boundary of the large heterochromatic region of Yq. The three classes of euchromatic sequences are shown, together with heterochromatic sequences (modified from Skaletsky et al. 2003).

The euchromatic portion amounts to about 22.5 Mb, including 8 Mb on the short arm (Yp) and 14.5 Mb on the long arm (Yq), and three discrete sequence classes can be identified within: the X-transposed, the X-degenerate and the ampliconic region (figure 1).

The X-transposed sequences are divided into two blocks on the Y chromosome short arm. Their total length is about 3.4 Mb and they show a 99% similarity to DNA sequences in Xq21. This class is the product of a X-to-Y transposition event which occurred about 3-4 million years ago (Ross et al. 2005), after the divergence of the human and chimpanzee lineages. Subsequently, the transposed region experienced an inversion within the MSY short arm (Skaletsky et al. 2003).

The X-degenerate sequences are the relics of the autosomes from which the sex chromosomes co-evolved (Lahn and Page 1999; Lahn et al. 2001; Skaletsky et al. 2003; Ross et al. 2005; Graves 2006) and display between 60% and 96% nucleotide sequence similarity to their paralogous regions on the X chromosome.

Finally, the ampliconic portion is 10.2 Mb long and is composed of sequences which are very similar, showing up to 99% identity.

Variation in the human Y chromosome

Biallelic polymorphisms

Biallelic polymorphisms are markers that either show their ancestral state or their derived one. The first MSY biallelic polymorphism was discovered about thirty years ago (Casanova et al. 1985). In the second half of the 1990s, the introduction of DHPLC (Denaturing High Performance Liquid Chromatography) and improvements in sequencing techniques made it possible to discover hundreds of new biallelic markers, including Single Nucleotide Polymorphisms (SNPs), deletions, and Alu insertions (Hammer 1994; Seielstad et al. 1994; Whitfield et al. 1995; Jobling et al. 1996; Underhill et al. 1997, 2000, 2001; Shen et al. 2000, 2004; Hammer et al. 2001, 2003; Hammer and Zegura 2002; Cruciani et al. 2002, 2004, 2006, 2007, 2008, 2010, 2011; Y Chromosome Consortium 2002; Kayser et al. 2006; Mohyuddin et al. 2006; Underhill and Kivisild 2007; Karafet et al. 2008; Chiaroni et al. 2009; Scozzari et al. 2012; Mendez et al. 2013). Then, in recent years, the advent of Next Generation Sequencing (NGS) technologies greatly increased the number of newly discovered polymorphisms (Xue et al. 2009; 1000 Genomes Project Consortium 2010; Francalacci et al. 2013; Poznik et al. 2013, 2016; Wei et al. 2013; Scozzari et al. 2014; Hallast et al. 2015; Karmin et al. 2015; Trombetta et al. 2015a).

The MSY shows a lower genetic variation when compared to the autosomes, the X chromosome or mitochondrial DNA (mtDNA) (International SNP Map Working Group 2001). This can be explained by at least two factors:

1. The smaller MSY effective population size (Hammer 1995; Underhill et al. 1996) compared to the rest of the genome; this effect is also enhanced by male mortality (due to wars, hunting, etc.) or cultural phenomena such as polygyny. A consequence of this is that genetic drift has a greater effect on the MSY, causing the faster fixation of alleles and loss of diversity;
2. Lack of recombination, which might have led to the fixation of alleles associated to variants under positive selection, a phenomenon known as hitchhiking (Rice 1987; Whitfield et al. 1995).

The MSY biallelic polymorphisms are considered "stable" in an evolutionary perspective. Their mutation rate, recently estimated at less than 10^{-9} events/position/year (Mendez et al. 2013; Fu et al. 2014; Scozzari et al. 2014; Helgason et al. 2015; Trombetta et al. 2015b) is remarkably lower than the mtDNA mutation rate. Therefore, it is very unlikely that any given site has mutated more than once in recent evolution (Jobling et al. 2013). Thus, if different chromosomes show the same derived state at a site, they probably descend from a common ancestor. Following this reasoning, chromosomes sharing a derived state at one or more sites can be grouped into monophyletic entities, called haplogroups, arranged in a univocal phylogeny (Karafet et al. 2008).

Alu insertions represent a particular kind of biallelic polymorphism. Alu elements are retrotransposons of the SINE (Short Interspersed Element) class, approximately 300 bp long and named after a restriction site for the AluI enzyme within them (Houck et al. 1979). They show an average nucleotide diversity which is higher than the diversity in the rest of the genome, due to the presence of abundant CpG sites (Batzer and Deininger 2002). Their mechanism of transposition passes through an RNA intermediate, but in the genome only few Alu elements, called "master" copies, can retrotranspose. The variant alleles within the "master" elements are passed to all of their copies and this makes it

possible to classify the Alu elements in three families (*AluJ*, *AluS* and *AluY*) and several subfamilies. The expansion of Alu elements occurred in recent times, so some of these sequences are polymorphic for presence/absence in the human genome. In the MSY, the only Alu polymorphism to have been identified to date is the YAP (Y Alu Polymorphism), located in the proximal Yq (Hammer and Horai 1995).

Multiallelic polymorphisms

The human Y chromosome hosts several classes of multiallelic polymorphisms (microsatellites, minisatellites, telomeric repeats, etc.).

Microsatellites are the most utilized in research and are composed of tandem repeats of 1-6 bp stretches. Most of them are polymorphic for the number of repeats (Jobling et al. 2013).

Microsatellites usually mutate through the gain or loss of a single repeat (Weber and Wong 1993; Di Rienzo et al. 1994; Kayser et al. 2000, 2004; Gusmão et al. 2005) following the stepwise model described by Ohta and Kimura (1973). Large variations in the number of repeats are rarely possible (Di Rienzo et al. 1994, Malaspina et al. 1998, 2000).

From a molecular point of view, the "slipped strand mispairing" model (Levinson and Gutman 1987) explains the variation of the number of repeats caused by the slippage of one of the two DNA strands during replication. Their mutation rate (approximately 2×10^{-3}) is orders of magnitude higher than the SNP mutation rate. For this reason, identity by state (equal number of repeats) does not necessarily correspond to identity by descent. As a consequence, the phylogenetic relations reconstructed on the basis of microsatellite variation are unreliable, but these markers can be used in the analysis and dating of recent microevolutionary events (Jobling et al. 2013).

Copy number variations

Copy Number Variations (CNVs) are a class of variants which describe different numbers of copies of sequences that are longer than 1 kb (Feuk et al. 2006). These variants can be both biallelic and multiallelic polymorphisms and are the result of different molecular mechanisms involving segmental duplications (sequences more than 1 kb long with a homology greater than 90%) or short homologous regions (2-15 bp). Some of these mechanisms are physiological responses to single or double strand breaks, which can also lead to chromosomal aberrations and copy number variations.

Due to its haploid nature, the human Y chromosome has a larger number of segmental duplications than the rest of the genome and this has favoured the accumulation of CNVs. CNVs can be placed within the Y chromosome phylogeny, allowing the definition of the ancestral or derived state and, subsequently, the rate of CNV generation (Jobling 2008).

Uniparental markers and phylogenetic trees

Most of our genome shows a biparental transmission, with a whole haploid genome transmitted by each parent, and is subjected to recombination during meiosis. However, MSY and mtDNA represent two exceptions to this rule, because they are inherited from only one parent (male and female, respectively) and they do not experience meiotic crossing-over.

In the MSY, the genetic diversity is due essentially to the sequential accumulation of new mutations (Rozen et al. 2003; Skaletsky et al. 2003), excluding the intrachromosomal or X-Y gene conversion (Rozen et al. 2003; Rosser et al. 2009; Trombetta et al. 2010, 2014; Hallast et al. 2013). The variants are then passed

down along the patrilineages, allowing the recognition of haplogroups. During the world-wide dispersal of *Homo sapiens*, haplogroups underwent molecular differentiation, so each lineage is present in one or few geographic areas. Taking into account the phylogenetic relations and the ethno-geographic distribution of haplogroups, it is possible to reconstruct some human demographic events. This approach is known as phylogeography (Jobling and Tyler-Smith 2003, Chiaroni et al. 2009).

Haplogroups are usually defined by SNPs, so they can be considered as stable and organized in an unambiguous phylogenetic tree. A tree with a good resolution and including all the lineages known at that time was published in 2008 (Karafet et al. 2008). It contains 20 main clades, indicated with letters from A to T (figure 2). Since then, the resolution of the MSY has increased with the discovery of new SNPs. In particular, the structure of the deepest portion of the tree, separating haplogroup A from the rest of the phylogeny (Karafet et al. 2008), was dramatically modified by the discovery of new relations among the basal clades (Cruciani et al. 2011, Scozzari et al. 2012). After these modifications, the former internal A1b lineage became the deepest-rooting branch and haplogroup A was no longer considered a monophyletic lineage. A few years later, an even deeper and rare lineage, called A00, was discovered (Mendez et al. 2013) and, consistently with the nomenclature of A00, A1b was renamed as A0 (figure 3). Apart from this important rearrangement, the tree presented by Karafet et al. (2008) has been only slightly modified, so it can be still considered an adequate representation of the general structure of the Y phylogeny.

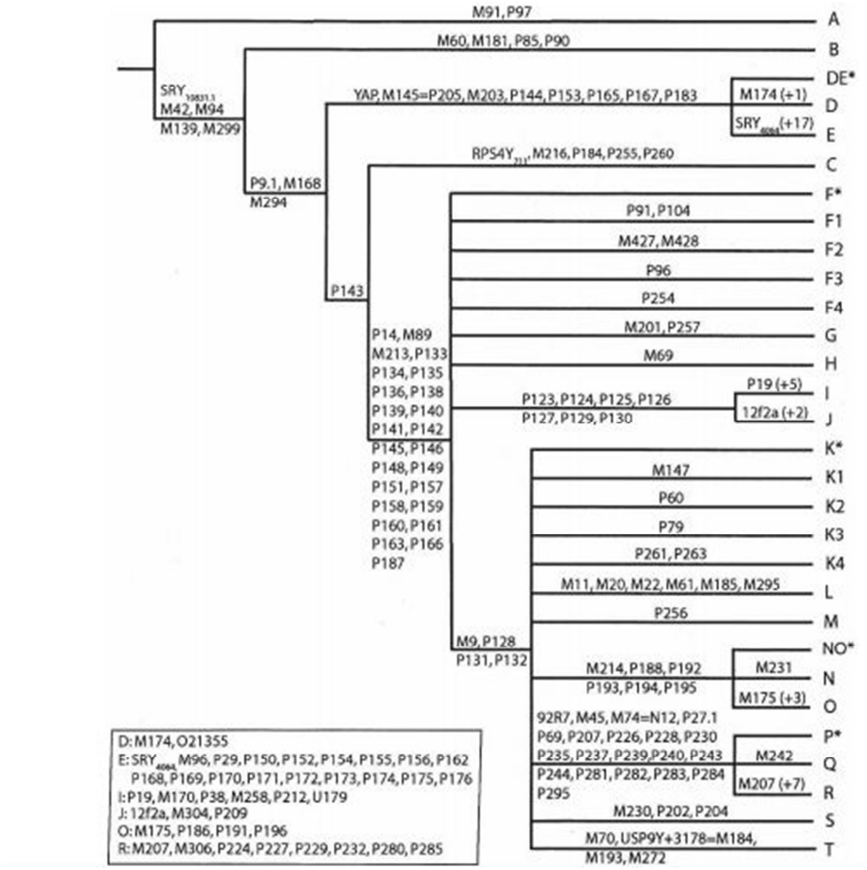


Figure 2: Phylogenetic tree of the human Y chromosome. Haplogroups and defining markers are indicated on the tip and above the branches, respectively (from Karafet et al. 2008).

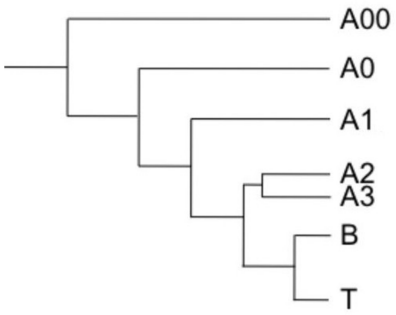


Figure 3. Structure of the Y tree root and relationships of the basal clades on the basis of the discoveries reported in Cruciani et al. (2011) and Mendez et al. (2013) (modified from Mendez et al. 2013).

Geographic distribution of Y chromosome haplogroups

The geographic distribution of the main haplogroups of the MSY phylogenetic tree is represented in figure 4.

The basal clades of the Y phylogeny (A00, A0, A1, A2, A3 and B) are found, with some exceptions, in sub-Saharan Africa with low and intermediate frequencies (Cruciani et al. 2002, 2011; Wood et al. 2005; King et al. 2007a; Chiaroni et al. 2009; Batini et al. 2011; Jobling et al. 2013; Mendez et al. 2013).

The downstream lineages are further grouped within two macro-haplogroups: DE and CF.

Macro-haplogroup DE, defined by the presence of the YAP, shows a world-wide geographic distribution. Haplogroup D is mainly observed in central and south-eastern Asia (Karafet et al. 2001), while haplogroup E is the most common haplogroup in Africa, but is also found at significant frequencies in Europe and the Middle East.

Haplogroup CF is further split into haplogroup C and macro-haplogroup F. The former can be observed at high frequency in New Guinea and Australia (Underhill et al. 2001) and at lower

frequency in southern and eastern Asia (Zhong et al. 2010), while the latter is widely distributed over the world and contains other lineages (G, H, IJ and macro-haplogroup KT).

Haplogroup G is present in the Mediterranean basin and the Caucasian region, haplogroup H is mostly present in India, haplogroup I is characteristic of the European populations, while haplogroup J shows a wider distribution, being present in Europe, northern Africa, the Middle East, India and central Asia. (Hammer et al. 2000; Underhill et al. 2001; Di Giacomo et al. 2004; Rootsi et al. 2004; Sengupta et al. 2006; Battaglia et al. 2009; Chiaroni et al. 2010).

Different clades are grouped within haplogroup K and they are observed mainly in India, Oceania and Indonesia, but some of them display a different geographic pattern, being present in northern Eurasia, central Asia, Africa and the Middle East (Y Chromosome Consortium 2002; Jobling and Tyler-Smith 2003; Sanchez et al. 2005; Kayser et al. 2006; King et al. 2007b; Mona et al. 2007; Rootsi et al. 2007; Karafet et al. 2008). Haplogroup Q is typically observed in the American continent (Karafet et al. 2008), while haplogroup R is one of the most common lineages in the European population (Balaresque et al. 2010; Myres et al. 2010; Underhill et al. 2010).

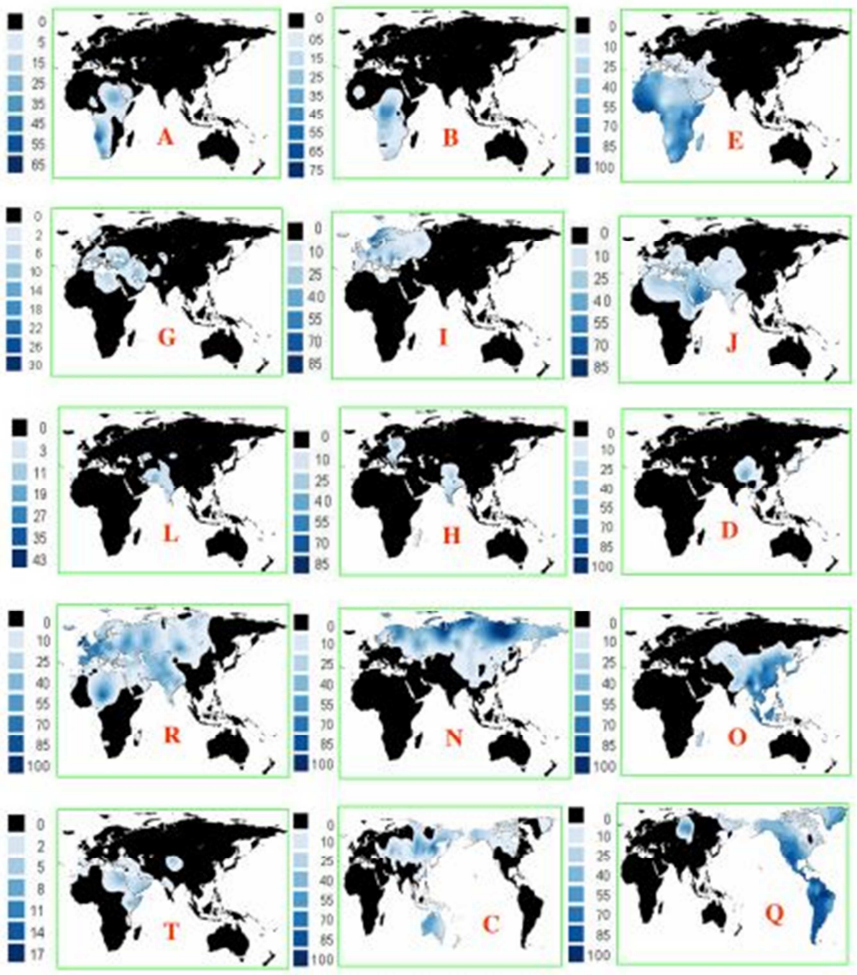


Figure 4. Distribution of the main MSY haplogroups. The geographic frequency distribution maps for the major MSY haplogroups are shown, with the frequency scale indicated on the left. Panel "A" collectively represents the information for haplogroups A00, A0, A1, A2 and A3 (from Chiaroni et al. 2009).

Time estimates for the MSY phylogeny

The splitting times of lineages in a phylogeny provide important information about population dynamics. The nodes of the tree can be dated assuming that the number of mutations in a lineage increase with time. Therefore, accumulated genetic variability represents a measure of the time since the origin of that lineage, assuming that the mutation rate is known.

In order to estimate the Time to the Most Recent Common Ancestor (TMRCA) of two or more lineages, different methods have been developed. These methods are based on the amount of SNPs downstream from the node of interest or on the variability in the number of microsatellite repeats.

The Average of the Squared Distance (ASD) between alleles (Goldstein et al. 1995; Slatkin 1995) is a method used to estimate microsatellite variability and is linearly related to mutation rate (μ) and time (t) (Jobling et al. 2013).

$$ASD = \mu \times t$$

However, problems arise when using microsatellite variability in time estimations linked to the high mutation rate of these markers, which cause a high rate of recurring mutations. In addition, microsatellite variability reaches a plateau. The combination of these two factors leads to an underestimation of the microsatellite diversity with a subsequent underestimation of the TMRCA, in particular for ancient events (Busby et al. 2012).

The mutation rate of SNSs (Single Nucleotide Substitutions) is significantly lower (Kong et al. 2012; Mendez et al. 2013), so the occurrence of recurring mutations is rare, making these markers more suitable for dating purposes. The Rho (ρ) statistic measures the average number of nucleotide differences (Forster et al. 1996; Saillard et al. 2000) and is proportionally related to mutation rate (μ) and time (t):

$$\rho = \mu \times t$$

However, the use of SNSs for time estimates within the Y phylogeny is complicated by their low amount of variability that is more difficult to discover when compared to the microsatellite variability.

The use of NGS has hugely increased the number of newly identified variants in recent years (Xue et al. 2009; 1000 Genomes Project Consortium 2010; Francalacci et al. 2013; Poznik et al. 2013, 2016; Wei et al. 2013; Scozzari et al. 2014; Batini et al. 2015; Hallast et al. 2015; Karmin et al. 2015; Trombetta et al 2015a, 2015b), overcoming all the difficulties linked to the SNSs and making it possible to use these markers to estimate node ages.

Peopling of Africa

According to archaeological, paleoanthropological and genetic data, *Homo sapiens* originated in Africa less than 300 kya (Kilo Years Ago). Later, our species left the African continent and colonised the rest of the world, a scenario called "Out of Africa" (Jobling et al. 2013).

Distribution of Y chromosome haplogroups in Africa

Within the African continent, different lineages have different geographic distributions.

The basal clades A00 and A0 have been observed at very low frequencies only in small populations in central Africa (Cruciani et al. 2011; Mendez et al. 2013), in contrast with the most ancient human fossil remains that were found in eastern Africa.

A1 has been found at low frequencies in an area ranging from Morocco to Senegal and Niger (Cruciani et al. 2002; Gonçalves et al. 2003; Vallone and Butler 2004; Wood et al. 2005; King et al. 2007a; Rosa et al. 2007).

A2, defined by the M6 marker, is present in Khoisan populations from southern Africa (Underhill et al. 2000; Cruciani et al. 2002; Batini et al. 2011).

On the contrary, its sister clade, A3, is present in different areas of the African continent, with its internal lineages showing a clear geographic differentiation. The rare A3-M28 is only found in eastern Africa, while its sister clade is more frequent and is further subdivided into two main sub-haplogroups, A3-M51 and A3-M13. They are found in southern Africa and central-eastern Africa, respectively. The A3 distribution suggests that eastern Africa was its possible place of origin, with a recent expansion to the central and southern part of the continent (Underhill et al. 2000; Cruciani et al. 2002; Semino et al. 2002; Batini et al. 2011).

The sub-Saharan haplogroup B, defined by the M60 marker, is subdivided into two main branches, B1-M236 and B2-M182. B1-M236 has been observed at very low frequencies in the western-central part of the continent (Underhill et al. 2000; Cruciani et al. 2002), while B2-M182 has a wider distribution, being present in eastern, central and southern Africa, with some of its sub-haplogroups reaching frequencies of up to 70% in some populations (Underhill et al. 2000; Cruciani et al. 2002; Wood et al. 2005; Gomes et al. 2010; Batini et al. 2011).

The most common haplogroup in Africa is haplogroup E, which is also frequent in Europe and the Middle East. The distribution of this clade and its internal lineages within Africa is not homogeneous.

Its E-M33 branch is found in western Africa (Wood et al. 2005), while E-P2 sub-haplogroup is present in the entire continent. This clade is further subdivided into two lineages, E-V38 and E-M215.

Each of them include a rare eastern African lineage (E-M329 and E-V16, respectively) and a more common branch (E-M2 and E-M35, respectively), which is further subdivided into different lineages. E-M2 and E-M35 are found in different geographic areas: the first is present mainly in sub-Saharan Africa, while the second shows a wider geographic distribution, being present at significant frequencies outside Africa in the Mediterranean area (Scozzari et al. 1999; Underhill et al. 2000; Cruciani et al. 2002, 2004, 2007; Semino et al. 2002, 2004; Knight et al. 2003; Luis et al. 2004; Beleza et al. 2005; Wood et al. 2005; Rosa et al. 2007; Henn et al. 2008; Battaglia et al. 2009; Gomes et al. 2010; Trombetta et al. 2015a).

Haplogroup E-M75 is subdivided into an eastern African clade (E-M41) and a more widespread lineage (E-M54) (Wood et al. 2005).

Within the macro-haplogroup K, the G and J lineages are quite frequent in the northern part of the continent (Cruciani et al. 2002) and their presence in this area has been put down to a Neolithic migration from the Middle East (Arredi et al. 2004).

The remaining lineages are very rare in Africa, with the exceptions of haplogroup T and a sub-haplogroup within R (R-V88). The former reaches a frequency of 18% among the Fulbe from Cameroon and a 11% frequency among the Oromo from Kenya (Cruciani et al. 2002; Wood et al. 2005). R-V88 is quite frequent in west-central Africa, where its presence has been explained as a migration back to Africa from Asia (Cruciani et al. 2002, 2010).

The role of the Sahara in the peopling of Africa

The Sahara desert is the widest hot desert on Earth and covers about one third of the African continent, from the Atlantic coast to the Red Sea. Over the past millennia, Sahara underwent strong climatic fluctuations, alternating arid and humid phases. During the humid periods, referred as "Green Sahara" or "African humid periods", the landscape was characterised by the presence of savannah, forests and an extensive system of rivers and lakes (Brooks et al. 2005; Drake et al. 2011; Larrasoaña et al. 2013; Skonieczny et al. 2015). A large amount of paleoecological and paleoanthropological evidence indicates that the fertile environment probably enabled the occupation of the Saharan area by fauna and hominins since the Miocene (about 23-5 Ma) (Pachur and Hoelzmann 2000; Smith et al. 2004, 2007; Osborne et al. 2008; Lahr 2010; Lebatard et al. 2010; Carrión et al. 2011; Blome et al. 2012; Larrasoaña et al. 2013; Scerri et al. 2014).

The most recent Green Sahara period occurred in the Holocene, in a time frame from about 12 kya to about 5 kya. This phase has been denominated "Holocene climatic optimum" and is the most well-documented past climatic change (Gasse et al. 1990; deMenocal et al. 2000; Adkins et al. 2006; Drake et al. 2011; McGee et al. 2013; Tierney and deMenocal 2013). The human settlement across the Sahara in this period is testified by archaeological evidence, such as rock engravings, lithic and bone tools and pottery (Sutton 1977; Yellen et al. 1998; Sereno et al. 2008).

After the African humid period, the climatic conditions became rapidly hyper-arid and the Green Sahara was replaced by the desert, which acted as a strong geographic barrier against human movements between northern and sub-Saharan Africa.

A consequence of this is that there is a strong differentiation in the Y haplogroup composition between the northern and sub-Saharan regions of the African continent. In the northern area, the predominant Y lineages are J and E-M81, with the latter reaching a

frequency as high as 80% in some populations (Arredi et al. 2004; Semino et al. 2004). The massive presence of these lineages in northern Africa has been explained as a consequence of a Neolithic migration from Middle East (Arredi et al. 2004). On the contrary, sub-Saharan Africa is characterised by a completely different genetic landscape, with lineages within E-M2 and haplogroup B comprising the most of the Y chromosomes. In most regions of sub-Saharan Africa, the observed haplogroup distribution has been linked to the Bantu expansion. According to this hypothesis, about 3 kya, the demic diffusion of Bantu agriculturalists brought E-M191, E-U209 (two E-M2 sub-haplogroups) and B-M150 from central Africa to the East and to the South, while E-M191 also spread westward (Cruciani et al. 2002; Beleza et al. 2005; Wood et al. 2005; Berniell-Lee et al. 2009; Gomes et al. 2010; Batini et al. 2011; de Filippo et al. 2011; Montano et al. 2011; Ansari Pour et al. 2013; Scozzari et al. 2014; Poznik et al. 2016).

