



SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE
XXVI Ciclo
(A.A. 2012/2013)

HUMAN Y CHROMOSOME VARIATION
AND THE PEOPLING OF THE AFRICAN CONTINENT

Dottorando
Andrea Massaia

Docente guida
Prof. Fulvio Cruciani

Tutore
Prof. Andrea Novelletto

Coordinatore
Prof. Irene Bozzoni

Andrea Massaia

TABLE OF CONTENTS

Abbreviations	p. 4
Summary	p. 5
Introduction	p. 7
The human Y chromosome	p. 7
Uniparental markers and phylogenetic trees	p. 13
The origin of <i>Homo sapiens</i>	p. 22
Aim of the study	p. 31
Results	p. 33
Discussion	p. 66
Materials and methods	p. 78
References	p. 88
Appendices	p. 112
Recent publications	p. 113

ABBREVIATIONS

AMH = Anatomically modern human

ASD = Average of the squared distance

CNV = Copy number variation

kya = Kilo years ago

MP = Maximum parsimony

MRCA = Most recent common ancestor

MSY = Male-specific region of the Y chromosome

mtDNA = Mitochondrial DNA

NGS = Next generation sequencing

PAR = Pseudoautosomal region

SINE = Short interspersed element

SNP = Single nucleotide polymorphism

SNS = Single nucleotide substitution

TMRCA = Time to the most recent common ancestor

YAP = Y Alu polymorphism

SUMMARY

The analysis, by Next Generation Sequencing, of 1.5 Mb of the Male-Specific region of the Y chromosome (MSY), in a sample carefully selected to represent a wide range of diversity and antiquity among MSY lineages, led to the identification of 2,386 variable positions, 80% of which were novel. Many aspects of this pool of variants resembled the pattern observed among genome-wide *de novo* events, suggesting that in the MSY a large proportion of newly arisen alleles have survived in the phylogeny. Some degree of purifying selection emerged in the form of an excess of private missense variants.

We used these markers to reconstruct a phylogenetic tree, which showed remarkable differences with the one known in literature, although recapitulating the previously known topology. The relative lengths of the tree branches have been notably altered, and the time estimates associated with the tree nodes have moved towards more ancient times. Keeping into account the present day distribution of patrilineages, and the fossil remains of *Homo sapiens* found to date, our data enabled us to draw hypotheses on the evolutionary events that involved the human species, since its origin, up to its migration out of the African continent.

Andrea Massaia

INTRODUCTION

The human Y chromosome

Structure of the human Y chromosome

The Y chromosome is one of the smallest chromosomes in the human genome, measuring about 58 Mb in length (Harris et al. 1986; Morton 1991; Foote et al. 1992; Group NCCM 1992, International Human Genome Sequencing Consortium 2001; Skaletsky et al. 2003).

The telomeric portions of the Y chromosome are known as PseudoAutosomal Regions (PARs), named PAR1 and PAR2 on the short and on the long arm, respectively. These regions constitute about the 5% of the length of the chromosome, with 29 genes identified inside them to date, and show a high recombination rate with the allelic X chromosome regions during male meiosis (Ross et al. 2005).

The PARs flank a vast region, named as the Male-Specific region of the Y chromosome (MSY), accounting for 95% of the Y chromosome length, which follows a male uniparental inheritance pattern.

The MSY's euchromatic DNA sequences total roughly 22.5 Mb, including 8 Mb on the short arm (Yp) and 14.5 Mb on the long arm (Yq) (figure 1).

Three heterochromatic blocks are also found on the Y chromosome, with the largest (approximately 40 Mb long)

comprising the bulk of the distal long arm. Another smaller block, approximately 400 kb long, comprises 3,000 tandem repeats of 125 base pairs (bp), and interrupts the euchromatic sequences of proximal Yq. A third block may be found in the centromeric region (a feature of all nuclear chromosome), and is approximately 1 Mb long (Skaletsky et al. 2003).

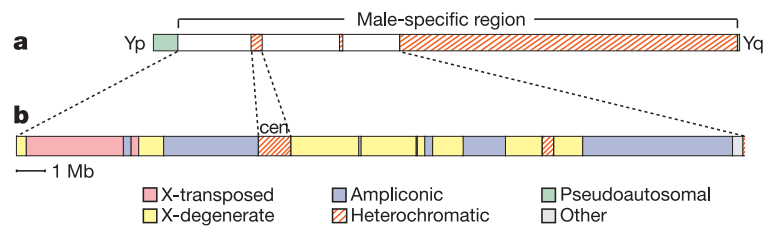


Figure 1: Structure of the MSY. a) Schematic representation of the whole chromosome. b) Enlarged view of a 24-Mb portion of the MSY, extending from the proximal boundary of the Yp pseudoautosomal region to the proximal boundary of the large heterochromatic region of Yq. Shown are three classes of euchromatic sequences, as well as heterochromatic sequences. A 1-Mb bar indicates the scale of the diagram (modified from Skaletsky et al. 2003).