However, in spite of their Y haplogroup differentiation, northern and sub-Saharan Africa share at least four lineages at different frequencies, namely A3-M13, E-M2, E-M78 and R-V88.

A3-M13 is typical of eastern Africa, where it is found with a frequency as high as 40% (Gomes et al. 2010) and is prevalent in the Nilo-Saharan populations, in particular among Nilotic pastoralists (Hassan et al. 2008; Gomes et al. 2010; Batini et al. 2011). A3-M13 chromosomes have also been found in central and northern Africa, with frequencies ranging from 1% to 7% (Cruciani et al. 2002; Luis et al. 2004; Wood et al. 2005; Batini et al. 2011). Outside Africa, this haplogroup has been found at very low frequency both in the Middle East (Nebel et al. 2001; Cinnioğlu et al. 2004; Luis et al. 2004; Shen et al. 2004; Flores et al. 2005) and Sardinia, where its presence has been explained as a consequence of historical events, such as the Roman or Vandalic dominations (Semino et al. 2000; Underhill et al. 2000; Francalacci et al. 2013, 2015).

As described above, E-M2 is a clade which has been often associated with the Bantu expansion in the sub-Saharan area. However, E-M2 chromosomes have also been found at low frequency (2-10%) in northern Africa (Cruciani et al. 2002; Arredi et al. 2004; Semino et al. 2004; Robino et al 2008; Fadhlaoui-Zid et al. 2013; Triki-Fendri et al. 2015) where their presence has been hypothesized to be a consequence of recent movements (Luis et al. 2004).

E-M78 is a widespread lineage, with significant frequencies in Africa, Europe and the Middle East (Cruciani et al. 2004, 2007). Within the African continent, the E-M78 sub-clades show different frequencies in different regions. E-V22 is mainly an eastern African sub-haplogroup, with frequencies of more than 80% in the Saho population from Eritrea, but it has also been reported in Egypt and Morocco (Cruciani et al. 2007; Trombetta et al. 2015a). E-V12 is relatively frequent in northern and eastern Africa, but it has also been reported outside Africa at lower frequencies (Cruciani et al. 2004, 2007; Trombetta et al. 2015a). The vast majority of the eastern African E-V12 chromosomes belong to the internal clade E-V32, which has also been observed in northern and central Africa at very low frequencies (Wood et al. 2005; Cruciani et al. 2004, 2007; Trombetta et al. 2015a). The sister clade of E-V12 is defined by the V264 marker and is further subdivided into two sub-clades, E-V65 and E-V259. The former is very frequent in northern Africa, while the latter includes few central African chromosomes (Cruciani et al. 2004, 2007; Trombetta et al. 2015a). The overall E-M78 geographic distribution in Africa has been explained as the result of migrations along the Nile valley between eastern and north-eastern Africa (Cruciani et al. 2007).

R-V88 can be observed at high frequencies in the central Sahel (northern Cameroon, northern Nigeria, Chad and Niger) and it has also been reported at low frequencies in north-western Africa (Cruciani et al. 2010). Outside the African continent, two specific and rare R-V88 sub-lineages, namely R-M18 and R-V35, have

been observed. R-M18 has been reported at very low frequencies in Lebanon, Corsica and Sardinia (Zalloua et al. 2008; Cruciani et al. 2010; Morelli et al. 2010; Francalacci et al. 2013, 2015), while R-V35 is restricted to Sardinia (Cruciani et al. 2010; Francalacci et al. 2013, 2015). Within the African continent, R-V88 is very frequent in the central African populations that speak the Chadic language, a branch of the Afroasiatic language family (Cruciani et al. 2010). Two different hypotheses have been proposed to explain the arrival of Chadic language in central Africa from north-eastern Africa, which is the probable place of origin of Proto-Afroasiatic language. According to the "inter-Saharan" hypothesis, the Chadic language followed an east-to-west course along the Sahel belt (Blench 2006). On the contrary, according to the "trans-Saharan" hypothesis, the Proto-Chadic language arose in northern Africa and arrived in central Africa crossing the Sahara (Ehret 1995). Because of its ethno-geographic distribution, R-V88 has been linked to the Chadic spread, but to date both hypotheses are considered equally likely and another trans-Saharan migration route has also been proposed, in the opposite direction compared to the previous one (from central to northern Africa) (Cruciani et al. 2010; González et al. 2013).

AIMS

The last Green Sahara period, which took place during the Holocene climatic optimum, is one of the most studied past climatic events. Its effects on human settlements are testified by considerable archaeological and paleoanthropological evidence scattered throughout the Sahara. Nonetheless, the extent, the routes and the time frames of human movements are still largely unknown, because of the difficulties linked to the presence of the Sahara desert.

From a genetic point of view, the use of the present day MSY variability to infer past population dynamics across the Sahara is complicated by two major factors:

1. the onset of the hyper-arid conditions caused the depopulation of the Sahara;
2. the regions immediately northward and southward of the Sahara have experienced extensive demographic expansions after the African humid period, which have led to the increase in frequency of different Y haplogroups, replacing the pre-existing genetic composition.

In this context, rare Y lineages with a relic geographic distribution can be highly informative regarding human migrations across the Sahara. Thus, considering their frequency distribution, the four trans-Saharan lineages A3-M13, E-M2, E-M78 and R-V88 could represent the remains of the Saharan MSY genetic landscape before the desertification, contrary to the usual interpretation involving recent gene flow events such as the trans-Saharan Arab slave trade.

In order to understand the effects of the last Green Sahara period (12-5 kya) on human movements, we used high-coverage (> 50×) NGS to study 104 Y chromosomes, 77 of which belonging to the four trans-Saharan haplogroups and the remaining 27

representing the main Y tree lineages. We included our sample in a wider phylogenetic context adding 28 high-coverage Y sequences from Complete Genomics (Drmanac et al. 2010) and from a recent study concerning world-wide Y chromosome variation (Karmin et al. 2015). Finally, we added as calibration points the Y chromosomes from four radiocarbon-dated ancient specimens (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015) in order to estimate an accurate Y mutation rate to be used for dating the informative splits in the phylogeny.

We reconstructed the phylogenetic relations among the whole set of 150 Y chromosomes and we selected a set of 108 informative markers among the newly identified SNPs. These polymorphisms were analysed in more than 5000 African and non-African males from our lab collection and from two recent studies (Francalacci et al. 2015; Poznik et al. 2016) in order to gain additional information about the ethno-geographic distribution of the trans-Saharan lineages.

In this thesis, we will discuss our results, focusing in particular on the implications of the time estimates, phylogenetic relationships and geographic distribution of the trans-Saharan lineages regarding the climatic changes that took place in Africa during the Holocene.

RESULTS

We deep-sequenced about 3.3 Mb of the X-degenerate portion of the MSY in 104 Y chromosomes from our lab collection, belonging to informative lineages of the Y phylogenetic tree. To our set, we added 46 publicly available high-coverage Y chromosome sequences. We used the identified SNPs to reconstruct the relations among the samples and to estimate the coalescence age of the nodes within the phylogeny. An informative subset of SNPs was also selected to be further analysed in a wider sample.

Targeted next generation sequencing

Sample selection

For the phylogenetic tree reconstruction and time estimates, we used 150 Y chromosomes, belonging to different datasets.

Samples from our lab collection

First of all, from our lab collection, we selected 104 Y chromosomes to be analysed by next generation sequencing, on the basis of their Y haplogroup affiliation (Cruciani et al. 2002, 2004, 2007, 2010, 2011; Scozzari et al. 2012, 2014; Trombetta et al. 2011, 2015a). We mainly focused our analysis on the four trans-Saharan clades (A3-M13, E-M2, E-M78 and R-V88), taking into account the samples' place of origin. We also included a set of chromosomes from other important phylogenetic lineages in order to obtain more accurate estimates of tree node age (table 1).

ID	Former ID	Haplogroup	Region	Country	Population
S101		A00-L1086	Western/Central Africa	Cameroon	General population
S102		A0-V148	Western/Central Africa	Cameroon	General population
S103		A0-V148	Northern Africa	Algeria	Mozabite Berbers
S104	S07	A1-M31	Western/Central Africa	Mali	General population
S105	S08	A2-PN3	Southern Africa	Angola	!Kung
S106	S75	A2-PN3	Western/Central Africa	Cameroon	General population
S107	S73	A3-M28	Eastern Africa	Eritrea	Nara
S108		A3-M51	Southern Africa	Angola	!Kung
S109		A3-M51	Southern Africa	Angola	!Kung
S110	S10	A3-M13*	Europe	Italy	Sardinians
S111		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S112		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S113		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S114		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S115		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S116		A3-M13/V3	Eastern Africa	Ethiopia	Ethiopian Jews
S118		A3-M13/V317	Eastern Africa	Ethiopia	Amhara
S119		A3-M13/V317	Eastern Africa	Ethiopia	Oromo
S120		A3-M13*	Eastern Africa	Kenya	Maasai
S121		A3-M13*	Eastern Africa	Kenya	Maasai
S122		A3-M13*	Western/Central Africa	Cameroon	Fulbe
S123		A3-M13/V67	Western/Central Africa	Nigeria	Hausa
S124		A3-M13*	Western/Central Africa	Chad	Shuwa Arabs
S125		A3-M13*	Western/Central Africa	Chad	Ngambai
S126	S77	A3-M13*	Western/Central Africa	Chad	Massa
S127		A3-M13*	Western/Central Africa	Benin	General population
S128		A3-M13*	Northern Africa	Morocco	Souss Berbers
S129	S76	A3-M13*	Northern Africa	Egypt	Northern Egyptians
S130		B-M60	Western/Central Africa	Chad	Gor
S131	S14	B-M236/M146	Western/Central Africa	Burkina Faso	General population
S132		B-M150/M109	Western/Central Africa	Cameroon	Moundang
S133		B-M150/M109	Western/Central Africa	Chad	Toupouri
S134	S74	B-M112	Western/Central Africa	Cameroon	General population
S135		E-V44	Eastern Africa	Ethiopia	Amhara
S136		E-V257* (xM81)	Western/Central Africa	Chad	Massa
S137		E-M78/V259	Western/Central Africa	Cameroon	Daba
S138		E-M78/V259	Western/Central Africa	Cameroon	Guidar
S139		E-M78/V12*	Western/Central Africa	Cameroon	Mandara
S140		E-M78/V32	Western/Central Africa	Cameroon	Moundang
S141		E-M78/V32	Western/Central Africa	Chad	Massa
S142		E-M78/V32	Western/Central Africa	Chad	Goulaye
S143		E-M78/V32	Western/Central Africa	Chad	Goulaye
S144		E-M78/V32	Western/Central Africa	Chad	Madjingay
S145		E-M78/V32	Western/Central Africa	Chad	Goulaye
S146		E-M78/V12*	Northern Africa	Egypt	General population
S147		E-M78/V12*	Northern Africa	Egypt	General population
S148		E-M78/V12*	Northern Africa	Morocco	Moroccan Jews
S149	S25	E-M78/V65	Northern Africa	Libya	Libyan Jews

ID	Former ID	Haplogroup	Region	Country	Population
S150		E-M78/V65	Northern Africa	Egypt	Egyptian Berbers from Siwa
S151		E-M78/V65	Northern Africa	Egypt	Egyptian Berbers from Siwa
S152		E-M78/V65	Northern Africa	Morocco	Moroccan Arabs
S153		E-M78/V65	Northern Africa	Morocco	Moroccan Arabs
S154		E-M78/V32	Eastern Africa	Ethiopia	Amhara
S155		E-M78/V32	Eastern Africa	Eritrea/Ethiopia	Tigrai
S156		E-M78/V32	Eastern Africa	Ethiopia	Ethiopian Jews
S157		E-M78/V32	Eastern Africa	Kenya	Borana
S158		E-M78/V32	Eastern Africa	Kenya	Luhya
S159		E-M78/V32	Eastern Africa	Kenya	Maasai
S160		E-V68/V2009	Western/Central Africa	Cameroon	Fulbe
S161		E-V68/V2009	Northern Africa	Morocco	Souss Berbers
S162		E-M78/V22	Eastern Africa	Eritrea	Saho
S163		E-M2*	Northern Africa	Egypt	Egyptians from Baharia
S164		E-M2*	Northern Africa	Egypt	Egyptians from Baharia
S165		E-M2*	Northern Africa	Egypt	Egyptian Berbers from Siwa
S166		E-M2*	Northern Africa	Egypt	Egyptian Berbers from Siwa
S167		E-M2*	Northern Africa	Morocco	Ouarzazate Berbers
S168		E-M2*	Northern Africa	Morocco	Ouarzazate Berbers
S169		E-M2*	Northern Africa	Morocco	Asni Berbers
S170		E-M2*	Northern Africa	Morocco	Bouhria Berbers
S171		E-M2*	Western/Central Africa	Burkina Faso	General population
S173		E-M2*	Western/Central Africa	Senegal	Mandenka
S175		E-M2*	Western/Central Africa	Cameroon	Ngambai
S176		E-M2*	Western/Central Africa	Cameroon	Fali
S177		E-M2*	Western/Central Africa	Niger	Songhai
S178	S38	C-V20	Europe	Italy	General population
S179		C-M8	Asia	Japan	General population
S181		J-M172	Near East	Turkey	Sephardic Turkish
S182		J-M267/P58	Western/Central Africa	Chad	Gor
S183		R-V88*	Western/Central Africa	Benin	General population
S184		R-V88*	Western/Central Africa	Benin	General population
S185		R-V88*	Western/Central Africa	Cameroon	Ouldeme
S186		R-V88/V69	Western/Central Africa	Cameroon	Moundang
S187		R-V88/V69	Western/Central Africa	Cameroon	Fulbe
S188		R-V88/V69	Western/Central Africa	Cameroon	Toupouri
S189		R-V88/V69	Western/Central Africa	Chad	Toupouri
S190		R-V88*	Western/Central Africa	Chad	Madjingay
S191		R-V88/V69	Western/Central Africa	Chad	Toupouri
S192		R-V88*	Western/Central Africa	Chad	Toupouri
S193		R-V88*	Western/Central Africa	Chad	Moundang
S194		R-V88*	Western/Central Africa	Chad	Massa
S195		R-V88*	Western/Central Africa	Chad	Gor
S196		R-V88*	Western/Central Africa	Cameroon	Ewondo
S197		R-V88*	Europe	Bulgaria	Sephardic Bulgarians
S198		R-V88*	Northern Africa	Algeria	Mozabite Berbers
S200		R-V88*	Northern Africa	Morocco	Ouarzazate Berbers
S201		R-V88*	Northern Africa	Egypt	General population

ID	Former ID	Haplogroup	Region	Country	Population
S202		R-V88*	Northern Africa	Egypt	General population
S203	TV18	R-V88*	Northern Africa	Egypt	General population
S204		R-V88*	Northern Africa	Egypt	Northern Egyptians
S206		J-M267/P58	Eastern Africa	Ethiopia	Amhara
S207		J-M267* (xP58)	Eastern Africa	Ethiopia	Oromo
S208		J-M267/P58	Northern Africa	Algeria	Mozabite Berbers
S209		J-M267* (xP58)	Near East	Yemen	Yemenites
S210		J-M267* (xP58)	Eastern Africa	Ethiopia	Amhara

Table 1: Samples from our lab collection analysed by next generation sequencing. Haplogroup affiliation is expressed according to the nomenclature "by marker", except for samples belonging to the basal clades (A00, A0, A1, A2'3), which are defined by a mixed nomenclature ("by lineage" and "by marker"). For the subjects that had already been analysed by NGS in previous studies (Scozzari et al. 2014; Trombetta et al. 2015a, 2015b), the former ID is reported. The DNA samples are from blood, saliva or cell lines.

Publicly available whole Y chromosomes

In order to increase the power of resolution of the present study, we included a set of publicly available Y high-coverage sequences (20 - 40×), selected according to their haplogroup affiliation. More specifically, we added 28 samples from the Complete Genomics diversity set (Drmanac et al. 2010) and 14 subjects among those published by Karmin et al. (2015) (table 2).

ID	Haplogroup	Region/population	Reference
GS16204	A3-M13/V243	Ethiopian Jews	Karmin et al. 2015
GS35245	E-M2/U174	Congo-pygmies	Karmin et al. 2015
GS16179	E-V32	Iranians	Karmin et al. 2015
GS16217	E-M4145	Arab-Christian	Karmin et al. 2015
GS16206	E-M34	Arab-Christian	Karmin et al. 2015
GS13741	J1c-PF7256	Azeri	Karmin et al. 2015
GS14421	J1a1-B232	Tabas-saran	Karmin et al. 2015
GS13724	J1a2-B234	Lezgin	Karmin et al. 2015
GS35126	J1b6-PR6622	Armenian	Karmin et al. 2015
GS35125	J1b5-CTS11284	Armenian	Karmin et al. 2015
GS14474	J1b4-B235	Jordanian	Karmin et al. 2015
GS16136	J1b2-L829	Druze	Karmin et al. 2015
GS16180	J1b1-B243	Arabian	Karmin et al. 2015
GS35124	J1b3-B382	Armenian	Karmin et al. 2015
NA18940	D2-M55	Eastern Asia, Japan, JPT	Complete Genomics
NA19239	E1a-P110	Western Africa, Nigeria, YRI	Complete Genomics
NA18504	E1b1a-U174	Western Africa, Nigeria, YRI	Complete Genomics
NA19026	E1b1a-U174	Eastern Africa, Kenya, LWK	Complete Genomics
NA19834	E1b1a-U174	Northern America, USA, ASW	Complete Genomics
NA19703	E1b1a-U181	Northern America, USA, ASW	Complete Genomics
NA18501	E1b1a-U209* (xU290)	Western Africa, Nigeria, YRI	Complete Genomics
NA19020	E1b1a-U209* (xU290)	Eastern Africa, Kenya, LWK	Complete Genomics
NA19025	E1b1a-U209* (xU290)	Eastern Africa, Kenya, LWK	Complete Genomics
NA19700	E1b1a-U290* (xU181)	Northern America, USA, ASW	Complete Genomics
NA20510	E1b1b-V13	Europe, Italy, TSI	Complete Genomics
NA21732	E1b1b-V22	Eastern Africa, Kenya, MKK	Complete Genomics
NA21737	E1b1b-V22	Eastern Africa, Kenya, MKK	Complete Genomics
NA19670	G-U8	Northern America, California, MXL	Complete Genomics
NA06994	I1-M253	Northern America, Utah, CEU	Complete Genomics
NA12891	I1-M253	Northern America, Utah, CEPH/Utah Pedigree 1463	Complete Genomics
NA20511	I1-M253	Europe, Italy, TSI	Complete Genomics
NA18558	N-M231	Eastern Asia, China, CHB	Complete Genomics
NA19735	Q1a-M3	Northern America, California, MXL	Complete Genomics
NA20846	R1a-M17	Northern America, Texas, GIH	Complete Genomics
NA20850	R1a-M17	Northern America, Texas, GIH	Complete Genomics
NA10851	R1b1-M529* (xM222)	Northern America, Utah, CEU	Complete Genomics
NA12889	R1b1-P312* (xDF27,U152,M529,L238)	Northern America, Utah, CEPH/Utah Pedigree 1463	Complete Genomics
HG00731	R1b1-U152	Central America, Puerto Rico, PUR	Complete Genomics
NA07357	R1b1-U152	Northern America, Utah, CEU	Complete Genomics
NA19649	R1b1-U152	Northern America, California, MXL	Complete Genomics
NA20509	R1b1-U152	Europe, Italy, TSI	Complete Genomics
NA20845	R2-M124	Northern America, Texas, GIH	Complete Genomics

Table 2: Selected publicly available samples. The information regarding the samples is reported according to the original study, except for the haplogroup affiliation of the Complete Genomics subjects, which is the same as in Trombetta et al. (2015b).

Ancient specimens

We needed at least one calibration point within the phylogeny to calculate accurate time estimates, so we added four precisely dated ancient specimens (Fu et al. 2014; Lazaridis et al. 2014; Jones et al.

2015). These samples were selected on the basis of the average sequence coverage of the Y chromosome regions analysed in this study (table 3).

ID	Radiocarbon dating	Average Y coverage	Haplogroup	Region	Reference
Ust'-Ishim	~ 45.0 kya	22×	NO[a]	Siberia	Fu et al. 2014
Bichon	~ 13.7 kya	4.7×	I2a[b]	Switzerland	Jones et al. 2015
Kotias	~ 9.7 kya	7.7×	J2[b]	Georgia	Jones et al. 2015
Loschbour	~ 8.0 kya	11×	I2[a]	Luxembourg	Lazaridis et al. 2014

Table 3: Radiocarbon dated ancient specimens included in this study. The subjects are listed from the most ancient to the most recent. a. Haplogroup affiliation as deduced from the phylogeny in Trombetta et al. (2015b). b. Haplogroup affiliation according to the original study.

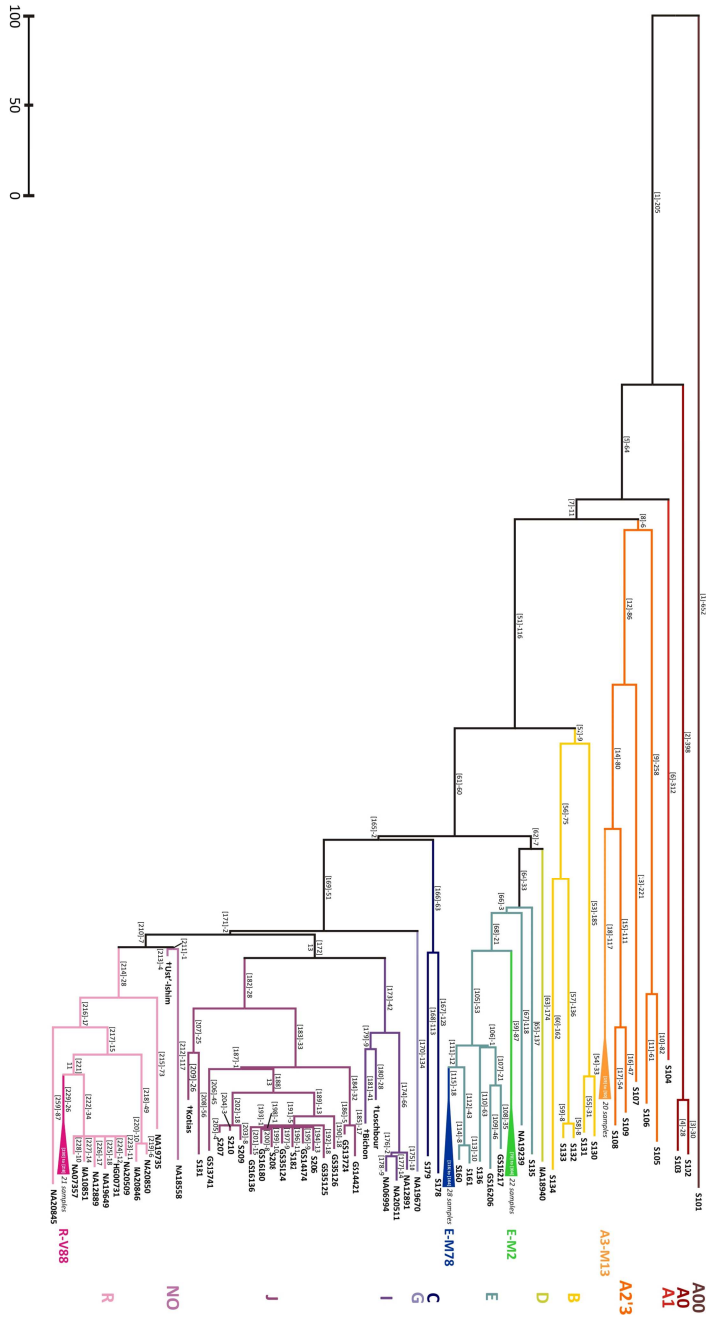
Region selection

We performed a targeted next-generation sequencing of about 4.4 Mb of the X-degenerated portion of the human Y chromosome on 104 samples selected from our lab collection (table 1). In order to guarantee reliable data, to the initial target region, we applied a filter based on the average sequence depth in each subject (see Materials and Methods), obtaining a final set of 3.3 Mb. We extracted the same set of bases from the publicly available whole Y chromosomes from Complete Genomics (Drmanac et al. 2010) and Karmin et al. (2015) (table 2). Then, we checked the average coverage for the final set of bases in the ancient specimens (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015) to ensure that these samples could be analysed exactly for the same regions as the other subjects (see Materials and Methods).

Phylogenetic tree and time estimates

General features of the phylogeny

We compared the sequences of the 104 subjects from our lab collection (table 1) with the reference sequence of the human Y chromosome (Feb. 2009 - GRCh37/hg19 assembly) and we identified 5966 nucleotide positions which showed a different base compared to the reference in at least one of our sequences. Then, we also added the variants present in the 46 publicly available samples (tables 2-3), reaching a total of 7544 differences from the reference (Appendix 1). We used all these variants to reconstruct a maximum parsimony tree (figure 5), which was found to be coherent with the recently published world-wide Y phylogenies (Karmin et al. 2015; Poznik et al. 2016).


100
50
0
A00
A0
A1
A2'3
A3-M13
B
D
E
E-M2
E-M78
C
G
I
J
NO
R
R-V88

(Previous page) Figure 5: Maximum parsimony tree reconstructed with 7544 variable positions. Different haplogroups are represented in different colours. Above or near each branch, the name (in squared brackets) and the number of mutations are indicated. At the tip of the terminal branches, the ID of the corresponding subject is reported. The four trans-Saharan clades are contained in the form of triangles, with a length proportional to the average number of SNPs in the lineages included. The names of the branches and the number of samples belonging to the collapsed clades are reported within and near each triangle, respectively. The ancient specimens are labelled with the symbol "†" before their names.

Considering the phylogenetic relations among the subjects, we were able to recognise the ancestral and the derived state for most of the variant positions and, consequently, it was possible to assign them univocally to one branch (or more branches, in cases of recurrent variants).

However, in some cases, we could not decide the direction of the mutational event. 857 positions (including one triallelic and three recurrent mutations) were different from the reference in all the samples except for the A00 subject. Since the reference sequence is composed of a large portion belonging to haplogroup R and a small portion belonging to haplogroup G, the difference from the reference of the A00 lineage can put down to either an A00-specific mutational event or a mutation which occurred at the root of A0-T. For this reason, we defined these two branches together as "branch 1" and the resulting tree is unrooted.

Moreover, we identified two recurrent mutations which could be interpreted in two different ways (table 4). One of them could be assigned to branches 1 and 6 looking at the corresponding positions on the chimpanzee Y chromosome, assuming that the chimpanzee maintained the ancestral state. The other mutation could not be assigned considering the chimpanzee sequence (red in Appendix 1). We decided to interpret it as a mutation in branch 12 and a reversion in branch 18, considering two different possible explanations for this pattern:

1) the mutation "T" to "C" on branch 12 formed a CpG site; CpG sites show a higher mutation rate compared to the rest of the genome (Ségurel et al. 2014; Makova and

Hardison 2015) so our position could have experienced a revertion in branch 18;

2) the variant position fell within a Y portion showing about 90% of similarity with the X chromosome, which had a "T" in the corresponding position; as a consequence, the reversion on branch 18 could also be explained as the result of a X-Y gene conversion event.

Position (hg19)	Reference base	Alternative base	Interpretation 1	Interpretation 2	chimp base (panTro4)	Final assignment
7582257	T	C	double hit in branches 1 and 6	double hit in branches 2 and 7	T	Interpretation 1
14044307	T	C	mutation in branch 12 and reversion in branch 18	double hit in branches 13 and 15	T	Interpretation 1

Table 4. Two recurring mutations interpretable in two different ways based on their phylogeny. The unresolved position is labelled in grey (see the text for more details).

We also identified 11 triallelic variants and 23 recurring mutations (respectively, green and yellow in Appendix 1). Among these, one was found three times in our phylogeny, four could be recognised as reversions and two could be interpreted either as double-hit mutations or reversions because one of the branches involved was branch 1 and so the SNPs could not be univocally assigned. Finally, we identified 21 polymorphisms which were different from the reference sequence but invariant in all the samples. These variants were interpreted as reference-specific mutations (text in underlined italics in Appendix 1).

Comparison with literature

We compared the SNPs identified in the present work with the markers reported in other public datasets of human Y chromosome variation (Hallast et al. 2015; Batini et al. 2015; Francalacci et al. 2015; ISOGG, date: 27 October 2015; Karmin et al. 2015; Trombetta et al. 2015a,b; Poznik et al. 2016) (Appendix 1).

The following image (figure 6) represents the comparison between the 5966 SNPs identified in our 104 samples and the

variants identified in four recent papers (Francalacci et al. 2013; Hallast et al. 2015; Karmin et al. 2015; Poznik et al. 2016).

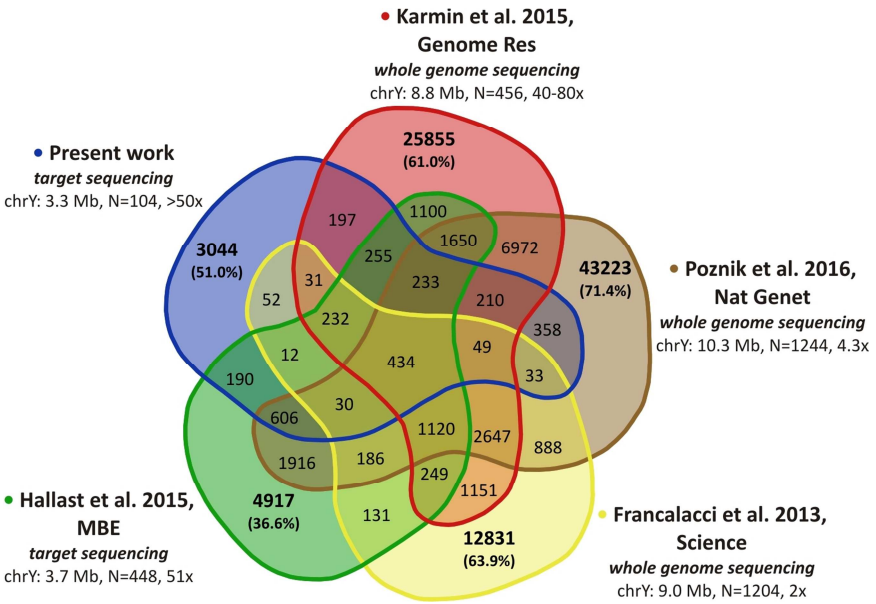


Figure 6: Comparison among the variants reported in the present study and in four recent works. For each dataset, we report: the reference, the experimental approach, the amount of sequenced bases on the human Y chromosome, the number of analysed subjects and the average depth. The number of variants reported for the first time in each study is also expressed as a percentage of the total. The intersection areas of the Venn diagram are not proportional to the number of shared variants.

Among our variants, more than half (51%) was only reported in the present work. This figure is only slightly less than the amount of novel variants found in large studies involving the whole genome sequencing of many hundreds of subjects (Francalacci et al. 2013; Karmin et al. 2015; Poznik et al. 2016). Moreover, our proportion of novel variants is significantly greater (Chi-squared test: $p < 2.2 \times 10^{-16}$) than the one reported in Hallast et al. (2015) (51% *vs.* 36.6%), despite the fact the experimental approach was similar to the one used in this thesis (target sequencing) and the number of sequenced samples was about four times higher.

Dating

We dated the tree nodes using rho statistics, assuming a mutation rate of 0.735×10^{-9}/site/year, which we obtained by calibration with four archeologically-dated specimens (Fu et al. 2014, Lazaridis et al. 2014, Jones et al. 2015). Our rate is coherent with previously published estimates (Fu et al. 2014, Trombetta et al. 2015b). In table 5, the rho values and time estimates, with their standard deviations (SD), are reported for each node, after the exclusion of the four ancient samples from the tree (see Materials and Methods).

Node	Rho	SD Rho	Time (kya)	SD Time (kya)	Node	Rho	SD Rho	Time (kya)	SD Time (kya)
1	420.48	14.81	171.55	6.04	**118**	14.00	2.75	5.71	1.12
2	29.00	3.81	11.83	1.55	**120**	0.50	0.50	0.20	0.20
5	359.57	12.71	146.71	5.19	**123**	39.08	3.78	15.95	1.54
7	348.91	12.36	142.35	5.94	**124**	24.53	2.86	10.01	1.17
8	310.72	14.64	126.77	5.97	**125**	22.00	2.74	8.98	1.12
9	71.50	5.98	29.17	2.44	**128**	1.00	0.71	0.41	0.29
12	224.09	12.89	91.43	5.26	**131**	14.69	1.76	5.99	0.72
14	144.18	10.06	58.83	4.10	**132**	9.50	1.58	3.88	0.65
15	50.50	5.02	20.60	2.05	**135**	7.00	1.87	2.86	0.76
18	26.35	2.23	10.75	0.91	**138**	11.33	2.05	4.62	0.84
19	25.09	2.36	10.24	0.96	**140**	9.50	2.18	3.88	0.89
21	18.67	2.25	7.62	0.92	**143**	12.67	2.68	5.17	1.09
23	15.80	2.09	6.45	0.85	**145**	5.60	1.41	2.28	0.58
26	10.00	1.83	4.08	0.74	**147**	4.50	1.32	1.84	0.54
30	14.75	2.41	6.02	0.98	**149**	4.00	1.33	1.63	0.54
32	13.00	2.47	5.30	1.01	**151**	1.00	0.71	0.41	0.29
34	3.00	1.22	1.22	0.50	**154**	30.14	4.04	12.30	1.65
37	20.67	2.97	8.43	1.21	**155**	6.00	1.73	2.45	0.71
38	7.00	1.87	2.86	0.76	**158**	3.60	0.94	1.47	0.38
41	13.43	2.62	5.48	1.07	**162**	3.50	1.32	1.43	0.54
43	5.83	1.01	2.38	0.41	**162**	175.74	9.32	71.70	3.80
45	3.00	1.22	1.22	0.50	**166**	118.00	7.68	48.14	3.13
51	238.79	9.90	97.42	4.04	**169**	124.54	6.52	50.81	2.66
52	222.00	9.39	90.58	3.83	**171**	122.35	6.49	49.92	2.65
53	32.00	4.00	13.06	1.63	**172**	106.89	7.37	43.61	3.01
56	150.00	8.96	61.20	3.65	**173**	12.33	2.13	5.03	0.87
57	8.00	2.00	3.26	0.82	**176**	11.50	2.40	4.69	0.98
61	180.13	6.91	73.49	2.82	**182**	76.38	6.68	31.16	2.73
62	175.36	9.70	71.55	3.96	**183**	43.07	4.19	17.57	1.71
64	142.39	8.03	58.09	3.28	**184**	11.00	2.35	4.49	0.96
66	139.48	7.98	56.91	3.26	**187**	42.08	4.64	17.17	1.89
68	118.87	6.71	48.50	2.74	**188**	28.83	3.45	11.76	1.41
69	27.05	1.95	11.03	0.79	**189**	17.78	2.44	7.25	1.00
71	25.81	1.76	10.53	0.72	**191**	12.75	1.51	5.20	0.62
72	9.33	2.11	3.81	0.86	**193**	11.00	1.27	4.49	0.52
73	2.00	1.00	0.82	0.41	**198**	9.00	2.12	3.67	0.87
77	18.00	2.45	7.34	1.00	**202**	5.00	1.29	2.04	0.53
81	0.50	0.50	0.20	0.20	**210**	116.81	8.74	47.66	3.57
84	17.00	2.31	6.94	0.94	**214**	88.77	7.30	36.22	2.98
86	14.75	2.11	6.02	0.86	**216**	72.30	6.31	29.50	2.57
87	11.50	2.40	4.69	0.98	**217**	56.79	5.25	23.17	2.14
90	12.00	2.45	4.90	1.00	**218**	8.00	2.00	3.26	0.82
93	20.63	2.37	8.42	0.97	**221**	45.78	4.52	18.68	1.85
94	15.75	2.02	6.43	0.82	**222**	13.67	1.51	5.58	0.62
97	11.00	2.35	4.49	0.96	**229**	19.24	2.20	7.85	0.90
100	9.50	1.54	3.88	0.63	**233**	14.06	1.19	5.73	0.49
105	69.10	5.63	28.19	2.30	**234**	11.50	2.40	4.69	0.98
106	62.00	5.03	25.30	2.05	**240**	13.75	2.30	5.61	0.94
107	40.50	4.50	16.52	1.84	**242**	11.67	2.29	4.76	0.93
111	57.70	5.11	23.54	2.08	**244**	9.50	2.18	3.88	0.89
112	9.00	2.12	3.67	0.87	**247**	12.86	1.95	5.25	0.79
115	40.11	3.42	16.36	1.40	**253**	2.00	0.82	0.82	0.33
116	37.25	4.59	15.20	1.87					

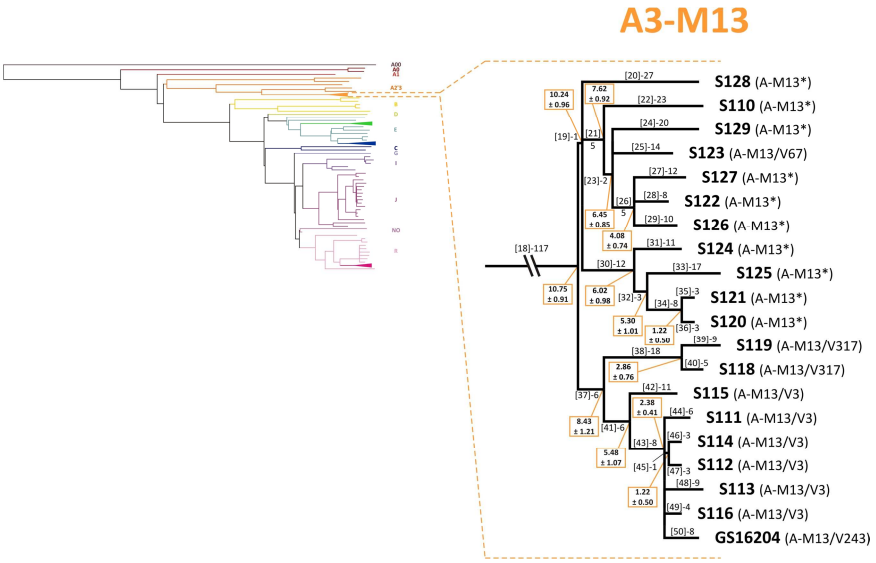
Table 5: Dating of the tree nodes. Each node (in bold) is named after the corresponding branch, according to figure 5.

The four trans-Saharan haplogroups

The phylogenetic relations among samples belonging to the four trans-Saharan clades (A3-M13, E-M2, E-M78 and R-V88) are reported in the following images (figures 7-10).

A3-M13

A3-M13 was characterised by a polytomic topology, with A3-V67, A3-V3 and A3-V317 being sister clades (Scozzari et al. 2012). This topology was partially resolved in a recent NGS study (Scozzari et al. 2014), but this clade remained poorly investigated. Our results showed a completely different structure, with a first bifurcation which distinguished branch 19 from branch 37 (figure 7).



(Previous page) Figure 7: Maximum parsimony tree of the A3-M13 haplogroup. Left: the same tree as figure 5; right: enlarged representation of the structure obtained for the A3-M13 clade. For each branch, the name (in squared brackets) and the number of mutations are reported. The coalescence age and its standard deviation is reported for each node in orange blocks (see table 5). The known haplogroup affiliation of the samples is reported at the tip (in brackets).

The latter included all the sequenced samples defined as A3-V3 and A3-V317, along with a subject (GS16204) published by Karmin et al. (2015) and only defined by V243, which is phylogenetically equivalent to M13 (Scozzari et al. 2012). In our set of SNPs, we also found the V3 marker in branch 41, so we were able to assign GS16204 to the A3-V3 sub-clade. On the contrary, branch 19 included all the sequenced chromosomes belonging to the former A3-M13* paragroup, together with one A3-V67 chromosome.

E-M2

E-M2 has been widely studied due to its high frequency in sub-Saharan Africa. However, previous studies have often focused on its main and most frequent sub-lineages E-M191 and E-U209. Here, we reconstructed the relations among chromosomes belonging to the E-M2* (×M191, U209) paragroup (figure 8).

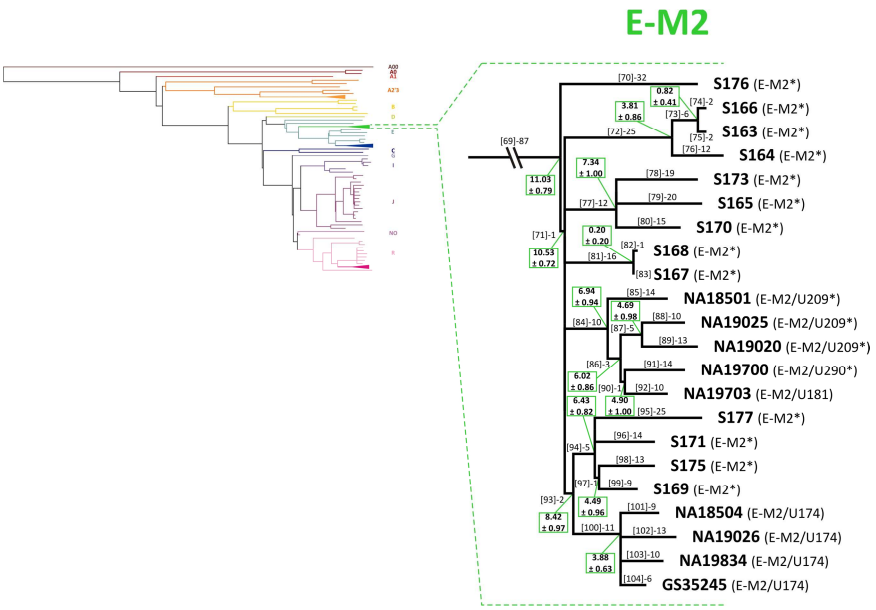
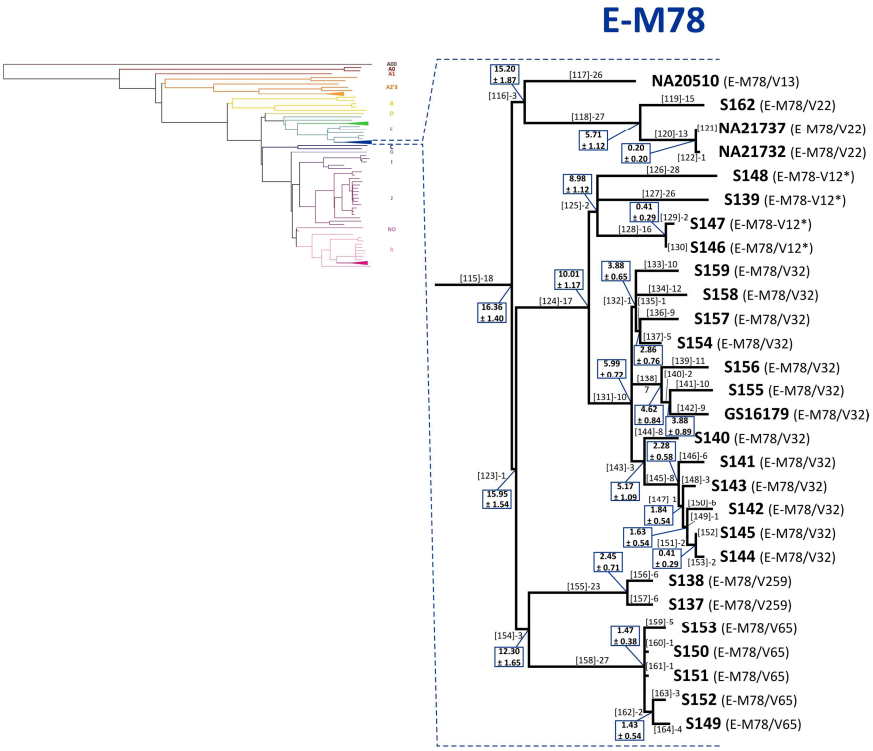


Figure 8: Maximum parsimony tree of the E-M2 haplogroup. Left: the same tree as figure 5; right: enlarged representation of the structure obtained for the E-M2 clade. For each branch, the name (in squared brackets) and the number of mutations are reported. The coalescence age and its standard deviation is reported for each node in green blocks (see table 5). The known haplogroup affiliation of the samples is reported at the tip (in brackets).

The former E-M2* Y chromosomes were found to be arranged in different sub-clades. The first split separated branch 70 from the short branch 71, which included all the other sequenced subjects, arranged in a multiforcation of 5 sister lineages. One of these clades, namely branch 84, corresponded to the E-U209 sub-haplogroup. E-M191 coincided with branch 100, which was found to be a sister clade of a cluster of four former E-M2* chromosomes (branch 94), downstream from the short branch 93. The phylogenetic relations among the E-M2 chromosomes here reported is coherent with the structure published by Poznik et al. (2016) as part of Phase 3 of the 1000 Genomes Project (1000 Genomes Project Consortium 2015).

E-M78

The relationships among the major E-M78 sub-haplogroups were resolved in a recent study (Trombetta et al. 2015a). However, we obtained further information concerning the phylogenetic relations within some sub-lineages of interest (figure 9). First, our results showed that the former E-V12* chromosomes form a monophyletic cluster that is sister to the E-V32 subclade, which in turn is subdivided into three sister clades (branches 132, 138 and 143). Within branch 154, corresponding to E-V264, we focused on the E-V65 subclade, which showed an internal multiforcated structure that seems to have arisen in very recent times (less than 1.5 kya).



(Previous page) Figure 9: Maximum parsimony tree of E-M78 haplogroup. Left: the same tree as figure 5; right: enlarged representation of the structure obtained for the E-M78 clade. For each branch, the name (in squared brackets) and the number of mutations are reported. The coalescence age and its standard deviation is reported for each node in blue blocks (see table 5). The known haplogroup affiliation of the samples is reported at the tip (in brackets).

R-V88

The known structure of R-V88 was formed by a paragroup R-V88* and four sister sub-clades, defined by the M18, V8, V35 and V69 SNPs (Cruciani et al. 2010). In this study, we did not select R-M18 and R-V35 chromosomes and we found the V8 marker in our set of SNPs within branch 241 (figure 10). The first split within our reconstructed phylogeny separated branch 233 from three subjects arranged in three different sister lineages. Branch 233 showed a "star-like" topology, with 18 chromosomes arranged in 8 sister clades, five of which made up by only one subject. Branch 247, one of the 8 sister clades, was defined by only one SNP and included all the sequenced R-V69 chromosomes, so we can assume that this branch corresponds to R-V69 even though this marker was outside our selected MSY regions.

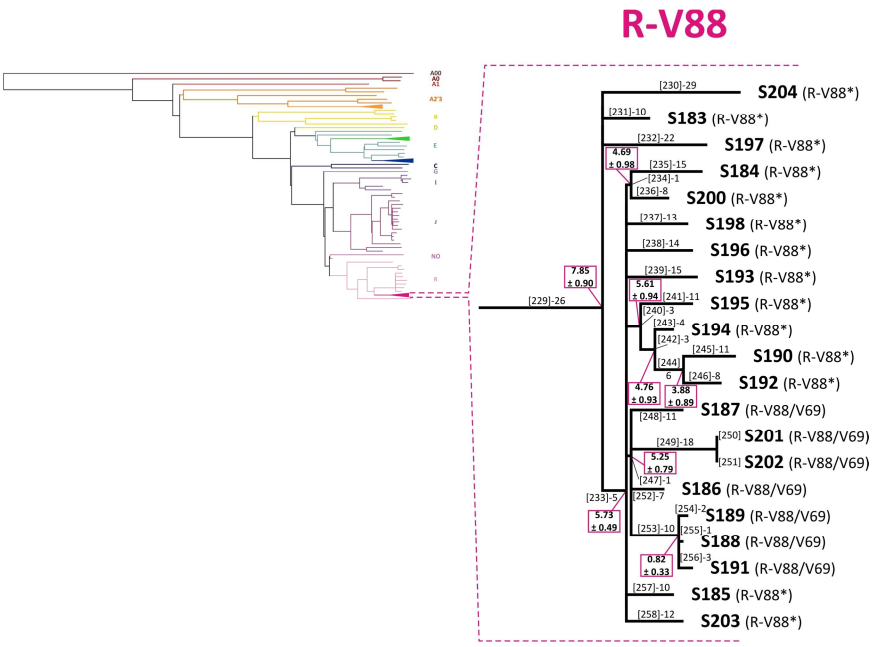


Figure 10: Maximum parsimony tree of R-V88 haplogroup. Left: the same tree as figure 5; right: enlarged representation of the structure obtained for the R-V88 clade. For each branch, the name (in squared brackets) and the number of mutations are reported. The coalescence age and its standard deviation is reported for each node in purple blocks (see table 5). The known haplogroup affiliation of the samples is reported at the tip (in brackets).

Geographic distribution and further molecular dissection of the trans-Saharan clades

In order to gain more information about the ethno-geographic distribution of the four trans-Saharan haplogroups, we selected a total of 108 informative SNPs from those belonging to these lineages and genotyped them in a wider sample (tables 6-9). Our aim was to understand the population movements and interactions during the last African humid period, so we avoided the markers belonging to the clades downstream from more recent nodes (table

5 and figures 7-10). Furthermore, we also checked and considered the information reported in literature for the clades of interest.

The selected SNPs were analysed in more than 5500 males from 124 populations. 87 out of 124 populations were African ethnic groups, which made it possible to obtain detailed data about the distribution of the trans-Saharan clades across different regions of the African continent. The remaining groups we selected were from Europe and the Near East because they are important to understand the dynamics of the trans-Saharan clades outside Africa. We then extracted the distribution data relative to the 108 selected markers from relevant populations reported in other re-sequencing studies. In this way, we were able to add 16 more populations selected among those of the 1000 Genomes Project (Poznik et al. 2016), subdivided in 6 African, 3 European and 7 admixed groups. Moreover, we also added the Sardinian population analysed by Francalacci et al. (2015), for a total of 7690 Y chromosomes from 141 populations (Appendix 2).

On the basis of the genotyping results, we analysed the geographic pattern of the sub-clades within the four trans-Saharan haplogroups.

Molecular dissection of A3-M13

As shown in figure 7, A3-M13 haplogroup had a clear structured phylogeny, which was more evident considering the place of origin of the deep-sequenced chromosomes (figure 11). Branch 37 included all the sequenced chromosomes from the Horn of Africa, while branch 19 showed a more widespread distribution, harbouring lineages from within and outside the African continent.