Three discrete sequence classes can be identified in the MSY euchromatin: X-transposed, X-degenerate and ampliconic.

The X-transposed sequences are located in two blocks on the short arm, with a combined length of 3.4 Mb and 99% identity to DNA sequences in Xq21; they are the product of a massive X-to-Y transposition event that took place about 3-4 million years ago (Ross et al. 2005), after the divergence of the human and chimpanzee lineages, followed by an inversion within the MSY short arm (Skaletsky et al. 2003).

The X-degenerate sequences are considered as 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 et al. 2006), and display between 60% and 96% nucleotide sequence identity to their X-linked paralogous regions.

The third class, the ampliconic sequences, are composed largely of sequences that exhibit marked similarity – as much as 99.9% identity over tens or hundreds of kilobases – to other regions in the MSY, and are located in seven segments whose combined length is 10.2 Mb.

Variation in the human Y chromosome

Biallelic polymorphisms

Biallelic polymorphisms are markers that only have two states, one being the ancestral and the other the derived. While the first MSY polymorphism in this class was discovered almost thirty years ago (Casanova et al. 1985), only few of such markers, including Single Nucleotide Polymorphisms (SNPs), deletions, and *Alu* insertions, were discovered until 1996 (Hammer 1994; Seielstad et al. 1994; Whitfield et al. 1995; Jobling et al. 1996). The introduction of DHPLC (Denaturing High Performance Liquid Chromatography) and improvements in the sequencing techniques led to the discovery, in the subsequent years, of over 600 new MSY biallelic polymorphisms (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, 2010a, 2011a; 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 brought new life to the search of variants in the MSY, leading to the discovery of thousands of new polymorphisms (Xue et al. 2009; The 1000 Genome Project Consortium 2010; Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013).

Several studies highlighted the lower genetic variation of the MSY when compared to the autosomes, the X chromosome, or the mtDNA (International SNP Map Working Group 2001). This can be explained by two factors at least:

1. The smaller MSY effective population size (Hammer 1995; Underhill et al. 1996) when compared to the rest of the genome, an effect enhanced by male mortality (due to wars, hunting, etc.) or cultural phenomena like polygyny. This exposes the MSY to a stronger effect of genetic drift, which leads to faster fixation of alleles and loss of diversity;
2. The lack of recombination, which might have led to the fixation of alleles associated to positively selected variants, a phenomenon known as hitchhiking (Rice 1987; Whitfield et al. 1995).

The MSY biallelic polymorphisms are usually regarded as “stable” in an evolutionary perspective. Their mutation rate, recently estimated at less than 10^{-9} events/position/year (Mendez et al. 2013), is remarkably lower than the mtDNA mutation rate, and makes it very unlikely that any site is hit by repeated mutations more than once in recent evolution (Jobling et al. 2013). This means that different chromosomes displaying the same derived state at a site probably descend from a common ancestor. Chromosomes sharing the derived state at one or more sites are thus gathered in haplogroups, which constitute a univocal phylogeny (Karafet et al. 2008).

Other than Single Nucleotide Substitutions (SNSs), *Alu* insertions represent another important class. *Alu* elements, approximately 300 bp long, are retrotransposons of the SINE (Short Interspersed Element) class, and are named after a restriction site for the *AluI* enzyme within them (Houck et al. 1979). They show higher average nucleotide diversity when compared to the rest of the genome, due to the presence of abundant CpG sites (Batzer and Deininger 2002). Their transposition features the retrotranscription of an RNA intermediate,

although only few “master” copies in the genome are competent for retrotransposition. As variant alleles in any of these “master” elements are passed to all of its copies, they identify different *Alu* families and subfamilies. The expansion of *Alu* elements is a recent event, with some of them being polymorphic for presence/absence. Only one such polymorphism has been so far identified in the MSY (Hammer and Horai 1995).

Multiallelic polymorphisms

Several classes of multiallelic polymorphisms are present in the Y chromosome, with microsatellites being the most widely employed in research.

Microsatellites are tandem repeats of 1-6 bp stretches, and most of them are polymorphic for the number of repeats, with a number of alleles usually greater than two (Jobling et al. 2013). Their mutation rate (approximately 2×10^{-3}) is several orders of magnitude higher when compared to the SNPs mutation rate, and this impairs the use of microsatellites in the reconstruction of phylogenetic relations, as identity by state (equal number of repeats) does not necessarily correspond to identity by descent. Nevertheless, microsatellites are employed in the analysis and dating of recent microevolutionary events (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 a stepwise model described by Ohta and Kimura (1973), with the possibility of rare, large variation of the number of repeats (Di Rienzo et al. 1994, Malaspina et al. 1998, 2000). On molecular grounds, the “slipped strand mispairing” model (Levinson and Gutman 1987) explains the alteration of the number of repeats with the slippage of one of the two DNA strands during replication.