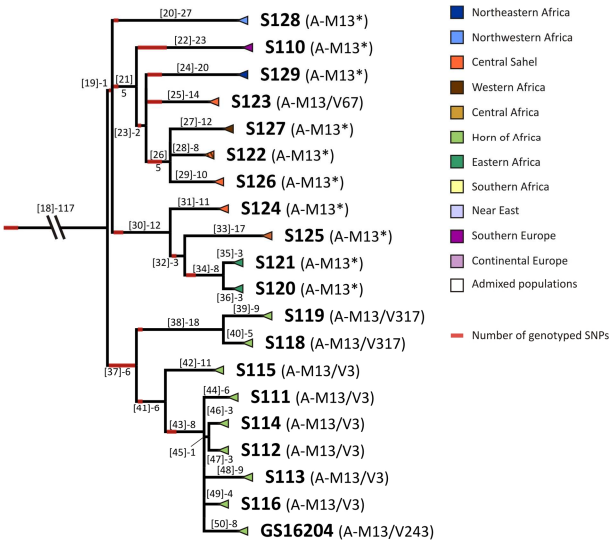


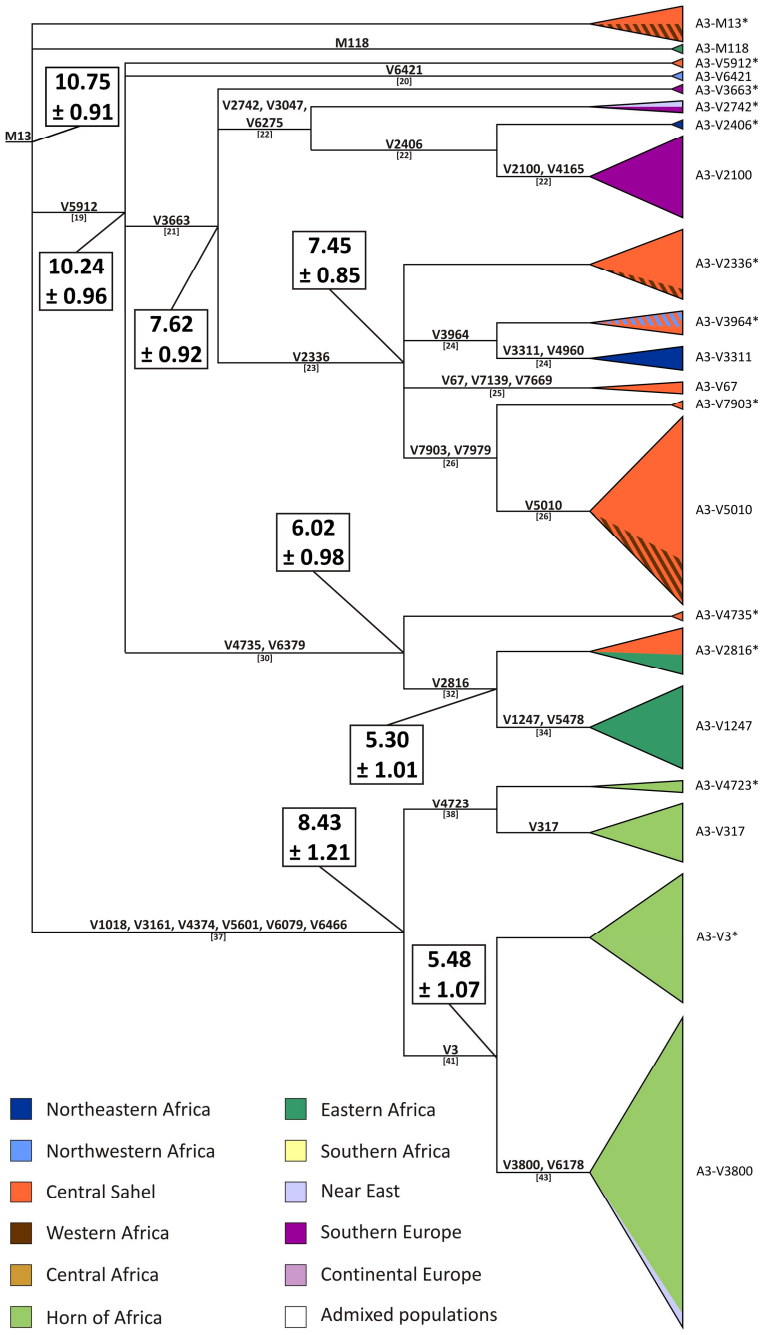
Figure 11: Maximum parsimony tree of A3-M13 with the geographic and genotyping information. The place of origin of each branch is represented by the triangle at the tip, and coloured according to the legend (top right). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions (Appendix 1). Above each branch, the name (in brackets) and the number of SNPs are reported. The red bars on the branches are proportional to the number of markers we selected to be genotyped in a wider sample set.

Considering the topological and geographical features of this clade, we selected 40 SNPs (table 6 and figure 11). Besides the markers within the deep-sequenced samples, we also analysed informative polymorphisms reported in literature but not found in our list of SNPs because their position on the Y chromosome falls outside our targeted regions or because they define lineages not included in our sample set (Appendix 1).

SNP	Direction of mutation	Branch (present study)	Notes
M13	G to C	-	
V1283	C to T	18	
V2734	G to C	18	
V3392	A to G	18	
M118	A to T	-	
V5912	C to T	19	
V6421	C to T	20	reported in this study for the first time
V3663	C to T	21	
V2100	T to C	22	
V2406	C to T	22	
V2742	C to T	22	
V3047	A to G	22	
V4165	A to T	22	
V6275	C to T	22	
V2336	C to T	23	
V3311	T to C	24	
V3964	A to T	24	
V4960	T to C	24	reported in this study for the first time
V67	G to A	25	
V7139	A to G	25	reported in this study for the first time
V7669	T to C	25	reported in this study for the first time
V5010	T to C	26	reported in this study for the first time
V7903	T to C	26	reported in this study for the first time
V7979	T to C	26	reported in this study for the first time
V4735	A to G	30	
V6379	T to A	30	reported in this study for the first time
V2816	A to C	32	
V1247	C to A	34	
V5478	C to G	34	reported in this study for the first time
V1018	C to T	37	
V3161	C to T	37	
V4374	A to C	37	
V5601	C to T	37	
V6079	C to G	37	
V6466	G to A	37	
V4723	C to A	38	reported in this study for the first time
V317	G to A	-	
V3	T to C	41	
V3800	C to T	43	
V6178	G to C	43	

Table 6: Selected SNPs within the A3-M13 haplogroup genotyped in a wider sample. For each SNP, the direction, the branch of the tree in the present study and information about the newly reported polymorphisms are indicated.

The results of the genotyping of these markers are reported in Appendix 2 and summarised in the following image (figure 12).


A3-M13*
M118
A3-M118
V6421
[20]
A3-V5912*
A3-V6421
A3-V3663*
10.75
± 0.91
V2742, V3047,
V6275
[22]
A3-V2742*
A3-V2406*
M13
V2406
[22]
V2100, V4165
[22]
A3-V2100
V5912
[19]
V3663
[21]
7.45
± 0.85
A3-V2336*
10.24
± 0.96
7.62
± 0.92
V3964
[24]
A3-V3964*
V3311, V4960
[24]
A3-V3311
V2336
[23]
V67, V7139, V7669
[25]
A3-V67
A3-V7903*
V7903, V7979
[26]
V5010
[26]
A3-V5010
6.02
± 0.98
A3-V4735*
V4735, V6379
[30]
A3-V2816*
V2816
[32]
V1247, V5478
[34]
A3-V1247
5.30
± 1.01
A3-V4723*
V4723
[38]
8.43
± 1.21
V317
A3-V317
V1018, V3161, V4374, V5601, V6079, V6466
[37]
A3-V3*
5.48
± 1.07
V3
[41]
V3800, V6178
[43]
A3-V3800
Northeastern Africa
Northwestern Africa
Central Sahel
Western Africa
Central Africa
Horn of Africa
Eastern Africa
Southern Africa
Near East
Southern Europe
Continental Europe
Admixed populations

(Previous page) Figure 12: Molecular dissection of the A3-M13 haplogroup. The branches downstream from the last genotyped SNPs are contained in the form of triangles proportional to the number of derived Y chromosomes. Each triangle is subdivided into areas proportional to the number of subjects belonging to each macroregion (Appendix 2) and coloured according to the legend (bottom left). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions. For each branch, the analysed SNPs and the corresponding branch number of figure 11 (in squared brackets) are reported. Grey branch numbers: their phylogenetic position was deduced on the basis of the genotyping results, although the analysed SNPs were not present among the markers identified by NGS. The SNPs V1283, V2734 and V3392 were found to be phylogenetically equivalent to M13 (not shown).

The geographic differentiation shown by the A3-M13 phylogeny from the NGS data (figure 11) was confirmed and enhanced by the results of the genotyping (Appendix 2 and figure 12). Most clades were restricted to one geographic area and it is worth noting that the southern European clade A3-V2100 was restricted to Sardinia, including all the seven A3-M13 chromosomes reported by Francalacci et al. (2013, 2015). When more regions were represented within the same sub-haplogroup, there were no more than two. In these cases, the lineages were not shared among African and non-African samples, apart from the little near-eastern component within the A3-V3800, which was mainly from the Horn of Africa. Furthermore, nodes harbouring both northern African and sub-Saharan lineages showed time estimates within the last Green Sahara period.

Analysing in more detail the frequency information reported in Appendix 2 and represented in figure 12, A3-M13 was found with frequencies as high as 40% in the African continent, which is consistent with the data from literature. The highest frequency was observed in the populations from the Horn of Africa, completely represented by the sub-clade A3-V1018. Its sister clade A3-V5912 was more widespread. It reaches the highest frequencies in the central Sahel, but it has also been observed at lower frequencies in Kenya, Sardinia and Egypt.

Molecular dissection of E-M2

Figure 8 represents how the former E-M2* chromosomes belong to different sister lineages. Considering the place of origin of the deep-sequenced samples from our lab collection (table 1), we observed that the majority of the northern African sequenced chromosomes (7 out of 8) were clustered in three subclades (branches 72, 77, 81), which were mainly or exclusively distributed in this geographic area (figure 13).

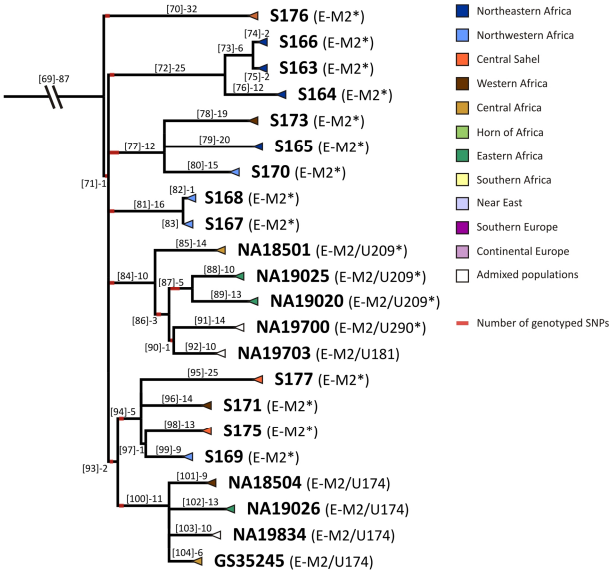


Figure 13: Maximum parsimony tree of E-M2 with the geographic and genotyping information. The place of origin of each branch is represented by the triangle at the tip, coloured according to the legend (top right). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions. (Appendix 1). Above each branch, the name (in brackets) and the number of SNPs are reported. The red bars on branches are proportional to the number of selected and genotyped markers.

When selecting the polymorphisms to be genotyped, we paid particular attention to those variants shared with the 262 E-M2

chromosomes from the 1000 Genomes Project (Poznik et al. 2016) in order to increase the discrimination power of our analysis. Through this comparison, we identified some branches (namely branch 72 and its internal clades, branch 79, branch 81 and its sub-lineages, branches 95, 98 and 99) which shared no markers with the 1000 Genomes Project chromosomes (Appendix 1). While branches 79, 95, 98 and 99 represented terminal clades downstream from branches sharing SNPs with the 1000 Genomes Project lineages, branches 72 and 81 were two deep sister lineages within the E-M2 main multifurcation. Interestingly, both of them included deep-sequenced chromosomes from northern Africa. Similarly, the other lineages not found in the 1000 Genomes Project were mainly represented by northern African subjects (tables 1-2 and figure 13).

To obtain a detailed picture of the distribution of E-M2 sub-haplogroups in Africa, in a wider sample set, we genotyped 15 potentially informative SNPs (figure 13 and table 7), chosen taking into account our resequencing data, information from the 1000 Genomes Project and data from literature (Karafet et al. 2008).

SNP	Type of mutation	Branch (present study)	Notes
M2	A to G	-	
V4257	T to C	70	
M4727	A to C	71	
M10	T to C	-	
V5001	C to T	72	reported in this study for the first time
Z15941	T to G	77	
Z15943	C to T	77	
A186	C to T	81	
U209	C to T	84	
M58	G to A	-	
V1891	G to A	93	
L516	A to C	94	
M191	T to G	100	
U174	G to A	100	
8019527[a]	G to T	-	

Table 7: Selected SNPs within the E-M2 haplogroup genotyped in a wider sample. For each SNP, the direction, the branch of the tree in the present study and information about the newly reported polymorphisms are indicated. The markers absent in our dataset and reported in other studies without a name are indicated with their Y chromosome position according to the Feb. 2009 (GRCh37/hg19) assembly of the human genome.

The results of the genotyping analysis of the E-M2 sub-clades (Appendix 2 and figure 14) showed how the geographic spread of E-M2 was mainly due to the E-M4727 sub-haplogroup, while the rare E-V4257 was scattered across western Africa and central Sahel. Within E-M4727, E-U209 and E-U174 were the most frequent clades, being present at average-high frequencies in all the sub-Saharan region, from the Sahara to the southern part of the African continent. Interestingly, the sister clade of E-U174, namely E-8019527, was very rare and scattered. The other lineages with a high number of subjects were E-Z15941 and E-L516. The first was one of the sister lineages within the E-M2 multifurcation and showed a strong western African component, with also a large number of Fulbe subjects from Nigeria and northern Cameroon/Chad. Interestingly, it also harboured a good proportion of northern African samples, in particular from the western areas. E-L516 was the sister clade of E-M191 and mainly included central Sahelian chromosomes. E-V5001, corresponding to branch 72, remained restricted to north-eastern Africa, while E-A186

(branch 81) also included some nomadic Fulbe subjects. It is worth noting that the lineages mainly distributed in western Africa also harboured a relatively high proportion of subjects from American admixed populations with African ancestry.

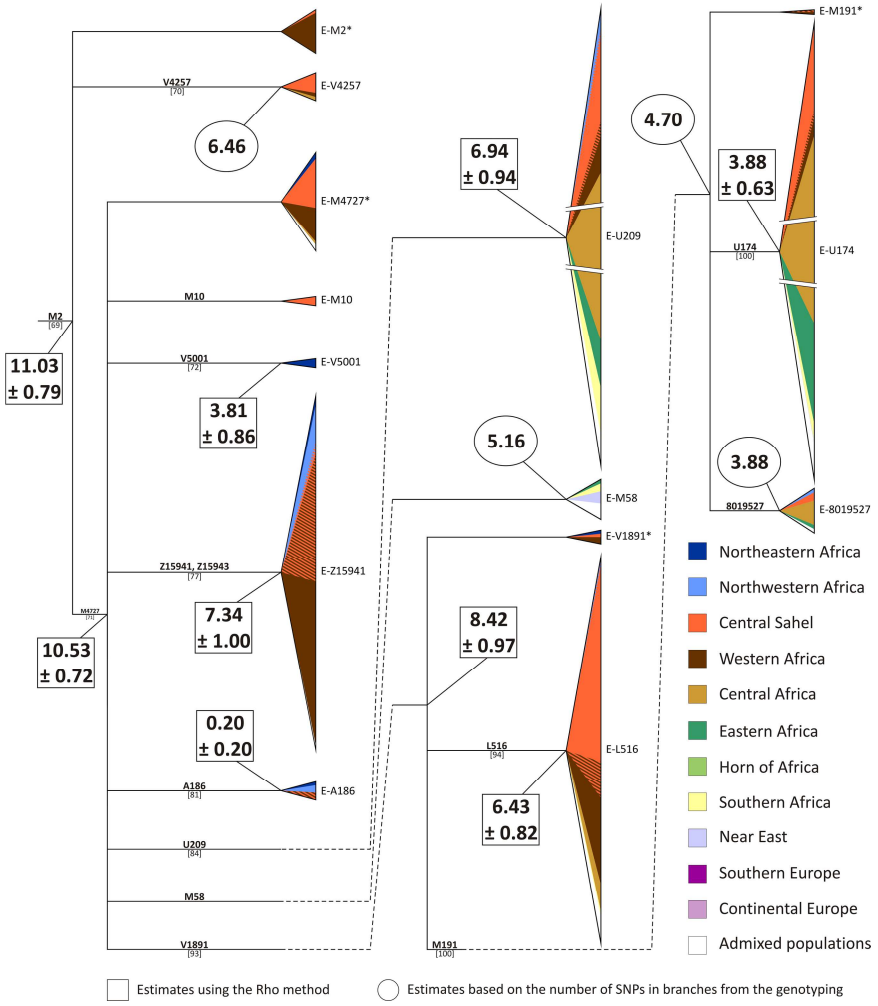
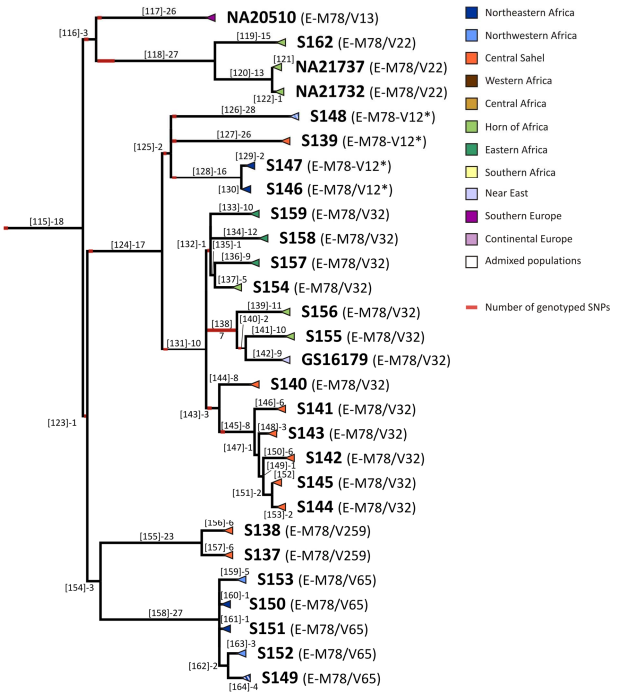


Figure 14: Molecular dissection of the E-M2 haplogroup. The branches downstream from the last genotyped SNPs are contained in the form of triangles proportional to the number of derived Y chromosomes. Each triangle is subdivided into areas proportional to the number of subjects belonging to each macroregion (Appendix 1) and coloured according to the legend (top left). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions. If the clades harboured more than 100 subjects, they are represented by interrupted triangles. For each branch, the analysed SNPs and the corresponding branch number of figure 13 (in squared brackets) are reported. Grey branch numbers: their phylogenetic position was deduced on the basis of the genotyping results, although the SNPs analysed were not present in the markers identified by NGS.

Molecular dissection of E-M78

As described above, the E-M78 topology here reported is coherent with the structure obtained in a recent study (Trombetta et al. 2015a). Considering its geographic distribution, E-M78 is one of the few haplogroups present in Africa, the Near East and Europe at relevant frequencies. However, its global distribution is due to the geographic restricted distribution of its internal lineages. This feature is evident in figure 15, where the place of origin of the deep-sequenced chromosomes is reported.



(Previous page) Figure 15: Maximum parsimony tree of E-M78 with the geographic and genotyping information. The place of origin of each branch is represented by the triangle at the tip, coloured according to the legend (top right). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions (Appendix 1). Above each branch, the name (in brackets) and the number of SNPs are reported. The red bars on branches are proportional to the number of selected and genotyped markers.

We chose not to analyse the E-V13 sub-lineage, because past studies have shown that it was involved in the Neolithic transition in the Near East (Cruciani et al. 2007), and this is not the focus of the present thesis. With the exception of a single E-V13 sample, all the other E-M78 subjects in figure 15 came from the African continent. It is worth noting that the only lineage which harboured samples from northern and sub-Saharan regions was the cluster formed by the former E-V12* chromosomes. All the other lineages are restricted to the north or to the south of the Sahara and the nodes joining them dated during the last African humid period.

These interesting features were further explored choosing 31 markers which were analysed in the wider sample. The informative SNPs were selected on the basis of their phylogenetic position and/or geographic distribution, taking into account also data from literature (Cruciani et al. 2004, 2007; Karafet et al. 2008; Karmin et al. 2015; Trombetta et al. 2015a). In particular, we decided to focus mainly on the E-V22 and E-V12 clades.

SNP	Type of mutation	Branch (present study)	Notes
M78	C to T	115	
V1477	G to C	-	
V1083	C to G	116	
V13	G to A	117	
V22	T to C	118	
CTS5479	A to G	118	
V3536	C to A	-	
17082060[a]	G to A	-	
V3262	G to A	118	
V3948	A to C	118	
V1129	T to C	123	
V12	A to G	124	
CTS693	G to A	125	
V7165	T to A	126	
V4859	C to T	127	
V2629	G to A	128	
V32	G to C	131	
V3746	C to T	132	
V4381	A to G	138	
V4679	C to T	138	
V5184	C to T	138	reported in this study for the first time
V5712	T to A	138	
V5767	C to T	138	
V6888	T to C	138	
V6892	C to G	138	
V7591	G to C	140	
V6873	G to A	143	reported in this study for the first time
V7399	T to C	145	reported in this study for the first time
V264	C to T	-	
V259	C to T	-	
V65	C to A	-	

Table 8: Selected SNPs within the E-M78 haplogroup genotyped in a wider sample. For each SNP the direction, the branch of the tree in the present study and information about the newly reported polymorphisms are indicated. The markers absent in our dataset and reported in other studies without a name are indicated with their Y chromosome position according to the Feb. 2009 (GRCh37/hg19) assembly of the human genome.

The results are reported in Appendix 2 and shown in figure 16.

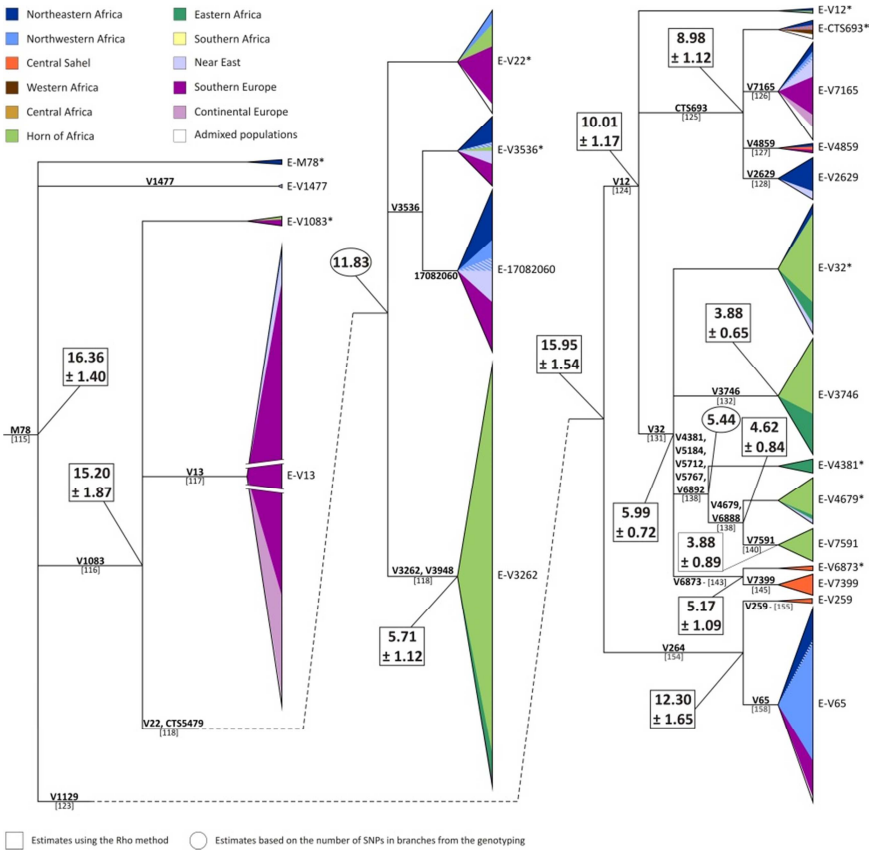


Figure 16: Molecular dissection of the E-M78 haplogroup. The branches downstream from the last genotyped SNPs are contained in the form of triangles proportional to the number of derived Y chromosomes. Each triangle is subdivided into areas proportional to the number of subjects belonging to each macroregion (Appendix 1) and coloured according to the legend (top left). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions. If the clades harboured more than 100 subjects, they are represented by interrupted triangles. For each branch, the analysed SNPs and the corresponding branch number of figure 15 (in squared brackets) are reported. Grey branch numbers: their phylogenetic position was deduced on the basis of the genotyping results, although the SNPs analysed were not present in the markers identified by NGS.

For the genotyping of variants within the E-V22 clade, we selected two SNPs (V3536 and chrY:17082060) reported for the E-V22 “mixed” sample in the work by Karmin et al. (2015) and one

SNP (V3262) within our branch 118 and absent in the dataset of Karmin et al. (2015). Through this approach, we were able to identify two different subclades from different geographic areas. In fact, most of the eastern African samples belonged to the E-V3262 lineage, which did not contain subjects from other geographic areas. Interestingly, it reached a frequency as high as 80% in the Saho from Eritrea. On the contrary, chromosomes from other African, European and Near Eastern regions were found to belong to the E-V3536 lineage or to the paragroup E-V22*. Considering the node age (figure 16 and table 5), these clades split before 5 kya, during the Green Sahara.

An analogous scenario was obtained from the genotyping of informative SNPs within the E-V12 sub-haplogroup, with the E-V32 restricted to sub-Saharan Africa and the E-CTS693 mainly present in the Mediterranean regions. E-CTS693 showed very low frequencies, with a maximum of 14% in the Baharia oasis in Egypt. On the contrary, its sister clade, E-V32, reached high frequencies in eastern Africa and low frequencies in southern Chad, where its presence is due to the E-V6873 sub-clade. Similarly to the E-V22 sub-haplogroups, these lineages separated during the Green Sahara time frame (figure 16 and table 5).

Finally, even the E-V264 showed a sharp differentiation between sub-Saharan Africa, represented by the E-V259 clade, and northern Africa, represented by E-V65, with the most recent common node dating back to the beginning of the Green Sahara period (figure 16 and table 5).

Molecular dissection of R-V88

Among the four trans-Saharan haplogroups, R-V88 was the one with a less structured internal topology and its lineages did not show any clear geographic pattern (figures 10 and 17).

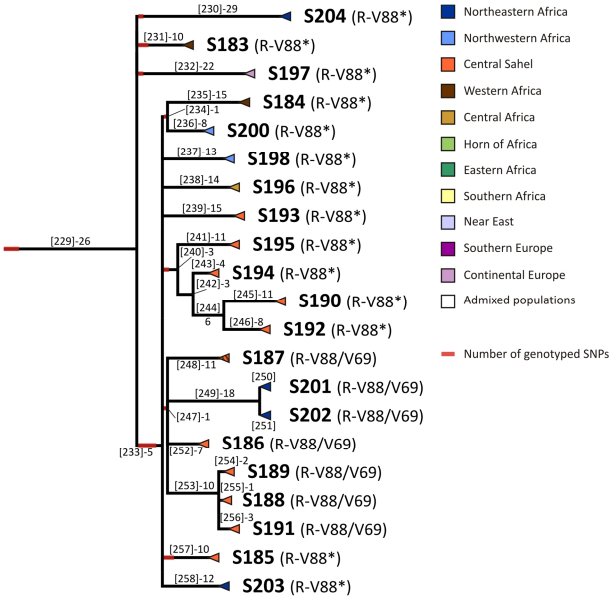


Figure 17: Maximum parsimony tree of R-V88 with the geographic and genotyping information. The place of origin of each branch is represented by the triangle at the tip, coloured according to the legend (top right). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions (Appendix 1). Above each branch, the name (in brackets) and the number of SNPs are reported. The red bars on branches are proportional to the number of selected and genotyped markers.

Considering the star-like structure of this clade, we decided to genotype at least one marker for each of the four most basal clades, which were branches 230, 231, 232 and 233. Within branch 233, which formed another star-like structure, we decided to focus on the sub-haplogroups composed of more than one subject to which

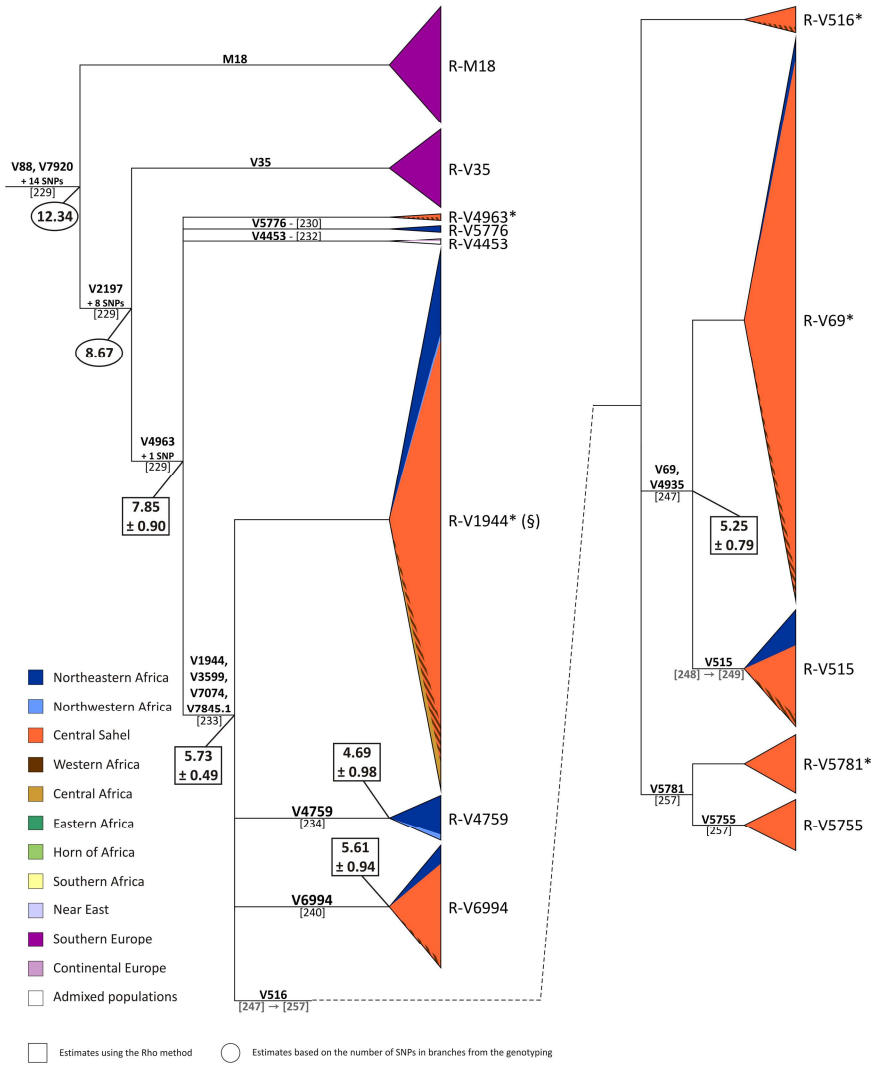
we added one of the other internal lineages, specifically branch 257 (table 9)

SNP	Type of mutation	Branch (present study)	Notes
V88	C to T	-	
M18	ins AA	-	
V35	T to A	-	
V2197	G to A	229	
V4963	A to G	229	
V7920	C to G	229	
V5776	T to C	230	reported in this study for the first time
V4238	T to A	231	reported in this study for the first time
V7260	C to T	231	reported in this study for the first time
V4453	G to A	232	
V1944	G to A	233	
V3599	T to C	233	
V7074	T to C	233	reported in this study for the first time
V7845.1	C to A	233	reported in this study for the first time
V4759	G to A	234	reported in this study for the first time
V6994	G to C	240	reported in this study for the first time
V516	A to G	-	reported in this study for the first time
V4935	T to C	247	reported in this study for the first time
V69	C to T	-	
V515	G to T	-	reported in this study for the first time
V5755	T to C	257	reported in this study for the first time
V5781	G to C	257	reported in this study for the first time

Table 9: Selected SNPs within the R-V88 haplogroup genotyped in a wider sample. For each SNP, the direction, the branch of the tree in the present study and information about the newly reported polymorphisms are indicated. The markers absent in our dataset and reported in other studies without a name are indicated with their Y chromosome position according to the Feb. 2009 (GRCh37/hg19) assembly of the human genome.

The genotyping results enhanced the R-V88 features highlighted by the NGS data (figure 18). However, although this haplogroup maintained a poorly differentiated geographic pattern, it was possible to note some peculiarities. First of all, R-V88 was mainly an African clade, but it had two basal internal lineages (R-M18 and R-V35) which were restricted to southern Europe, more specifically to Sardinia. All the African subjects belonged to the R-V4963 subclade, which also included the non-African R-V4453 clade. R-V1944 was the first R-V88 sub-haplogroup restricted to the African continent and it harboured most of the R-V88 subjects

(Appendix 2). The ages of the nodes joining northern African and sub-Saharan lineages fell in a time frame of 8 - 5 kya (figure 18 and table 5).



(Previous page) Figure 18: Molecular dissection of the R-V88 haplogroup. The branches downstream from the last genotyped SNPs are contained in the form of triangles proportional to the number of derived Y chromosomes. Each triangle is subdivided into areas proportional to the number of subjects belonging to each macroregion (Appendix 1) and coloured according to the legend (bottom left). Bi-coloured areas represent nomadic populations (Tuareg and Fulbe) currently dispersed over different geographic macroregions. If the clades harboured more than 100 subjects, they are represented by interrupted triangles. For each branch, the analysed SNPs and the corresponding branch number of figure 17 (in squared brackets) are reported. Grey branch numbers: their phylogenetic position was deduced on the basis of the genotyping results, although the SNPs analysed were not present in the markers identified by NGS. § The lineage R-V1944* also includes branches 237, 238, 239 and 258, which were not selected for genotyping (figure 17).

The feature emerging from the frequency distribution (Appendix 2) was that R-V88 and, more specifically, its most frequent subclade R-V1944 were fundamentally saharan/sahelian haplogroups, being present from northern Egypt to lake Chad and further south, although with variable frequencies. Within R-V1944, R-V6994 reached its maximum in the central Sahel (Chad), while its main sister clade R-V516 showed a dual distribution. In fact, it reached frequencies of more than 80% in Cameroon and Chad, while it was found to be more widespread but much less frequent (~ 5%) in north-eastern Africa. This geographic pattern is mainly due to its internal lineage R-V69. The sister clade of R-V69, namely R-V5781, is restricted to the Ouldeme people in northern Cameroon.

DISCUSSION

The advantages of the targeted sampling approach

With the introduction of NGS techniques, an increasing amount of data concerning the whole genome human variability is becoming available (1000 Genomes Project Consortium 2015; Gurdasani et al. 2015; Malaspinas et al. 2016; Mallick et al. 2016; Pagani et al. 2016). The study of the human Y chromosome also took a great leap forward thanks to NGS, but there were important differences compared to autosomes. In fact, because of its peculiar features, the Y chromosome has a greater inter-population variability. For this reason, the sampling method is of primary importance when interpreting data about Y variability. This aspect is highlighted in figure 6, where the Y chromosome variability reported in three whole genome sequencing projects (Francalacci et al. 2013; Karmin et al. 2015; Poznik et al. 2016) and in one target sequencing project (Hallast et al. 2015) was compared with the data from the present study. All the previous projects used an unbiased sampling method, analysing subjects chosen independently from their haplogroup affiliation and belonging to few African populations. On the contrary, we used a targeted method, choosing samples known to have a peculiar haplogroup affiliation and which are distributed in many different populations. A consequence of this sampling strategy is that 51% of the mutations we discovered were new variants, while Hallast et al. (2015), who analysed about the same amount of bp in more samples, reported only 37% of variants which were not shared with other studies (figure 6).

Due to the fact the focus of the present thesis concerns the analysis of four haplogroups which are quite frequent in some African regions and quite rare in other areas of Africa, an unbiased method would have failed to take into account the variability embedded in rare sub-lineages. The E-M2 haplogroup is a perfect

example of this, because it has been extensively analysed both in the present work and in Phase 3 of the 1000 Genomes Project (Poznik et al. 2016). Poznik et al. (2016) analysed 262 complete E-M2 chromosomes from 5 African ethnic groups and 5 American populations of African ancestry. On the contrary, we obtained NGS data for 3.3 Mb of the Y chromosome from 22 subjects belonging to 13 African populations. Despite the lower number of both the deep-sequenced subjects and bases, we were able to find 11 new branches. They included subjects from northern Africa, a region that was not covered in the 1000 Genomes Project sampling. Most of the subjects analysed by Poznik et al. (2016) were found to belong to E-U209 and E-M191, which are the most frequent E-M2 sub-lineages in the African continent and, more specifically, in the ethnic groups analysed in that study.

In conclusion, the random sampling of few groups in a broad area cannot be considered to be a good representation of Y variability in the whole region, due to the peculiar features of the human Y chromosome. Moreover, even with patterned sampling, rare lineages run the risk of being missed, with a consequential loss of information. Nonetheless, low-frequency lineages could represent the footprints of important events of the human history and therefore they could be highly informative. In such cases, a targeted sampling approach based on previous knowledge about haplogroup affiliation would be preferable.

The Green Sahara and the four trans-Saharan clades

In the present thesis, we found considerable evidence concerning the effects of Holocene climatic changes on the distribution of the four haplogroups here analysed.

In general, we found that the nodes that join lineages from northern and sub-Saharan Africa coalesce during the last African humid period (12-5 kya). On the contrary, most clades restricted to either northern or sub-Saharan Africa coalesce after 5 kya. These findings suggest that the present distribution of the trans-Saharan haplogroups cannot only be put down to very recent events, as suggested in previous studies (Luis et al. 2004; Rosa et al. 2007; Francalacci et al. 2013, 2015). On the contrary, it seems that the current distribution was heavily influenced by events that took place during the Green Sahara period.

Despite the general concordance of results for the four trans-Saharan haplogroups, there were differences in their specific history and geographic distribution. In fact, the onset of the arid conditions at the end of the Green Sahara period was more abrupt in eastern Africa compared to the central Sahel, where the presence of an extensive hydrogeological network buffered the climatic change, which, in fact, was not complete before ~ 4 kya (Lézine et al. 2011). Therefore, departure from eastern Sahara was more rapid than from central Sahara, a fact testified by the different density of archaeological evidence (Manning and Timpson 2014).

The central Sahara

The central part of the Saharan/Sahelian belt is characterised by the presence of Lake Chad. During the Green Sahara period, the lake occupied an area of ~ 400,000 km^2 (Lake Mega-Chad) and it

was linked to other northern megalakes by rivers and wetlands (Drake et al. 2011). This water network formed a corridor across the Sahara connecting present-day Algeria, Libya, Niger, Chad and the northern regions of Cameroon and Nigeria.

In this part of Africa, we found lineages belonging to three of the four trans-Saharan clades: A3-M13, E-M2 and R-V88.

A3-M13 during the Green Sahara

Based on its phylogeny and geographic distribution, the first northward movement linked to A3-M13 involved the A3-V5912 sub-clade and it took place between 10.75 ± 0.91 and 10.24 ± 0.96 kya from its homeland, which was probably somewhere in the sub-Saharan regions between central and eastern Africa (figures 12 and 19). This is coherent with the hypothesis of the occupation of central Sahara by sahelian hunter-gatherer groups, which moved northward following the spread of the savannah landscape (di Lernia 2002; Brooks 2006). After the arrival of A3-V5912 (figure 19 panel B) in a undefined region in the central Sahara, there was a hiatus of 3-4 kya, followed by the formation of several lineages including a large number of subjects (figure 12). In this context, we found the footprints of two main events: 1) a green-saharan expansion northward, represented by the lineage A3-V3663 (figure 19 panel D) and, more specifically, by its internal clade A3-V2336 (figure 19 panel E); 2) a sahelian movement from the central Sahel to eastern Africa represented by A3-V4735 (figure 19 panel F; described in detailed in a later section).

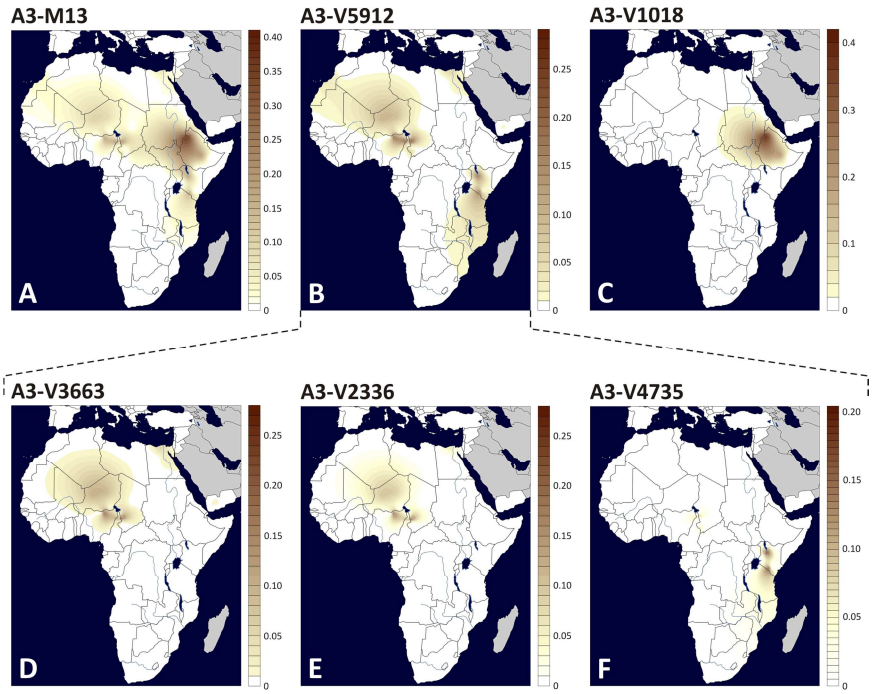


Figure 19: Frequency maps of A3-M13 and its internal clades. Each panel represents the geographic distribution of a haplogroup, indicated above. The countries not investigated in the present thesis are in grey. Please note that each lineage is represented with its own scale, adjusted according to its frequency distribution (to the right of each panel). When the panels represent the internal clades of another panel, they are enclosed by dashed lines.

After the arrival of A3-V5912 in the central Sahara, the internal clade, A3-V3663, differentiated between 10.24 ± 0.96 and 7.62 ± 0.92 kya. A3-V3663 seems to have followed two different evolutionary routes, with the A3-V2742 clade (figure 12) which arrived as far north as Sardinia, while A3-V2336 expanded in the central Sahel between 7.62 ± 0.92 and 7.45 ± 0.85 kya and arrived as far north as Egypt. Considering the topology of this clade and its frequency distribution, the centre of the expansion could be near lake Chad, where the highest frequencies are found.

We lack the information necessary to date the arrival in Egypt precisely, since we did not analyse all the 20 SNPs of branch 24

(figure 11). However, among the three random chosen markers genotyped in the whole sample, only one polymorphism (V3964) was found to be shared between the Egyptian and central samples and this finding can be interpreted as a clue which indicates a northward movement that occurred in not recent times (figure 12).

It is worth noting that during the early Green Sahara (more specifically, between 10.75 ± 0.91 and 8.43 ± 1.21 kya) a branch (A3-V1018) moved from the place of origin of A3-M13 towards the Horn of Africa, where it underwent an independent evolution (figure 12 and figure 19 panel C).

In conclusion, the strong geographic differentiation of the A3-M13 internal clades suggests that they have been separated for a long time and our time estimates indicate the end of the Green Sahara period as a possible cause of their separation. Moreover, the direction of the movements seems to have followed the northward spread of the fertile environment from the Sahel towards the central Sahara, according to the most common scenario reported in literature.

E-M2 during the Green Sahara

The signals found in the E-M2 haplogroup are less pronounced, because this haplogroup was heavily influenced by the Bantu expansion, which mainly involved the E-U209 and E-U174 sub-lineages (figures 14 and 20, panels E and H). Apart from these two sub-clades, we found at least two Saharan/Sahelian clades (E-Z15941 and E-L516) (figure 20, panels D and G) and one lineage restricted to northeastern Africa (E-V5001) (figure 14).

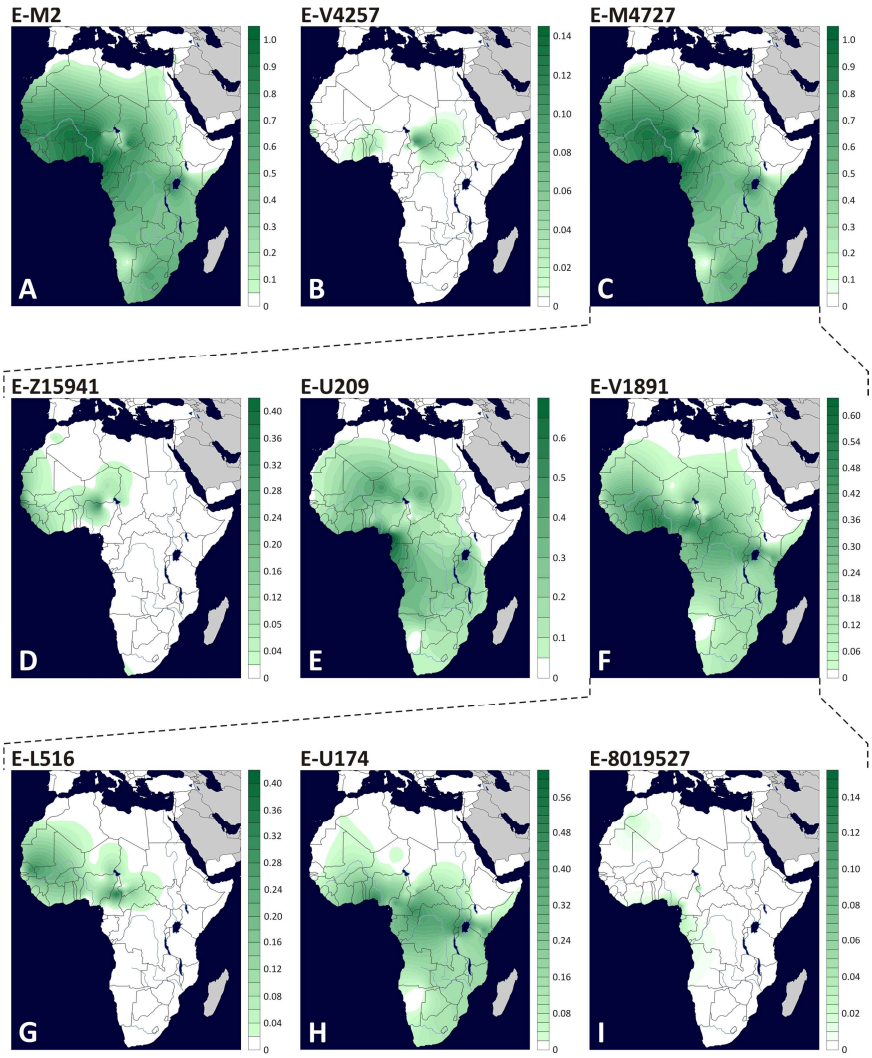


Figure 15: Frequency maps of E-M2 and its internal clades. Each panel represents the geographic distribution of a haplogroup, indicated above. The countries not investigated in the present thesis are in grey. Please note that each lineage is represented with its own scale adjusted according to its frequency distribution (to the right of each panel). When the panels represent the internal clades of another panel, they are enclosed by dashed lines.

The differentiation of E-Z15941 dates back to 10.53 ± 0.72 - 7.34 ± 1.00 kya. The frequency map of this clade (figure 20, panel D) shows that this haplogroup is found around lake Chad, along the Niger river and northward as far as Morocco and Egypt. In order to shed light on the past events of this lineage, we compared our SNPs with those found in the Y chromosomes from the 1000 Genomes Project (Appendix 1), which mainly come from western Africa (Poznik et al. 2016). In this *in silico* analysis, we only considered the 1000 Genomes Project SNPs within our sequenced 3.3 Mb. From our NGS, E-Z15941 was found to have three sub-lineages, namely branch 78, 79 and 80 (figure 13). Branch 79 comes from northeastern Africa and shares no SNPs with the 1000 Genomes Project samples, suggesting that its separation from the western/central populations occurred in ancient times. The subject representing branch 78 is a Mandenka from Senegal and shares only one SNP with a short sub-clade found in the Mandenka from Gambia analysed in the 1000 Genomes Project, suggesting that they arrived in the western regions long enough ago to differentiate within the same population. Finally, branch 80 comes from northwestern Africa and shares 1/15 SNPs with a Gambian terminal branch from the 1000 Genomes Project. This suggests that the separation of branch 80 and the Gambian lineage occurred in ancient times, possibly around 7 kya considering that each mutation occurred every ~ 408 years. All these findings suggest that the origin and first dispersal across the Sahara of E-Z15941 was linked to the African humid period. However, it reaches the highest frequencies among the Fulbe and the western populations and their possible impact on the spread of this lineage will be described in a later section.

The other Sahelian/Saharan clade, E-L516, differentiated between 8.42 ± 0.97 and 6.43 ± 0.82 kya in a broad area ranging from the central Sahel to the western coast. It was found with different frequencies in different populations with different linguistic affiliations, with no clear link with a single ethnic group (figure 20, panel G). By comparing the SNPs of the 1000 Genomes

Project with our variants (Poznik et al. 2016), we only found matches for branch 96 and 97 (Appendix 1). Branch 96 comes from western Africa and shares 2 SNPs with a 1000 Genomes Project sub-branch, including three western African samples. Branch 97 harbours two samples, one from northern Cameroon and the other from Morocco, and it corresponds to a 1000 Genomes Project short branch, including 3 American admixed subjects. The African ancestry found in American samples is mainly due to the Atlantic slave trade which took place between the XV and XIX centuries. Considering the Y chromosome variation, this recent event is hardly distinguishable from the present and the admixed populations are very similar to the source population from western Africa (Tishkoff et al. 2009). Interestingly, our subjects shared no polymorphisms with the admixed samples, even though they belong to the same sub-branch. This lack of common variants suggests that our samples separated from the admixed subjects before the Atlantic slave trade and just after the time estimates of node 97 (4.49 ± 0.96 kya). It is worth noting that not all the branches not identified by the 1000 Genomes Project (branch 95, 98 and 99) come from the western regions, but rather they include subjects from northern Sahel and northern Africa. All these findings suggest that E-L516, similarly to E-Z15941, was present in a Green Saharan population which settled in the broad area of the megalakes between northern Africa and the Sahel, with the first dispersal linked to the humid period.

Interestingly, we found a clade restricted to northeastern Africa, namely E-V5001. More specifically, this clade is found at low frequencies (1-5%) in the Siwa and Baharia oasis in the Egyptian Sahara, where it arrived between 10.53 ± 0.72 and 3.81 ± 0.86 kya (figure 14 and Appendix 2). This can be interpreted as the footprint of a wider and more ancient distribution across the Sahara, which was erased and segregated in the small saharan oasis after the desertification of the region.

Finally, E-V4257, the sister clade of the main E-M2 multiforcation, shows a scattered distribution from lake Chad to

the western coast (figure 20, panel B). This clade differentiated between 11.03 ± 0.79 and 6.46 kya, during the Green Sahara period. The present-day distribution possibly represents the relic of a more ancient and widespread presence throughout the central Sahara.

The geographic distribution and dating of E-Z15941, E-L516 and E-V5001 seems to suggest an extensive occupation of the Saharan/Sahelian region from lake Mega-Chad and northward during the Green Sahara in a time window from ~ 10 kya to ~ 4 kya, followed by a massive movement southward and, to lesser extent, northward due to the advancing desertification.

R-V88 during the Green Sahara

R-V88 shows its highest frequencies in the Saharan/Sahelian region. Taking into account the Eurasian distribution of other R lineages and considering that the R-V88 basal clades (R-M18 and R-V35, figure 18) are mainly found in Sardinia, the R-V88 homeland was possibly in Europe. The movement toward the central Sahel, possibly including an intermediate passage within an unknown northern African region, occurred between 8.67 and 7.85 ± 0.90 kya, considering the frequency distribution of the R-V4963 sub-clade (figures 18 and 21, panel B). In the central Sahel, the star like topology of the internal R-V1944 clade indicates a strong demographic expansion experienced by this clade just after 5.73 ± 0.49 kya (figure 18). R-V1944 is evenly distributed along a trans-Saharan axis from lake Chad towards the Siwa oasis, although the highest frequencies are found in northern Cameroon (figure 21, panel C).

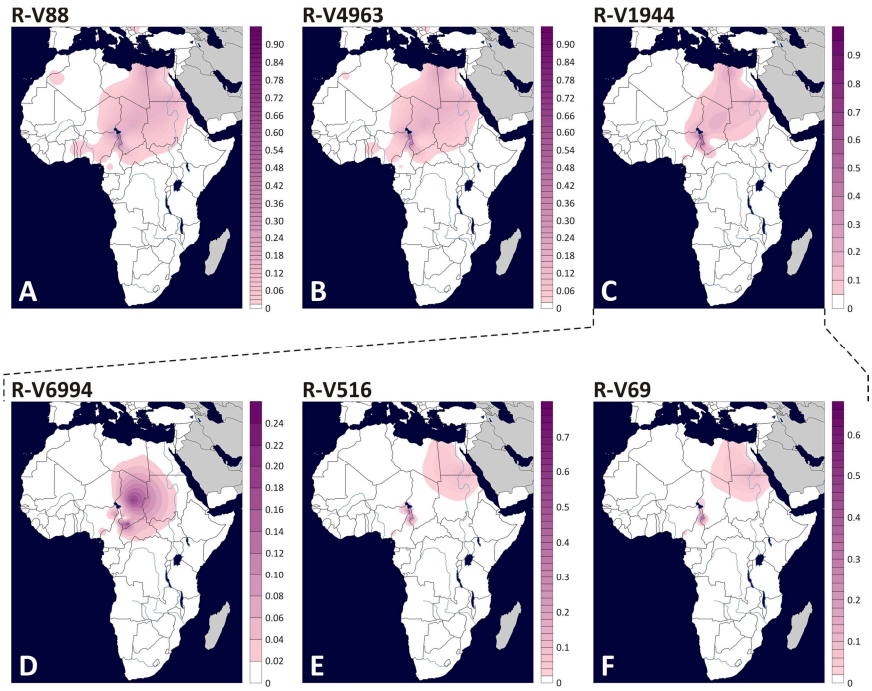


Figure 21: Frequency maps of R-V88 and its internal clades. Each panel represents the geographic distribution of a haplogroup, indicated above. The countries not investigated in the present thesis are in grey. Please note that each lineage is represented with its own scale adjusted according to its frequency distribution (to the right of each panel). When the panels represent the internal clades of another panel, they are enclosed by dashed lines.

Due to its star-like topology, R-V1944 lacks a clear geographic differentiation and it is difficult to understand the direction of its expansion, which, however, seems to have started somewhere between the central Sahel and northeastern Africa. In any case, we identified a north African clade, R-V4759, which arrived in the current area between 5.73 ± 0.49 and 4.69 ± 0.98 kya. Taking into account all these results, it is possible to conceive that the expansion started in the area of lake Mega-Chad, where the highest frequencies are to be found (Appendix 2). In this scenario, the expansion moved northward as far as northeastern Africa, where it

arrived in a time consistent with the coalescence age of the northern R-V4759 clade.

Finally, in literature, R-V88 has been linked to the Chadic arrival in the lake Chad basin (Cruciani et al. 2010; González et al. 2014). Our data seem to confirm the link between R-V88 and the Chadic language (Mann-Whitney test between R-V88 in the Chadic people *vs.* R-V88 in non-Chadic people: $p = 7 \times 10^{-4}$). More specifically, it seems that the internal lineage, R-V4963, was implicated in the arrival of the proto-Chadic language in the lake Chad basin between 7.85 ± 0.90 and 5.73 ± 0.49 kya, along a route from north to south coherent with the "trans-Saharan" hypothesis proposed by Ehret (1995). The alternative hypothesis of a trans-Saharan route in the opposite direction (Gonzalez et al. 2013) was proposed considering the R-V88 chromosomes found southward of the common area of this haplogroup. In fact, Gonzalez et al. (2013) reported 9 R-V88* (xM18, V8, V7, V69) chromosomes in Equatorial Guinea and Berniell-Lee et al. (2009) found, in Gabon, 46 R-P25* (xM18, M73, M269) subjects which are possibly R-V88. Furthermore, R-P25* chromosomes that probably harbour the V88 polymorphism have also been found in Sudan, reaching frequencies of about 40% among the Hausa (Hassan et al. 2008). Recently, R-V88 chromosomes have also been reported at low frequencies in the Saharawi and Mauritanians (Bekada et al. 2013). In the scenario proposed in the present thesis, the presence of R-V88 in these areas can be explained as the furthest offshoots of the R-V1944 expansion, with Equatorial Guinea and Gabon to the south, Mauritania to the west and Sudan to the east. The putative high frequencies of R-V88 among Hausa in Sudan, which arrived from Niger and Nigeria in recent times, seem to support our hypothesis. Finally, R-P25 subjects (possibly R-V88) were also reported in the Fulbe people from the central Sahel (Niger, Cameroon and Chad; Bučková et al. 2013), in sharp contrast with the paucity of R-V88 among the Fulbe people reported in this thesis. Given that the region is the same, the presence of R-V88 among Fulbe is possibly the result of contacts with the local

Chadic people. In support of this scenario, Bučková et al. (2013) found no trace of R-V88 among the Fulbe people from regions outside the R-V88 area, such as Mali and Burkina Faso. Only a high-resolution typing of the SNPs reported here can address the issue regarding the arrival of R-V88 proto-Chadic people in the central Sahel, the direction of their following expansion and the extent of any influence on the Fulbe people.

In conclusion, R-V88 and, more specifically, its sub-clade, R-V1944, seems to have a geographic distribution linked to the fertile environment during the Green Sahara period.

The eastern Sahara

The eastern part of the desert is thought to have experienced a more rapid desertification (Linstädter and Kröpelin 2004; Tierney and deMenocal 2013; Blanchet et al. 2014), as testified by the limited quantity of archaeological evidence after ~ 5 kya (Kuper and Kröpelin 2006).

In this region, we mainly found lineages belonging to E-M78. This haplogroup probably originated in northeastern Africa with a later dispersion within and outside Africa (Cruciani et al. 2007; Trombetta et al. 2015a). In the present study, we found three E-M78 sub-clades with a sharp internal differentiation: E-V22, E-V12 and E-V264 (figures 16 and 22, panels B, C and D).

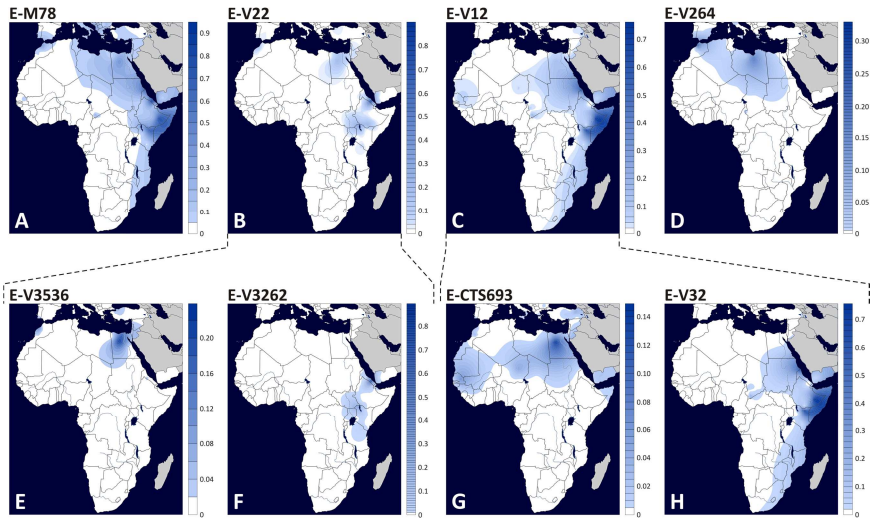


Figure 22: Frequency maps of E-M78 and its internal clades. Each panel represents the geographic distribution of a haplogroup, indicated above. The countries not investigated in the present thesis are in grey. Please note that each lineage is represented with its own scale adjusted according to its frequency distribution (to the right of each panel). When the panels represent the internal clades of another panel, they are enclosed by dashed lines.

The eastern corridor

The E-V22 haplogroup was found to be subdivided into two main clades: E-V3536 in the Mediterranean region (northern Africa, Near East and southern Europe) and E-V3262 in the Horn of Africa. Parallel to these two sub-haplogroups, we also found 23 Y chromosomes still belonging to the paragroup E-V22*, which only shows a minor eastern African component (Appendix 2). Although we were unable to make a precise estimate of the lower limit of the differentiation in the Mediterranean basin, the movement towards the Horn of Africa occurred between 11.83 (the coalescence age of E-V22) and 5.71 ± 1.12 kya (the coalescence age of E-V3262) (figure 16).

Interestingly, the same general pattern was also a feature of E-V12. This haplogroup has two lineages: the Mediterranean E-CTS693 and E-V32 in the Horn of Africa (figure 22, panels C, G

and H). The diversification in the Mediterranean area dates back to a time window of 10.01 ± 1.17 - 8.98 ± 1.12 kya, while the arrival in the Horn of Africa occurred between 10.01 ± 1.17 and 5.99 ± 0.72 kya (figure 16). In this context, it is worth noting that although E-V32 is mainly found in the Horn of Africa, it also includes a rare sub-clade (E-V32/V6873) which is restricted to the central Sahel. This finding can be interpreted as a result of a movement along the Sahelian belt, which will be described in-depth in a later section.

In conclusion, E-V22 and E-V12 seem to have experienced the same evolutionary processes.

Both E-V22/V3536 and E-V12/CTS693 contain samples from northeastern Africa, Near East and southern Europe, with no clear differentiation among these three regions. The diversification of these lineages occurred between ~ 11 and ~ 8 kya, possibly thanks to improvements in climate during the last African humid period.

Parallel to the Mediterranean clades, both lineages have an eastern African clade (E-V22/V3262 and E-V12/V32), which moved from northeastern Africa and which arrived in the Horn of Africa at about the same time, between ~ 11 and ~ 6 kya. After their arrival, both lineages went through an expansion and both are essentially restricted to this geographic region, with very few influences from other regions.

The distribution of E-V22 and E-V12 indicates movements on a northeast-to-east axis, along two possible main routes: 1) the Nile valley; 2) across the eastern Sahara. The corridor along the Nile was the route usually proposed to explain the distribution of these or other haplogroups (Cruciani et al. 2007). However, in the Nile valley, there is a general paucity of archaeological evidence dating back to the Green Sahara period, which is not surprising considering that this area was probably marshy and inhospitable at that time (Kuper and Kröpelin 2006). On the contrary, a large amount of archaeological sites have been found in the present-day Great Sand Sea between Libya and Egypt, which was transformed

into a grassland following the onset of the humid conditions (Kuper and Kröpelin 2006). The presence of both E-V12 and E-V22 in northern and western Sudan seems to indicate the eastern-Saharan route as the most plausible, although the dissection of these two clades is at low-resolution (Hassan et al. 2008).

Finally, it is worth noting that the onset of the arid conditions in the eastern Sahara and the consequential changes in food production probably contributed to the increasing social complexity and stratification at the basis of the formation of the Egyptian Dynastic civilisation (Brooks 2006).

The western corridor

Differently from E-V12 and E-V22, E-V264 shows a different geographic distribution (Appendix 2 and figure 22, panel D). It is mainly represented by the northern African E-V65, but it also harbours a central Sahelian clade (E-V259) found in two males from two different ethnic groups in Cameroon (1 Daba and 1 Guidar). The separation of E-V65 and E-V259 occurred about 12 kya, a period coherent with the beginning of wetter conditions in the eastern Sahara. Considering its distribution, E-V259 could represent the footprint of a trans-Saharan corridor from northeastern Africa towards the central Sahel. Interestingly, about 6% of E-M78 has been recently reported in the central Sahel (Bučková et al. 2013). Unfortunately, these chromosomes were not typed for the main E-M78 downstream lineages. On the basis of the Y-STRs, we can exclude the E-V32 affiliation because of the lack of the 11-repeat motif usually associated with this clade (Cruciani et al. 2006, and unpublished data). However, their microsatellite profiles are highly diversified, so it is not possible to make any haplogroup inferences. Considering the distribution of E-M78 in the central Sahel reported in the present study, these chromosomes could cluster with the E-V259 lineage (Appendix 2). Were this to be the case, the wider distribution of E-V259 would

support our hypothesis of a westward corridor across the Sahara during the last African humid period.

General overview of the Sahara

Since the peopling of the Green Sahara was a complex event with analogies and differences in different regions, in this section, we summarize the key findings of the present thesis, trying to paint a comprehensive picture of the events during the last African humid period.

1) First occupation of the Green Sahara

On the basis of their topology and their geographic distribution, A3-M13 and E-M2 seem to have arrived in the central Sahara from the southern Sahelian region, following the northward spread of the savannah environment. On the contrary, E-M78 and R-V88 seem to have followed the opposite route, arriving in the Sahara from northern Africa. These data fit with the two-way occupation of the Green Sahara proposed by Drake et al. (2011). However, the archaeological and paleoanthropological evidence shows great homogeneity in the material culture of the early Holocene Saharan populations. These findings suggest a rapid peopling of the Sahara and extensive contacts between different groups, facilitated by the network of lakes and rivers (Drake et al. 2011; Manning and Timpson 2014).

2) Expansions within the Green Sahara

We found evidence of a first expansion in the Sahara about 7 kya, when we observed the differentiation of the majority of the trans-saharan sub-lineages belonging to the A3-M13, E-M2 and R-V88 haplogroups. This finding can be interpreted both as a

consequence of a climatic event, and as the result of socio-economic improvements.

During the Green Sahara, a short period of aridity occurred around 8 kya (Brooks 2006; Manning and Timpson 2014). Interestingly, the diffusion of the lineages occurred after this arid parenthesis. In the same period, the Saharan populations abandoned hunting and gathering and adopted pastoralism, probably as an adaptive strategy against the arid conditions (Brooks 2006; Kuper and Kröpelin 2006; Manning and Timpson 2014). The exploitation of pastoral resources, along with the recovery of the wetter conditions, could have triggered the simultaneous population expansion testified by the increased number of contemporaneous archaeological sites throughout the Sahara (Manning and Timpson 2014).

Moreover, we found evidence of a second strong demographic expansion which occurred about 5.7 kya, involving the R-V88/V1944 haplogroup. It is difficult to state what event promoted this demographic explosion. One possible cause could have been the arrival of pastoralism as far as lake Mega-Chad, where probably R-V1944 arose and started its expansion. However, it is worth noting that at the same time, a general cultural shift was observed in other Saharan and sub-Saharan regions, possibly linked to the progression of desertification and the resulting economic, demographic and social changes (Brooks 2006).

3) Regional differences at the end of the Green Sahara

The humid period did not end at the same time throughout the whole Sahara. A consequence of this is that central and eastern Sahara show a different haplogroup distribution and different time estimates for the separation between northern and sub-Saharan Africa.

In particular, the eastern Sahara experienced a more rapid onset of arid conditions. Consistently, the E-M78 sub-clades distributed

in this area show a more ancient and evident differentiation between populations from the regions on the Sahara borders. On the contrary, in central Sahara, the shift from the Green Sahara to desert was gradual and was not complete until 5 - 4.5 kya. This delay is mirrored by the A3-M13, E-M2 and R-V88 distribution in this area, which in general shows a more recent and less sharp separation.

A completely different scenario has been proposed for the western littoral, which seems not to have been influenced by the Green Sahara dynamics, a fact deduced on the basis of the uninterrupted archaeological sites (Manning and Timpson 2014). Therefore, the coastal region could have allowed some contacts between western and northwestern Africa.

Beyond the Green Sahara: other movements within and outside the African continent

The aim of this thesis was to investigate the past human movements across the Sahara linked to the last African humid period. However, we also found evidence of other events which occurred outside the Green Sahara region.

The Mediterranean basin

Sardinia

In the island of Sardinia, all the European lineages can usually be found, although the most frequent Y haplogroup is I-M26, which is rare elsewhere. In addition, some African sub-haplogroups, such as A3-M13, E-M44 and R-V88, are found at low frequencies. The presence of the African lineages is usually

put down to recent events that occurred in historical times, such as the Roman and Vandalic dominations (Francalacci et al. 2013).

Consistently with published data, we found few Sardinian samples from our lab collection that belong to A3-M13 and R-V88 (Appendix 2). After the inclusion of the data from Francalacci et al. (2015), we were able to estimate when these lineages arrived in Sardinia (Figures 12 and 18). Interestingly, the Sardinian samples belonging to both A3-M13 and R-V88 clustered in internal lineages which originated in not recent times. However, the two haplogroups seem to have experienced a different evolution.

We analysed the A3-M13 Sardinian sample S110 (branch 22 in figure 11) through NGS and we found that it split early from all the other subjects, after 7.62 ± 0.92 kya. We selected and genotyped 6 SNPs of branch 22 in more than 7000 males, considering also the *in silico* analysis of the subjects from Francalacci et al. (2015) and Poznik et al. (2016) (table 6 and Appendix 2). We found that the Sardinian A3-M13 chromosomes are all included in the A3-V3663 sub-clade. More specifically, we found a lineage, namely A3-V3663/V2100, which is restricted to Sardinia and which includes all the Sardinian A3-M13 subjects, except one sample that forms the paragroup A3-V3663*.

The structure of the A3-M13 tree seems to suggest that the direction of the movement was from Africa to Sardinia. The route could have been both trans-Mediterranean, or through the Near East. In fact, the two paragroups upstream of A3-V2100 harboured respectively an Egyptian subject (A3-V2406*) and one Greek and one Yemenite sample (A3-V2742*).

We did not analyse all the SNPs in branch 22, so we are not able to precisely estimate the date of the arrival of the A3-V3663 sub-clades in the Mediterranean area and, more specifically, in Sardinia. However, the presence of A3-V3663 basal clades in southern Europe seems to indicate that the spread of A3-M13 sub-haplogroups in the Mediterranean basin was not due to recent historical events.

Differently from A3-M13, the R-V88 Sardinian clades form a nested topology at the base of this haplogroup. R-V88 arose within the R-P25 haplogroup, which also harbours R-V1274 found in Yemen and R-P297, which in turn includes the European R-M269. Moreover, the sister clade of R-P25 is R-SRY10831.2, which is present in Europe and Asia (Karafet et al. 2008; Trombetta et al. 2015b). This geographic distribution, along with the presence of two basal R-V88 lineages in southern Europe, seems to indicate that R-V88 originated somewhere in Europe between 18.68 ± 1.85 and 12.34 kya (table 5 and figure 18).

On the basis of these findings, the direction of human movements was probably opposite to the one suggested by the A3-M13. In fact, it seems plausible that R-V88 chromosomes moved from Europe to central Sahel during a time window between 8.67 and 7.85 ± 0.90 kya. The route toward the central African region probably passed through northeastern Africa rather than through Arabia and the Bab al Mandab strait, considering the absence of R-V88 in the Horn of Africa. The arrival in northern Africa could have come from the Near East or directly from southern Europe, crossing the Mediterranean Sea. Although one M18 chromosome was found in Lebanon (Zalloua et al. 2008), the absence of R-V88 chromosomes in the Balkans and Near East seems to suggest a trans-Mediterranean route for this clade.

In conclusion, the presence of A3-M13 and R-V88 in Sardinia and in the Mediterranean area is probably due to ancient events. The topology of A3-M13 probably indicates a movement outside Africa towards southern Europe. On the contrary, the R-V88 structure suggests a back-to-Africa migration. Probably, both lineages were widespread in southern Europe, where they have been replaced by the Y lineages brought by the recurrent migration waves from central Asia, which occurred in the following centuries.

The coastal region of northern Africa, the Near East and southern Europe.

The analysis of the E-M78 haplogroup revealed an interesting link between northern Africa, the Near East and southern Europe. In fact, we found that the chromosomes from these three areas clustered together after the Green Sahara period (figure 16).

We did not analyse all the SNPs included in the Mediterranean branches (E-V22/V3536, E-V12/CTS693 and E-V65), so we cannot absolutely exclude the presence of region-specific clades. However, the similar distribution of other Y haplogroups (for example, E-V257, E-M123, J-M267, J-M172, G-M201 and T-M70) and of mtDNA lineages in the Mediterranean area seems to suggest extensive contacts between northern Africa, the Near East and southern Europe (Bosch et al. 2001; Arredi et al. 2004; Cruciani et al. 2004; Semino et al. 2004; Francalacci et al. 2013; Fernández et al. 2014; Secher et al. 2014; Batini et al. 2015). Interestingly, concordant results have been obtained from genome-wide studies of modern or ancient humans, which found signals of near eastern/southern European admixture in northern Africa (Henn et al. 2012; Fadhlaoui-Zid et al. 2013; Busby et al. 2016; Fu et al. 2016; Lazaridis et al. 2016; Omrak et al. 2016).

In conclusion, this bulk of data indicate that the Mediterranean regions have been ceaselessly and extensively in contact over the last thousands of years. More specifically, after the end of the African humid period, northern Africa was influenced by the middle eastern/southern European events and *vice versa*. A consequence of this is that the northern African populations became more and more genetically differentiated from the sub-Saharan ethnic groups.

The Sahel

The Sahelian belt spans from the Atlantic Ocean to the Red Sea, immediately south of the Sahara. Its climate and ecology is intermediate between desert, typical of present-day Sahara, and tropical savanna, typical of southern regions. In this area, several dialects belonging to three of the four African linguistic families, namely Afro-asiatic, Nilo-Saharan and Niger-Congo, are spoken, confirming that the Sahel has been and still is an important crossroad in the African continent. In the present thesis, we found signals of at least two events that occurred along the Sahelian belt: the Fulbe movements and extensive contacts between central and eastern Africa. Because of the geographic position of the Sahel and considering the distribution and the time estimates of the haplogroups involved, it is possible that these events were directly or indirectly caused by the climatic changes of the last African humid period.

The Fulbe people

The Fulbe probably descend from ancient nomadic herders, although the majority of them has now abandoned pastoralism becoming sedentary farmers (Boulet et al. 1984). Their origins are still a matter of debate and have been traced back to different African or non-African people, such as western Africans, ancient Ethiopians, Nubians, Persians, Jews or Arabs (Hopen 1958; Murdock 1959; Cornevin 1962; Seligman 1966). In the XI century, they were settled in the mountains of Futa Djalon in Guinea and, in later centuries, they moved eastward along the Niger river and as far as lake Chad, where their first traces date back to the XV century (David 1971; Gauthier 1979; Newman 1995). Genetically, they seem to have admixed with the local populations during their spread from western Africa to the central Sahel (Tishkoff et al. 2009).

In this thesis, we found a haplogroup that is particularly frequent among Fulbe, namely E-M2/Z15941 (figure 14). As described in the previous section dedicated to the Green Sahara, the first dispersal of this clade seems to have occurred in a time window coherent with the last African humid period. However, its current ethno-geographic distribution perfectly traces the Fulbe migration from western Africa to the central Sahel.

The highest frequencies were reported among the western populations, in particular the Mandenka from Senegal and Gambia (~ 30 - 40%) and among Fulbe groups from Nigeria (> 50%) and from Cameroon (~ 14%). Its absence in Fulbe from Niger is probably due to the low sample size, since we have only 7 subjects belonging to this ethnic group. In other populations, E-Z15941 was absent, with the exception of a few Berbers (Appendix 2).

In conclusion, E-Z15941 followed a dual evolution. It firstly dispersed across the Green Sahara thanks to the improved climatic conditions, as testified by its probable ancient presence in the Egyptian Siwa oasis. On the other hand, its present-day diffusion in the Sahelian belt was heavily due to the recent movements of Fulbe people. The analysis of the markers in the branches downstream of E-Z15941, in particular the polymorphisms shared with the Gambian population of the 1000 Genomes Project (Appendix 1), can shed light on the time and trajectories of the first dispersal of this haplogroup. Furthermore, important information can emerge regarding the role of the Fulbe in this process and, possibly, about the ancient history of this population.

Links between eastern and central Africa

We found clear signals of connections between central and eastern Sahel on the basis of the distribution of two internal clades: A3-M13/V4735 and E-M78/V32 (figures 12 and 16).

A3-M13/V4735 originated between 10.24 ± 0.96 and 6.02 ± 0.98 somewhere in the central Sahel. Between 6.02 ± 0.98 and 5.30

± 1.01 kya, we observed a shift in the geographic distribution from the central Sahel to eastern Africa. In about the same period, between 5.99 ± 0.72 and 5.17 ± 1.09 kya, we found evidence of a movement along the same axis involving the internal lineages of E-M78/V32. In fact, in this time window, a clade restricted to the central Sahel (E-V32/V6873) split from the eastern E-V32 subjects, suggesting a migration from east to west, opposite to the putative movement of A3-M13/V4735.

Interestingly, both these lineages seem to be linked to the spread of the Nilo-Saharan languages, although in different ways.

The Nilo-Saharan is one of the four main linguistic families in Africa and it is spoken in an area ranging from the central Sahel to the Great Lakes region in eastern Africa, with some Nilo-Saharan groups as far as western Africa (such as the Songhai from Mali) (Tishkoff et al. 2009). It has been proposed that the Nilo-Saharan language originated in the eastern part of Sudan between ~ 15 and ~ 13 kya. In later centuries, bi-directional movements along the Sahel led to the diffusion of the Nilo-Saharan language from lake Chad to southern Sudan (Bender 1997; Tishkoff et al. 2009; Gomes et al. 2010). Several linguistic and genetic data suggest extensive contacts in the Ethiopian highlands between Nilo-Saharan people and Afroasiatic Cushitic groups, which started at least 20 kya and which intensified after 5 kya (Tishkoff et al. 2009; Blench 2006; Campbell et al. 2014; Shriner et al. 2016). The arrival of Afroasiatic proto-Chadic speakers from central Sahara in the lake Chad area about 7 kya probably caused a linguistic shift in the local Nilo-Saharan populations (Tishkoff et al. 2009).

Finally, it has been proposed that the spread of the Nilo-Saharan phylum was related to the Aqualithic culture in the early Holocene. This material culture consists of barbed points, which were used to fish and hunt aquatic species. It is worth noting that there is a high correspondence among the Nilo-Saharan distribution, the Aqualithic sites and the Holocenic distribution of aquatic species. On these bases, it has been hypothesized that the expansion of the

aquatic resources in the Green Sahara promoted the occupation by fishing people, who moved from east to west and northward. In this way, the Nilo-Saharan language could have arrived as far north as northern Africa, where isolated Nilo-Saharan populations are found, suggesting an ancient broader distribution than today (Drake et al. 2011).

A3-M13/V4735 was found to be significantly related to the Nilo-Saharan speaking groups (Mann-Whitney test between Nilo-Saharan *vs.* others: $p = 0.0016$), suggesting that the diffusion of this lineage is a consequence of the movements of Nilo-Saharan speakers along the Sahel from lake Chad eastward. In literature, the link between A3-M13 and the Nilo-Saharan speaking groups has already been proposed (Gomes et al. 2010; Batini et al. 2011). Here, we found a specific internal A3-M13 clade involved in the Nilo-Saharan diffusion, between ~ 6 and ~ 5 kya.

The link between E-M78/V32 and the Nilo-Saharan is less sharp. In fact, we did not observe any significant correlation considering the whole E-V32 (Mann-Whitney test between Nilo-Saharan *vs.* others: $p = 0.18$). However, the Nilo-Saharan language significantly correlates with the internal central Sahelian lineage E-V32/V6873, which is almost exclusively found among the Nilo-Saharans (Mann-Whitney test between Nilo-Saharan *vs.* others: $p = 0.02$). These findings, along with the concomitant time window, suggest that A3-V4735 and E-V6873 could have been involved in the same bi-directional movement between lake Chad and eastern Africa.

Finally, the eastern sister clades of E-V6873 are quite common among the Cushitic populations from the Horn of Africa (Appendix 2), confirming the previous findings of extensive contacts between Nilo-Saharans and Cushitic groups.

Sub-Saharan Africa

The Horn of Africa

The Horn of Africa includes Eritrea, Ethiopia, Djibouti and Somalia. This region shows considerable genetic and linguistic differentiation, with high levels of Eurasiatic ancestry from different sources, dating back to recent and ancient times (Hodgson et al. 2014; Gallego Llorente et al. 2015; Busby et al. 2016; Gandini et al. 2016).

In the present thesis, we found three lineages which were mainly distributed in the Horn of Africa.

A3-M13/V1018 split from the first multiforcation within A3-M13 and arrived in the Horn of Africa between 10.75 ± 0.91 and 8.43 ± 1.21 kya, where it is now confined (figure 12). In this region, this lineage differentiated into two sub-clades, A3-V317 and A3-V3, which in turn underwent a strong expansion between 8.43 ± 1.21 and 5.48 ± 1.07 kya.

E-V22/V3262 arrived in the Horn of Africa between 11.83 and 5.71 ± 1.12 kya and reaches its highest frequencies among the Saho people (Appendix 2).

The other frequent clade in the Horn of Africa is E-V32, which includes subjects mainly from this region with the exclusion of the rare central Sahelian E-V32/V6873 (figure 16). E-V32 variation in the Horn is subdivided into two sub-clades (E-V32/V3746 and E-V32/V4381) and one paragroup (E-V32*). Differently from A3-V1018 and E-V3262, the E-V32 lineages also harbour samples from Kenya, where they are usually found in non-Bantu people, with very few exceptions (Appendix 2). However, a separation between Kenya and the Horn of Africa emerged from the NGS of subjects from these two regions both in branch 132 (corresponding to E-V3746) and in branch 138 (corresponding to E-V4381) (figure 15). The genotyping analysis of all seven SNPs of branch 138

confirmed this pattern. Further analysis on the E-V32* and E-V3746 samples could lead to a complete differentiation between the Horn of Africa and the other eastern African regions. Taking into account all these findings, the arrival in eastern Africa dated back to a time window between 5.99 ± 0.72 and ~ 5.5 kya, with a subsequent separation between Kenya and the Horn of Africa which occurred before ~ 4.5 kya.

Interestingly, the expansion of the three clades in the Horn of Africa occurred in about the same period, ~ 5 kya, suggesting the same event as a possible cause of their present-day distribution. However, it is difficult to state which specific cultural changes have been involved in this expansion. In fact, pastoralism and the domestication of the first crops was present in the Horn of Africa since ~ 7 kya, while the arrival of Neolithic farmers from the Near East occurred later, ~ 3 kya (Hodgson et al. 2014). However, we found several signals of expansions and migrations that occurred about 5 kya in three haplogroups (A3-M13, E-M78 and R-V88) from different geographic areas, suggesting that the end of the Holocene climatic optimum was characterised by a massive demographic, socio-economic and cultural change which involved all the Sahara and neighbouring regions.

The Bantu expansion

About 5 kya, the Bantu branch of the Niger-Congo linguistic family originated in central Africa, in the southern regions of Nigeria and Cameroon. Thanks to the adoption of agriculture and, later, of iron technologies, these people spread from their homeland toward eastern and southern Africa, leading to the diffusion of agricultural skills and the Bantu language throughout sub-Saharan Africa (Nurse and Philippson 2003). Along with the cultural and technological features, this demic diffusion led to the spread of the Bantu genetic pool. The E-M2 and B-M150 Y haplogroups are usually linked to the Bantu expansion, considering their geographic distribution and their high frequencies among the

Niger-Congo speaking groups (Cruciani et al. 2002; Berniell-Lee et al. 2009; Batini et al. 2011; de Filippo et al. 2011).

The Bantu expansion is not the main focus of the present thesis, although the analysis of the E-M2 haplogroup made it possible for us to study E-M2/U209 and E-M2/M191. These sub-clades harbour a large number of E-M2 chromosomes and both include a frequent lineage, E-U290 and E-U174 respectively, which has been specifically linked to the Bantu expansion (Rosa et al. 2007; Batini et al. 2011; de Filippo et al. 2011; Ansari Pour et al. 2013). From the genotyping analysis, we obtained detailed distribution information for the E-U209, E-M191 and its sub-clade E-U174 (figure 14).

E-U209 and E-U174 show a strong central African component (gold in figure 14), consistently with their link with the Bantu demic diffusion. They are also found in central, southern and eastern Africa, among populations influenced by this expansion (Appendix 2 and figure 20). Interestingly, within E-M191, we found another sub-clade, namely E-8019527, which seems to be involved in the Bantu expansion, as well as its sister clade E-U174. E-8019527 is far less frequent than E-U174 and it has a scattered distribution from the Gulf of Guinea to Morocco. However, both lineages show their highest frequencies among Niger-Congo populations and their time estimates are coherent with the Bantu expansion.

We did not genotype U290 in all the samples, so we were not able to dissect the E-U209 variability in all the populations here analysed (Appendix 2). However, considering only the results of the NGS (figure 13), we dated the split of E-U290 from the other E-U209* (xU290) chromosomes between 6.02 ± 0.86 and 4.90 ± 1.00 kya. The diversification of E-U290 occurred after this date, consistently with a spread linked to the Bantu diffusion. Interestingly, the upper limit of the E-U209 radiation falls within the Green Sahara period, which may have triggered the first dispersal of this clade. The genotyping of further markers within E-

U209 could possibly distinguish between the sub-Saharan Bantu component and a more ancient northern African/Sahelian one.

MATERIALS AND METHODS

The sample

The 104 Y chromosomes analysed by NGS were selected on the basis of their haplogroup affiliation, which had been determined in previous studies (Cruciani et al. 2002, 2004, 2007, 2010, 2011; Scozzari et al. 2012, 2014; Trombetta et al. 2011, 2015a) (table 1). The DNA samples were obtained from peripheral blood, saliva or cultured cells.

Sample quality and quantity control

Target sequencing required specific quality and quantity parameters for the DNA to be analysed:

1) absence or low amount of DNA degradation;
2) quantity ≥ 3 µg;
3) concentration ≥ 37.5 ng/µl;
4) purity: A260/280 = 1.8 - 2.0.

Concentration and purity were measured using a NanoDrop 1000 spectrophotometer, produced by Thermo Fisher Scientific. Degradation was assessed by means of an electrophoretic run on a 1% agarose gel. In this way, we selected 45 Y chromosomes.

However, we also found 59 samples which respected criteria 1, 3 and 4 but in insufficient quantities. Since these subjects were crucial for this thesis due to their haplogroup affiliation and ethno-geographic origin, we decided to perform a Whole Genome Amplification (WGA) on them. This technique uses random hexamers as primers for the amplification of the whole genome. The disadvantage of this approach is that it is possible the different

regions of the genome are not amplified homogeneously. To perform the WGA, we used the GenomiPhi V2 DNA Amplification kit (GE Healthcare) according to the manufacturer's protocol.

Selection of the unique MSY regions

We selected 22 blocks within the X-degenerate portion of the Y chromosome, for a total of about 11 Mb which were characterized by a low degree of homology with the X chromosome or with the autosomes (table 10). The total number of targeted bases was expected to decrease to about 4 Mb after the exclusion of the repetitive elements. For the selection of the unique regions, we used the "Table browser" tool of the UCSC Genome browser, considering the human Feb. 2009 (GRCh37/hg19) assembly.

#	Start position (GRCh37/hg19)	End position (GRCh37/hg19)	Size
1	2649519	2917958	268440
2	6616599	7472224	855626
3	7540720	8751529	1210810
4	8835071	8875351	40281
5	9372551	9466033	93483
6	9640280	9648763	8484
7	9757200	10104553	347354
8	13193953	13239280	45328
9	13870436	16093531	2223096
10	16172354	17986737	1814384
11	18016823	18271431	254609
12	18537676	19567683	1030008
13	20802246	20841912	39667
14	21032069	22216158	1184090
15	22513118	23497661	984544
16	23509282	23514722	5441
17	23591735	23636626	44892
18	23766330	23772950	6621
19	23798490	23802086	3597
20	23897059	23901428	4370
21	24362824	24520948	158125
22	28457991	28819361	361371

Table 10: Human Y chromosome coordinates of the 22 regions selected for the target sequencing.

Targeted Next Generation Sequencing

Once the whole set of 104 DNA samples had been selected, we sent it to the BGI-Tech (Hong Kong). They performed the library preparation, targeting, sequencing and alignment.

Targeting and library preparation

Genomic DNA was sheared by means of a Covaris ultrasonicator in order to obtain DNA fragments of 200-300 bp. The fragments were purified, end-repaired and ligated to paired-end adapters, which had a short subject-specific tag, allowing the simultaneous analysis of multiple samples. The next step involved the selective enrichment of the targeted unique regions of the MSY

(table 10) using a capture array produced by Roche Nimblegen, composed of a set of short probes (about 200 bp in length) that overlapped the selected regions. The probes excluded almost all the repetitive elements from the 22 X-degenerated blocks, capturing a total of about 4.4 Mb.

Sequencing and alignment

The captured regions were loaded onto an Illumina HiSeq 2500 platform to produce a ≥ 50× mean depth for the targeted 4.4 Mb.

The low quality reads, contamination with adapters and repeated reads were discarded from the raw output, which was then sorted thanks to the subject-specific tags. The next step was to align the sequences of each subject to the human Y chromosome reference sequence (GRCh37/hg19) by means of the BWA (Burrows-Wheeler Aligner) software (Li and Durbin 2009). This process generated an alignment file in .bam format, which is the binary version of the .sam (Sequence Alignment/Map) file (Li et al. 2009).

Regional filtering

Analysis of the average depth

In order to obtain reliable data for all 104 subjects, we performed an analysis of the average sequence depth, through the extraction of some informative values from each .bam file using the SAMtools platform (Li et al. 2009, Li 2011). First of all, we extracted the depth value for each base of the captured region, after excluding the low quality reads. For this purpose, we set 30 as the lower threshold for the mapping quality (MQ) of the reads. Due to

the possible imbalance in the amplification of the regions introduced by the WGA method, we analysed the 59 WGA samples separately from the 45 genomic DNA samples. For both datasets, we calculated the moving average of the mean depth per position along the entire captured region, using sliding windows of 1000 bp moving 1 bp. Then, we discarded all the positions with a mean depth value which was in the lower 3‰ range of the average depth distribution in the genomic dataset or in the lower 4‰ range in the WGA dataset. We also discarded two blocks (chrY: 21152803-21154906; chrY: 28793241-28819317) with very high average depth values in the genomic dataset, because these values could be indicative of Y chromosome rearrangements. After the depth analysis, we rejected 29135 bp. We further refined the filtered regions removing the remaining ~ 0.70 Mb within the repetitive elements (using the Repeat Masker and Simple Repeats tracks from the Table browser tool of the UCSC Genome Browser).

We obtained a total of ~ 3.7 Mb which passed our quality and depth controls.

Analysis of putative deletions/duplications

We also used the depth information to identify candidate deletions or duplications within our target regions on the Y chromosome. We extracted the raw depth (without filtering for the MQ) for each of the 3.7 Mb obtained from the previous filtering steps. For each subject, we performed the moving average using sliding windows of 100 bp moving 1 bp. Then, we analysed separately the genomic and WGA samples, choosing for each dataset an arbitrary reference sample with a good coverage (S178 and S179, for the genomic and WGA group, respectively). For both datasets, the average depth value of each 100 bp interval was divided by the corresponding value of the reference sample. The identification of continuous clusters of positions with a ratio value ≤ 0.05 in at least one subject were marked as putative deletions. On

the contrary, putative deletions on the reference sample led to a very high (theoretically infinite) ratio in all the other samples.

The identification of putative duplications with the same method was difficult because of the strong oscillations of the observed depth value. Thus, we applied a double standardization method, dividing the moving average depth values by the total average depth of the same sample and then by the corresponding value of the reference subject. Despite this, the fluctuations were too strong to obtain reliable data in the WGA dataset, so we could only analyse the data from the group of the genomic samples. We identified clusters of positions with a value ratio ≥ 1.8 that is indicative of a duplication. Putative duplications on the reference sample, on the other hand, led to a very low (theoretically zero) ratio in all the samples. All these situations were checked on the .bam files.

Through this approach, we were able to identify ~ 0.36 Mb possibly involved in deletions or duplications, which need to be experimentally validated. However, these regions showed sub-optimal depth parameters, so we decided to discard them.

After all these analyses, we obtained a final set of ~ 3.3 Mb of unique Y chromosome regions.

These 3.3 Mb were also extracted from the Y chromosome alignment files (.bam) of the four ancient subjects (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015).

SNP calling and filtering

SNP calling

The variant positions were identified comparing our 104 sequences to the human Y chromosome reference sequence (Feb.

2009 - GRCh37/hg19 assembly) using the SAMtools platform (Li et al. 2009, Li 2011). The output was in the form of a VCF (Variant Call Format) file for each sample.

The same process was performed for the Y chromosome of the ancient samples (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015). On the contrary, the .bam files of the 42 modern public subjects from Complete Genomics (Drmanac et al. 2010) and Karmin et al. (2015) were not available, so we extracted the variant positions within the final ~ 3.3 Mb directly from their VCF files.

SNP filtering

In order to discard false positive calls, we applied different filtering criteria, which can be grouped into three different categories:

1) direct filtering: we used the information embedded in the VCF file to accept or discard the variant positions;
2) manual filtering: we manually checked the uncertain cases from the previous filtering step in the alignment (.bam) files;
3) cluster filtering: we checked for clusters of SNPs (i.e. groups of two or more SNPs occurring in close proximity and on the same branch of the Y phylogeny) and decided whether to maintain or discard them from the analyses.

Direct filtering

We analysed the variant positions identified in the three different datasets used in this thesis separately: our 104 sequenced subjects, the 42 publicly available complete Y chromosomes (Drmanac et al. 2010; Karmin et al. 2015) and the 4 ancient specimens (Fu et al. 2014; Lazaridis et al. 2014; Jones et al. 2015).

For our subjects, the VCF parameters considered were the quality ("QUAL" field), the depth ("DP" field) and the number of reads with the reference or the alternative base ("DP4" values within the "FORMAT" field). Using the information in the DP4, we calculated a new parameter used for a more accurate filtering:

$$FilDP4 = \frac{Number\ of\ reads\ with\ the\ ALT\ base - Number\ of\ reads\ with\ the\ REF\ base}{Total\ number\ of\ reads}$$

where FilDP4 was an abbreviation for "filter based on DP4", ALT for alternative base and REF for reference base.

We directly discarded variant positions with FilDP4 ≤ 0.3 and retained all the SNPs with FilDP4 > 0.8 and QUAL ≥ 100. In the other cases, we considered the known phylogenetic context. For SNPs shared among samples belonging to the same haplogroup, we applied less severe criteria (FilDP4 ≥ 0.6 and the number of ALT reads ≥ 2), while we discarded the private positions with DP < 2 or FilDP4 and DP less than 0.4 and 4, respectively. The remaining cases were manually checked in the alignment file.

Because we did not have the alignment files of the Y chromosomes from Complete Genomics (Drmanac et al. 2010) and from Karmin et al. (2015), we could not check the uncertain positions, so we decided to apply more restrictive criteria for the filtering to avoid false positives. Their VCF files had a different structure, so we used different parameters for the filtering, namely FT and AD. The former indicates if the position passed ("PASS") or not ("VQLOW") all the filtering criteria used for the SNP calling, while the latter reports the number of reads with the alternative base. If FT = VQLOW or AD ≤ 2, the presence of the same variant position in other phylogenetically related samples was checked: if present, the SNP was retained, otherwise it was discarded.

For all the other possible values of FT and AD, the SNPs were considered reliable.

Finally, the four ancient Y chromosomes came from different studies, so they showed different features. For this reason, they were analysed separately. The Ust'-Ishim (Fu et al. 2014) and the Loschbour (Lazaridis et al. 2014) samples had a quite high quality/depth, so the same filtering criteria applied for our 104 samples were used (see above).

On the contrary, Kotias and Bichon (Jones et al. 2015) were found to be more critical, so we defined specific filtering criteria. If FilDP4 $\leq$ 0 or DP < 2, the position was discarded, while it was accepted if ALT $\geq$ 2 and FilDP4 > 0.4. For the intermediate cases, the presence of the same variant position in other subjects was checked. Private positions were discarded if ALT < 2 or FILDP4 < 0.4 and DP < 4. The other cases were manually checked in the alignment files.

Manual filtering

All the unresolved cases from the direct filtering were checked manually in the alignment .bam file of the samples of interest. In the final decision, we considered several criteria such as phylogenetic context, the depth and the quality of the region in all the subjects, the proximity of repetitive elements or short indels, the presence of the same variant position with suboptimal parameters in other subjects and the mapping quality of the reads.

Cluster filtering

We checked for the presence of SNP clusters (with each cluster made up of 2 or more mutations), occurring in the same phylogenetic context and at closely spaced positions (less than 20 bp), in order to discard the groups of SNPs that could be generated by the same mutational event (for example, as a

consequence of a gene conversion event). We analysed all the called variant positions, paying more attention to the polymorphisms accepted by the direct or manual filtering steps. If the cluster was composed of 3 or more SNPs or of 2 polymorphisms separated by only one or two bp, the involved variant positions were discarded.

At the end of the filtering process, 5966 SNPs passed all the criteria in our 104 subjects. The number increased to 7544 SNPs in the whole set of 150 samples.

Tree reconstruction and validation

Reconstruction of the phylogenetic relations

The phylogenetic tree was reconstructed using the MEGA software (Tamura et al. 2011). The input file consisted in a table reporting the base for each of the 7544 variant positions (columns) in all the 150 samples (rows). This file was converted into a .meg file and was loaded onto the phylogenetic software, which produced a maximum parsimony tree. Because we did not assign univocally to A00 or A0-T the mutational events on branch 1 (see Results), the tree root was positioned by default to the midpoint. The same table was converted into an .rdf file to be loaded onto the Network software (Bandelt et al. 1999), which produced a median joining network used to obtain rho calculations, the list of variants for each branch and the positions of recurrent mutations. In this way, we identified 25 recurring mutations and 11 triallelic variants (Appendix 1 and Results), which were accurately checked in the alignment files.

Check of published data

The presence in our list of already identified variants in published papers (Hallast et al. 2014; Batini et al. 2015; Francalacci et al. 2015; ISOGG, date: 27 October 2015; Karmin et al. 2015; Trombetta et al. 2015a,b; Poznik et al. 2016) made it possible for us to check the efficiency of all the steps from the SNP calling to the tree reconstruction. Our data successfully passed all these control levels and no false negatives were found.

Mutation rate and dating

Mutation rate estimate

The estimate of the mutation rate was performed using the BEAST software (Drummond and Rambaut 2007). The input file was a NEXUS (.nex) file, containing a list of the genotype at the 7544 variable positions for all the 150 subjects and the structure of the maximum parsimony tree in the newick format. This file was loaded onto the BEAUTY suite and we assigned to the four ancient samples the calibrated radiocarbon dates, expressed in years before present:

- Loschbour (Lazaridis et al. 2014): 8055 years before present;
- Kotias (Jones et al. 2015): 9712 years before present;
- Bichon (Jones et al. 2015): 13,665 years before present;
- Ust'-Ishim (Fu et al. 2015): 44,890 years before present.

The other parameters were set as in Trombetta et al. (2015b). We used a GTR nucleotide substitution model under a strict clock and an expansion growth model for the population size. Then we set the priors:

- uniform clock rate;
- a lognormal current population size, with mean = 10 and standard deviation = 3;
- a uniform population growth rate, with the upper boundary of the human per-capita growth rate set to 4% (Hamilton et al. 2009);
- an exponential ancestral/current population size ratio, with mean = 0.2.

We used the default operator values, except for the weight of the expansion growth rate, which was set to 10. We used runs of 1 million steps, sampled every 10 thousand steps. The first 20% of each run was discarded as burn-in. The output was checked with the Tree Annotator and Tracer platforms.

The mutation rate for the 3.3 Mb analysed in this thesis was $0.735 \pm 0.03 \times 10^{-9}$ /site/year, corresponding to about 1 new mutational event every 408 years.

Time estimates

We applied different methods to estimate the age of the nodes of the tree on the basis of the available information for each node.

Nodes from NGS data

After the reconstruction of the phylogenetic relations among the 150 samples here analysed, we knew the precise number of SNPs downstream of each node. Therefore, the age of the nodes could be estimated using the ρ statistics, which is linearly related to time and mutation rate ($\rho = \mu \times t$). The ρ parameter is an estimation of the variability accumulated, measured as the average number of SNPs downstream of the node to be dated. For each node, the ρ statistic, its associated standard deviation and the corresponding

values expressed in years were calculated using the Network software (Bandelt et al. 1999).

Nodes from genotyping

By means of genotyping analysis, we identified several new lineages (represented by one or more samples not analysed by NGS) which defined new internal nodes. Since we lacked the information regarding the number of SNPs within the lineages identified from the genotyping, it was not possible to use the ρ method to date the new internal nodes. In these cases, we used two different methods:

- Let *n* be the new node identified, if we exactly know the number of markers in the branch upstream *n*, the number of mutations in the NGS branch downstream *n* and if the node downstream (*r*) was precisely dated by means of the ρ statistics, we simply counted the number of SNPs from *n* to *r* and multiplied the total by 408 years. We then added this time to the age of *r* to obtain the time estimate of *n*. The complete information about the number of SNPs downstream and upstream *n* was obtained by the genotyping of all the markers defining the involved branch or from literature (Francalacci et al. 2015).
- If *n* was also found in the phylogeny produced by the Phase 3 of the 1000 Genomes Project (Poznik et al. 2016) and there were no precisely dated downstream nodes to be used as a reference, we estimated the age of *n* by measuring its height, with the traversing approach used in Poznik et al. (2016). We measured the distances (*d*), expressed in number of SNPs, among three nodes: the node to be dated (*n*), a reference node precisely dated with the ρ method (*r*) and an ancestor node between them (*a*).

$$d_{rn} = d_{ra} - d_{an}$$

We then multiplied this value by 408 years and this time was summed to the age of the reference node to obtain the age of *n*.

Genotyping of informative markers

Selection of markers

We selected a total of 108 polymorphisms to be genotyped in the whole set of about 5500 men from the 124 populations of our lab collection. The 108 SNPs were chosen on the basis of their phylogenetic position (including also some known variants which did not fall within our NGS target region). We also considered the presence or absence of the same variant positions in other datasets (Hallast et al. 2014; Batini et al. 2015; Francalacci et al. 2015; ISOGG, date: 27 October 2015; Karmin et al. 2015; Trombetta et al. 2015a,b; Poznik et al. 2016), focusing on the SNPs which seemed to have a peculiar ethno-geographic distribution.

Analysis of the selected SNPs

Amplification

We designed PCR (Polymerase Chain Reaction) primers using the online Primer3 software (http://bioinfo.ut.ee/primer3-0.4.0/), on the basis of the Y chromosome reference sequence in the Feb. 2009 (GRCh37/hg19) assembly of the human genome, available on the UCSC Genome browser.

The PCR was performed in a final volume of 50 ml and the reaction mix was prepared with ~ 50 ng of genomic DNA, 200 mM of each deoxyribonucleotide, 2.4 mM $MgCl_2$, 1 unit Taq polymerase and 10 pmoles of both primers. We used a touch-down program, characterised by the decrease of the annealing temperature from 63° to 56° C, followed by an extension step of 40'' at 72°C, for 14 cycles. The final 30 cycles were performed setting standard amplification parameters: annealing at 56°C and extension at 72°C.

RFLP

When the mutation introduced or removed a cut site of a restriction enzyme, we genotyped the samples using the RFLP (Restriction Fragment Length Polymorphism) approach. The PCR products were digested by the respective restriction enzyme according to the manufacturer's protocol. The restriction pattern was checked by an electrophoretic run on a 2% agarose gel.

Sanger sequencing

When the RFLP was not possible, we genotyped the samples by Sanger sequencing, performed by the Eurofins Genomics (Vimodrone, Milan) or by the Bio-Fab Research (Rome) on an ABI 3730XL sequencing machine (Applied Biosystem).

Chromatograms were aligned to the human reference sequence (Feb. 2009 - GRCh37/hg19 assembly) using Sequencer 4.8 (Gene Codes Corporation, Ann Arbor, MI) and the variant positions were checked.

Population data from literature

We extracted the frequency distribution of the selected variants from the NGS data of one Sardinian population (Francalacci et al. 2015) and 16 populations from Phase 3 of the 1000 Genomes Project (Poznik et al. 2016) (Appendix 2). The geographic localization of the populations from literature and from our lab collection is reported in figure 23.

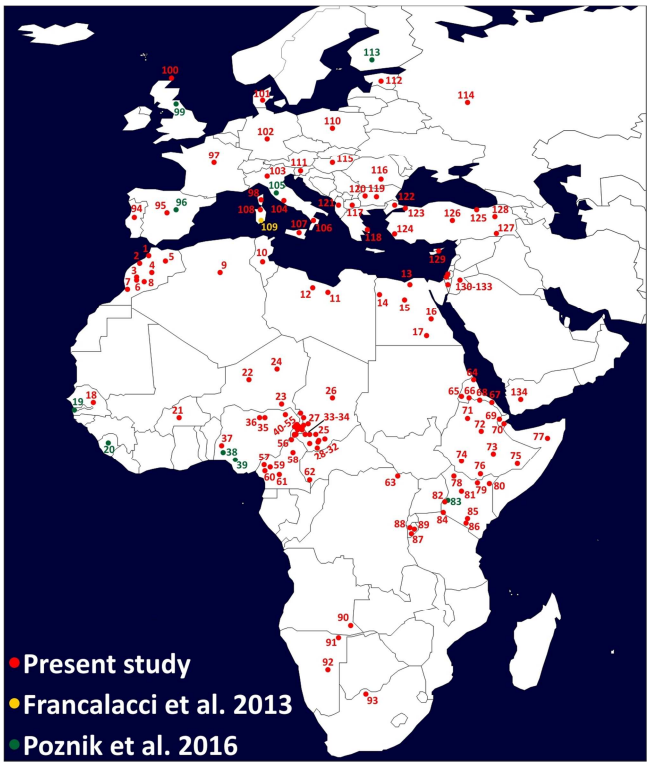


Figure 23: Geographic localization of the populations used for the genotyping. The populations belonged to three different sets, according to the legend (bottom left).

Frequency maps

Frequency maps were drawn using the Surfer 6.0 software (Golden Software, Inc., Golden, CO). The input file reported the three coordinates X, Y and Z (longitude, latitude and haplotype frequency, respectively) for each population. The missing Z values were extrapolated using the Kriging method (SURFER manual, Keckler 1997). For the frequency map, we used a grid with 100 rows × 78 columns.

REFERENCES

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

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

Adkins J et al. (2006) The ‘‘African humid period’’ and the record of marine upwelling from excess ^{230}Th in Ocean Drilling Program Hole 658C. *Paleoceanography* 21: PA4203.

Ansari Pour et al. (2013) Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people. *Eur. J. Hum. Genet.* 21: 423-429.

Arredi B. et al. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am. J. Hum. Genet.* 75: 338-345.

Balaresque P et al. (2010) A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* 8: e1000285.

Bandelt HJ et al. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16: 37-48.

Batini C et al. (2011) Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.* 28: 2603-2613.

Batini C et al. (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat. Commun.* 6: 7152.

Battaglia V et al. (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur. J. Hum. Genet.* 17: 820-830.

Batzer MA and Deininger PL (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3: 370-379.

Bekada A et al. (2013) Introducing the Algerian mitochondrial DNA and Y-chromosomes profiles into the North African landscape. *PLoS One* 8: e56775.

Beleza S et al. (2005) The genetic legacy of western Bantu migrations. *Hum. Genet.* 117: 366-375.

Bender ML (1997) The Nilo-Saharan languages: A comparative essay. *Lincom Europa*: *Munich.*

Berniell-Lee G et al. (2009) Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol. Biol. Evol.* 26: 1581-1589.

Blanchet CL et al. (2014) Asynchronous changes in vegetation, runoff and erosion in the Nile River watershed during the Holocene. *PLoS One* 9: e115958.

Blench R (2006) Archaeology, language, and the African past. *Altamira Press: Oxford.*

Blome MW et al. (2012) The environmental context for the origins of modern human diversity: a synthesis of regional variability in African climate 150,000-30,000 years ago. *J. Hum. Evol.* 62: 563-592.

Bosch E et al. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am. J. Hum. Genet.* 68: 1019-1029.

Boulet J et al. (1984) *Les groups humains.* In: *Le Nord du Cameroun*. *ORSTOM: Paris.* 103:157.

Brooks N et al. (2005) The climate-environment-society nexus in the Sahara from prehistoric times to present day. *J. North Afr. Stud.* 10: 253-292.

Brooks N (2006) Cultural responses to aridity in the Middle Holocene and increased social complexity. *Quat. Int.* 151: 29-49.

Bučková J et al. (2013) Multiple and differentiated contributions to the male gene pool of pastoral and farmer populations of the African Sahel. *Am. J. Phys. Anthropol.* 151: 10-21.

Busby GB et al. (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc. Biol. Sci.* 279: 884-892.

Busby GB et al. (2016) Admixture into and within sub-Saharan Africa. *eLife* 5: e15266.

Campbell MC et al. (2014) The peopling of the African continent and the diaspora into the new world. *Curr. Opin. Genet. Dev.* 29: 120-132.

Carrión JS et al. (2011) Early Human Evolution in the Western Palaearctic: Ecological Scenarios. *Quat. Sci. Rev.* 30: 1281-1295.

Casanova M et al. (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* 230: 1403-1406.

Chiaroni J et al. (2009) Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc. Natl. Acad. Sci. USA* 106: 20174-20179.

Chiaroni J et al. (2010) The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations. *Eur. J. Hum Genet.* 18: 348-353.

Cinnioğlu C et al. (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* 114: 127-148.

Cornevin R (1962) *Histoire des peoples de l'Afrique noire. Berger-Levrault: Paris.*

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

Cruciani F et al. (2004) Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* 74: 1014-1022.

Cruciani F et al. (2006) Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. *Hum. Mutat.* 27: 831-832.

Cruciani F et al. (2007) Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol. Biol. Evol.* 24: 1300-1311.

Cruciani F et al. (2008) Recurrent mutation in SNPs within Y chromosome E3b (E-M215) haplogroup: a rebuttal. *Am. J. Hum. Biol.* 20: 614-616.

Cruciani F et al. (2010) Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. *Eur. J. Hum. Genet.* 18: 800-807.

Cruciani F et al. (2011) A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am. J. Hum. Genet.* 88: 814-818.

David N (1971) The Fulani compound and the archaeologist. *World Archaeology* 3: 111-131.

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

deMenocal P et al. (2000) Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* 19: 347-361.

Di Giacomo F et al. (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum. Genet.* 115: 357-371.

di Lernia S (2002) Dry climatic events and cultural trajectories: adjusting Middle Holocene Pastoral economy of the Libyan Sahara. In: Hassan FA (Ed.), Droughts, Food and Culture. *Kluwer Academic/Plenum Publishers: New York.* 225-250.

Di Rienzo A et al. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91: 3166-3170.

Drake NA et al. (2011) Ancient watercourses and biogeography of the Sahara explain the peopling of the desert. *Proc. Natl. Acad. Sci. USA* 108: 458-462.

Drmanac R et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327: 78-81.

Drummond AJ and Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7: 214.

Ehret C (1995) Reconstructing Proto-Afroasiatic: consonants, vowels, tone and vocabulary. *University of California Press: Berkeley.*

Fadhlaoui-Zid K et al. (2013) Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. *PLoS One* 8: e80293.

Fernández E et al. (2014) Ancient DNA analysis of 8000 B.C. near eastern farmers supports an early neolithic pioneer

maritime colonization of Mainland Europe through Cyprus and the Aegean Islands. *PLoS Genet.* 10: e1004401.

Feuk L et al. (2006) Structural variation in the human genome. *Nat. Rev. Genet.* 7: 85-97.

Flores C et al. (2005) Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. *J. Hum. Genet.* 50: 435-441.

Foote S et al. (1992) The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* 258: 60-66.

Forster P et al. (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* 59: 935-945.

Francalacci P et al. (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* 341: 565-569.

Francalacci P et al. (2015) Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human Y chromosome from a Sardinian sample. *BMC Res. Notes* 8: 174.

Fu Q et al. (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514: 445-449.

Fu Q et al. (2016) The genetic history of Ice Age Europe. *Nature* 534: 200-205.

Gallego Llorente M et al. (2015) Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* 350: 820-822.

Gandini F et al. (2016) Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. *Sci. Rep.* 6: 25472.

Gasse F et al. (1990) The arid-humid transition in the Sahara and the Sahel during the last deglaciation. *Nature* 346: 141-146.

Gauthier JG (1979) *Archéologie du pays Fali, Nord Cameroun. Editions du CNRS: Paris.*

Goldstein DB et al. (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139: 463-471.

Gomes V et al. (2010) Digging deeper into East African human Y chromosome lineages. *Hum. Genet.* 127: 603-613.

Gonçalves R et al. (2003) Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers. *Hum. Genet.* 113: 467-472.

González M et al. (2013) The genetic landscape of Equatorial Guinea and the origin and migration routes of the Y chromosome haplogroup R-V88. *Eur. J. Hum. Genet.* 21: 324-331.

Graves JAM (2006) Sex chromosome specialization and degradation in mammals. *Cell* 124: 901-914.

Gurdasani D et al. (2015) The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517: 327-332.

Gusmão L et al. (2005) Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* 26: 520-528.

Hallast P et al. (2013) Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. *PLoS Genet.* 9: e1003666.

Hallast P et al. (2015) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol. Biol. Evol.* 32: 661-673.

Hamilton MJ et al. (2009) Population stability, cooperation, and the invasibility of the human species. *Proc. Natl. Acad. Sci. USA* 106: 12255-12260.

Hammer MF and Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet.* 56: 951-962.

Hammer MF and Zegura SL (2002) The human Y chromosome haplogroup tree: nomenclature and phylogeny of its major divisions. *Ann. Rev. Anthropol.* 31: 303-321.

Hammer MF (1994) A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* 11: 749-761.

Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature* 378: 376-378.

Hammer MF et al. (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci. USA* 97: 6769-6774.

Hammer MF et al. (2001) Hierarchical patterns of global human Ychromosome diversity. *Mol. Biol. Evol.* 18: 1189-1203.

Hammer MF et al. (2003) Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics* 164: 1495-1509.

Harris P et al. (1986) Determination of the DNA content of human chromosomes by flow cytometry. *Cytogenet. Cell Genet.* 41: 14-21.

Hassan HY et al. (2008) Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. *Am. J. Phys. Anthropol.* 137: 316-323.

Helgason A et al. (2015) The Y-chromosome point mutation rate in humans. *Nat. Genet.* 47: 453-457.

Henn BM et al. (2008) Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci. USA* 105: 10693-10698.

Henn BM et al. (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* 8: e1002397.

Hodgson JA et al. (2014) Early back-to-Africa migration into the Horn of Africa. *PLoS Genet.* 10: e1004393.

Hopen CE (1958) The pastoral Fulbé family in Gwandu. *International African Institute: London.*

Houck CM et al. (1979) A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* 132: 289-306.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-933.

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

Jobling MA (2008) Copy number variation on the human Y chromosome. *Cytogenet. Genome Res.* 123: 253-262.

Jobling MA et al. (1996) Recurrent duplication and deletion polymorphisms on the long arm of Y chromosome in normal males. *Hum. Mol. Genet.* 5: 1767-1775.

Jobling MA et al. (2013) Human evolutionary genetics (second edition). *Garland Science: New York.*

Jones ER et al. (2015) Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* 6: 8912.

Karafet T et al. (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* 69: 615-628.

Karafet T et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18: 830-838.

Karmin M et al. (2015) A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* 25: 459-466.

Kayser M et al. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* 66: 1580-1588.

Kayser M et al. (2004) A comprehensive survey of human Y chromosomal microsatellites. *Am. J. Hum. Genet.* 74: 1183-1197.

Kayser M et al. (2006) Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol. Biol. Evol.* 23: 2234-2244.

Keckler D (1997) Surfer Manual. *Golden software Inc: USA.*

King TE et al. (2007a) Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *Eur. J. Hum. Genet.* 15: 288-293.

King TE et al. (2007b) Thomas Jefferson's Y chromosome belongs to a rare European lineage. *Am. J. Phys. Anthropol.* 132: 584-589.

Knight A et al. (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr. Biol.* 13: 464-473.

Kong A et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488: 471-475.

Kuper R and Kröpelin S (2006) Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. *Science* 313: 803-807.

Lahn BT and Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286: 964-967.

Lahn BT et al. (2001) The human Y chromosome, in the light of evolution. *Nat. Rev. Genet.* 2: 207-216.

Lahr MM (2010) Saharan corridors and their role in the evolutionary geography of 'Out of Africa I'. In: Fleagle JG et al. (Eds) *Springer*: *Dordrecht.* 27-46.

Larrasoaña JC et al. (2013) Dynamics of green Sahara periods and their role in hominin evolution. *PLoS One.* 8: e76514.

Lazaridis I et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513: 409-413.

Lazaridis I et al. (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature* 536: 419-424.

Lebatard AE et al. (2010) Application of the authigenic 10 Be/ 9 Be dating method to continental sediments: Reconstruction of the Mio-Pleistocene sedimentary sequence in the early hominid fossiliferous areas of the northern Chad Basin. *Earth Planet Sci. Lett.* 297: 57-70.

Levinson G and Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4: 203-221.

Lézine AM et al. (2011) Sahara and Sahel vulnerability to climate changes, lessons from Holocene hydrological data. *Quat. Sci. Rev.* 30: 3001-3012.

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

Li H et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987-2993.

Linstädter J and Kröpelin S (2004) Wadi Bakht revisited: Holocene climate change and prehistoric occupation in the Gilf Kebir region of the Eastern Sahara, SW Egypt. *Geoarchaeology* 19: 753-778.

Luis JR et al. (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.* 74: 532-544.

Makova KD and Hardison RC (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* 16: 213-223.

Malaspina P et al. (1998) Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. *Am. J. Hum. Genet.* 63: 847-860.

Malaspina P et al. (2000) Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa. *Ann. Hum. Genet.* 64: 395-412.

Malaspinas AS et al. (2016) A genomic history of Aboriginal Australia. *Nature* 538: 207-214.

Mallick S et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538: 201-206.

Manning K and Timpson A (2014) The demographic response to Holocene climate change in the Sahara. *Quat. Sci. Rev.* 101: 28-35.

McGee D et al. (2013) The magnitude, timing and abruptness of changes in North African dust deposition over the last 20,000 yr. *Earth Planet. Sci. Lett.* 371-372: 163-176.

Mendez FL et al. (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am. J. Hum. Genet.* 92: 454-459.

Mona S et al. (2007) Patterns of Y-chromosome diversity intersect with the Trans-New Guinea hypothesis. *Mol. Biol. Evol.* 24: 2546-2555.

Montano V et al. (2011) The Bantu expansion revisited: a new analysis of Y chromosome variation in Central Western Africa. *Mol. Ecol.* 20: 2693-2708.

Morelli L et al. (2010) A comparison of Y-chromosome variation in Sardinia and Anatolia is more consistent with cultural rather than demic diffusion of agriculture. *PLoS One* 5: e10419.

Morton NE (1991) Parameters of the human genome. *Proc. Natl. Acad. Sci. USA* 88: 7474-7476.

Mohyuddin A et al. (2006) Detection of novel Y SNPs provides further insights into Y chromosomal variation in Pakistan. *J. Hum. Genet.* 51: 375-378.

Murdock GP (1959) Africa. Its peoples and their culture history. *Mc Graw-Hill Book Company, Inc: New York.*

Myres NM et al. (2010) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* 19: 95-101.

Nebel A et al. (2001) The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am. J. Hum. Genet.* 69: 1095-1112.

Newman JL (1995) The peopling of Africa: a geographic interpretation. *Yale University Press: New Haven.*

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258: 67-86.

Nurse D and Philippson G (2003) The Bantu Languages. Routledge Language Family Series. *Routledge: London.*

Ohta T and Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22: 201-204.

Omrak A et al. (2016) Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Curr. Biol.* 26: 270-275.

Osborne AH et al. (2008) A humid corridor across the Sahara for the migration of early modern humans out of Africa 120,000 years ago. *Proc. Natl. Acad. Sci. USA* 105: 16444-16447.

Pachur HJ and Hoelzmann P (2000) Late Quaternary palaeoecology and palaeoclimates of the eastern Sahara. *J. Afr. Earth Sci.* 30: 929-939.

Pagani L et al. (2016) Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* 538: 238-242.

Poznik GD et al. (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 341: 562-565.

Poznik GD et al. (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* 48: 593-599.

Rice WR (1987) Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* 116: 161-167.

Robino C et al. (2008) Analysis of Y-chromosomal SNP haplogroups and STR haplotypes in an Algerian population sample. *Int. J. Legal. Med.* 122: 251-255.

Rootsi S et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am. J. Hum. Genet.* 75: 128-137.

Rootsi S et al. (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur. J. Hum. Genet.* 15: 204-211.

Rosa A et al. (2007) Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol. Biol.* 7: 124.

Ross MT et al. (2005) The DNA sequences of the human X chromosome. *Nature* 17: 325-337.

Rosser ZH et al. (2009) Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. *Am. J. Hum. Genet.* 85: 130-134.

Rozen S et al. (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423: 873-876.

Saillard J et al. (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* 67: 718-726.

Sanchez JJ et al. (2005) High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *Eur. J. Hum. Genet.* 13: 856-866.

Scerri EML et al. (2014) Earliest evidence for the structure of *Homo sapiens* populations in Africa. *Quat. Sci. Rev.* 101: 207-216.

Scozzari R et al. (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* 65: 829-846.

Scozzari R et al. (2012) Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. *PLoS One* 7: e49170.

Scozzari R et al. (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. *Genome Res.* 24: 535-544.

Secher B et al. (2014) The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC Evol. Biol.* 14: 109.

Seligman CS (1966) Races of Africa. *Oxford University Press: Oxford.*

Ségurel L et al. (2014) Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* 15: 47-70.

Seielstad MT et al. (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum. Mol. Genet.* 3: 2159-2161.

Semino O et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* 290: 1155-1159.

Semino O et al. (2002) Ethiopians and Khoisan share the deepest clade of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* 70: 265-268.

Semino O et al. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroup E and J: inferences on the

neolithization of Europe and later migratory events in the Mediterranean area. *Am. J. Hum. Genet.* 74: 1023-1034.

Sengupta S et al. (2006) Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* 78: 202-221.

Sereno PC et al. (2008) Lakeside cemeteries in the Sahara: 5000 years of holocene population and environmental change. *PLoS One* 3: e2995.

Shen P et al. (2000) Population genetic implications from sequence variation in four Y chromosome genes. *Proc. Natl. Acad. Sci. USA* 97: 7354-7359.

Shen P et al. (2004) Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum. Mutat.* 24: 248-260.

Shriner D et al. (2016) Ancient Human Migration after Out-of-Africa. *Sci. Rep.* 6: 26565.

Skaletsky H et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequences classes. *Nature* 423: 825-837.

Skonieczny C et al. (2015) African humid periods triggered the reactivation of a large river system in Western Sahara. *Nat. Commun.* 6: 8751.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.

Smith JR et al. (2004) A reconstruction of Quaternary pluvial environments and human occupations using stratigraphy and geochronology of fossil-spring tufas, Kharga Oasis, Egypt. *Geoarchaeology* 19: 407-439.

Smith JR et al. (2007) New age constraints on the Middle Stone Age occupations of Kharga Oasis, Western Desert, Egypt. *J. Hum. Evol.* 52: 690-701.

Sutton JEG (1977) The African aqualithic. *Antiquity* 51: 25-34.

Tamura K et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739.

Tierney JE and deMenocal PB (2013) Abrupt shifts in Horn of Africa hydroclimate since the Last Glacial Maximum. *Science* 342: 843-846.

Tishkoff SA et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324: 1035-1044.

Triki-Fendri S et al. (2015) Paternal lineages in Libya inferred from Y-chromosome haplogroups. *Am. J. Phys. Anthropol.* 157: 242-251.

Trombetta B et al. (2010) Footprints of X-to-Y gene conversion in recent human evolution. *Mol. Biol. Evol.* 27: 714-725.

Trombetta B et al. (2011) A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. *PLoS One* 6: e16073.

Trombetta B et al. (2014) Inter- and intraspecies phylogenetic analyses reveal extensive X-Y gene conversion in the evolution of gametologous sequences of human sex chromosomes. *Mol. Biol. Evol.* 31: 2108-2123.

Trombetta B et al. (2015a) Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the African continent. *Genome Biol. Evol.* 7: 1940-1950.

Trombetta B et al. (2015b) Regional differences in the accumulation of SNPs on the male-specific portion of the human Y chromosome replicate autosomal patterns: Implications for genetic dating. *PLoS One* 10: e0134646.

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

Underhill PA et al. (1996) A pre-Columbian Y chromosome specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci. USA* 93: 196-200.

Underhill PA et al. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7: 996-1005.

Underhill PA et al. (2000) Y chromosome sequence variation and the history of human populations. *Nat. Genet.* 26: 358-361.

Underhill PA et al. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* 65: 43-62.

Underhill PA et al. (2010) Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur. J. Hum. Genet.* 18: 479-484.

Vallone PM and Butler JM (2004) Y-SNP typing of U.S. African American and Caucasian samples using allele-specific hybridization and primer extension. *J. Forensic Sci.* 49: 723-732.

Weber JL and Wong C (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2: 1123-1128.

Wei W et al. (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* 23: 388-395.

Whitfield LS et al. (1995) Sequence variation of the human Y chromosome. *Nature* 378: 379-380.

Wood ET et al. (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur. J. Hum. Genet.* 13: 867-876.

Xue Y et al. (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deeprooting pedigree. *Curr. Biol.* 19: 1453-1457.

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

Yellen JE (1998) Barbed bone points: Tradition and continuity in Saharan and sub-Saharan Africa. *Afr. Archaeol. Rev.* 15: 173-198.

Zalloua PA et al. (2008) Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am. J. Hum. Genet.* 82: 873-882.

Zhong H et al. (2010) Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *J. Hum. Genet.* 55: 428-435.

Web resources

ISOGG	http://isogg.org/
Primer 3:	http://bioinfo.ut.ee/primer3-0.4.0/
UCSC Genome browser:	http://genome.ucsc.edu/

APPENDICES

Appendix 1.

List of 7544 variants and their position in the phylogeny represented in figure 5.

The list is available as an Excel spreadsheet at:
https://www.dropbox.com/sh/74gsofkinbhk61s/AABMpvqPLY0lEYP08ppVpqZqa?dl=0

Appendix 2.

Frequencies of the four trans-Saharan haplogroups in the 124 populations analysed in the present thesis.

The results are available as an Excel spreadsheet at:
https://www.dropbox.com/sh/74gsofkinbhk61s/AABMpvqPLY0lEYP08ppVpqZqa?dl=0

LIST OF PUBLICATIONS

1. Iacovacci* G, D'ATANASIO* E, Marini* O, Coppa A, Sellitto D, Trombetta B, Berti A, Cruciani F. Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four Eastern African countries. *Forensic Sci Int Genet.* 27: 123-131.
2. Trombetta B, Fantini G, D'ATANASIO E, Sellitto D, Cruciani F. (2016) Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Sci. Rep.* 6: 28710.
3. Rapone* C, D'ATANASIO* E, Agostino A, Mariano M, Papaluca MT, Cruciani F, Berti A. (2016) Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler® Plus kit) in an Italian population sample. *Forensic Sci. Int. Genet.* 21: e1-5.
4. Trombetta B, D'ATANASIO E, Massaia A, Myres NM, Scozzari R, Cruciani F, Novelletto A. (2015) Regional Differences in the Accumulation of SNPs on the Male-Specific Portion of the Human Y Chromosome Replicate Autosomal Patterns: Implications for Genetic Dating. *PLoS ONE* 10: e0134646.
5. Trombetta* B, D'ATANASIO* E, Massaia A, Ippoliti M, Coppa A, Candilio F, Coia V, Russo G, Dugoujon JM, Moral P, Akar N, Sellitto D, Valesini G, Novelletto A, Scozzari R, Cruciani F. (2015) Phylogeographic Refinement and Large Scale Genotyping of Human Y Chromosome Haplogroup E Provide New Insights into the Dispersal of Early Pastoralists in the African Continent. *Genome Biol. Evol.* 7: 1940-1950.
6. Scozzari R, Massaia A, D'ATANASIO E, Myres NM, Perego UA, Trombetta B, Cruciani F. (2012) Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. *PLoS ONE.* 7: e49170.

*** These authors contributed equally to this work**