Copy number variations

Comprising both biallelic and multiallelic markers, the Copy Number Variations (CNVs) are a comprehensive class of variants, featuring different number of copies of sequences longer than 1 kb (Feuk et al. 2006). Reduction or increment of copy number may be the result of different mechanisms, which may involve large homologous sequences (mainly segmental duplications, regions more than 1 kb long and displaying identity over 90%), as well as sequences with short (2-15 bp) homologous regions. Some of these mechanisms are physiological responses to single or double strand breaks, but are also capable of leading to chromosomal aberrations and copy number variations.

Haploidy has led to the accumulation of a large number of segmental duplications on the Y chromosome. This, in turn, created the ideal landscape for the accumulation of CNVs. The Y chromosome also allows the application of the phylogenetic approach to the study of CNV, if their search is conducted in individuals for whom the phylogenetic relations are known. A robust phylogenetic contest allows the identification of CNVs alleles shared by descent, making it possible to count the minimum number of mutational events for a given CNV. The rate of CNV generation could then be deduced if the numbers of generations encompassed by the sampled chromosomes are known (Jobling 2008).

Uniparental markers and phylogenetic trees

The largest part of our diploid genome is evenly inherited from both parents. Each of them passes us a whole haploid genome, with a set of alleles (haplotypes) linked to each chromosome. Haplotypes are then reshuffled by recombination during meiosis. The mitochondrial DNA (mtDNA) and the MSY, being uniparentally inherited and not affected by meiotic recombination, represent the exceptions to these rules.

In the MSY, sequential accumulation of mutational events is the only source of intrapopulation genetic diversity (Rozen et al. 2003; Skaletsky et al. 2003). Derived alleles generated by mutational events are passed on along the male lineages. This process creates monophyletic entities, known as “haplogroups”, which show a strong geographic differentiation due to molecular divergence during the dispersal of mankind over the continents (Jobling and Tyler-Smith 2003). The phylogeographic approach keeps into account the phylogenetic relations among haplogroups and their ethno-geographic distribution, and it has allowed investigators to understand some demographic processes behind the origin of *Homo sapiens* and the dispersal of human populations (Chiaroni et al. 2009).

Haplogroups are stable entities, as they are defined by markers (usually SNPs) with a low mutation rate; they can thus be arranged in an unambiguous maximum parsimony phylogenetic tree. The most recent high-resolution tree to encompass all haplogroups (Figure 2) contains 20 main clades, indicated with letters from A to T (Karafet et al. 2008).

As new lineages are discovered, and larger amounts of sequence are surveyed, the structure of the MSY phylogenetic tree is perpetually adjusted to accommodate the new findings. In recent years, the deepest portion of the phylogeny was remarkably

susceptible to such changes. The deepest split, separating haplogroup A from the rest of the phylogeny (Karafet et al. 2008), was challenged by the discovery of a completely different structure (Cruciani et al. 2011a, Scozzari et al. 2012), which showed the polyphyletic nature of the lineages formerly grouped into haplogroup A. As a result, such lineages were no more grouped together in a single clade, and haplogroup A1b (recently renamed as A0) was identified as the deepest-rooting branch. An even more recent work (Mendez et al. 2013) discovered a very deep, rare lineage, named A00, whose split from the rest of the phylogeny long predates that of A1b. Such radical changes in structure directly lead to adjustments of time estimates and geographic inferences associated with the phylogenetic tree; these subjects will be discussed in detail in the following sections.

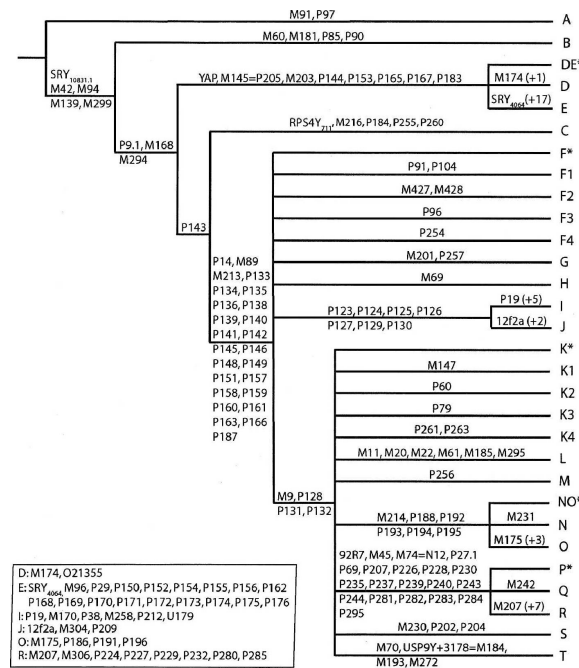


Figure 2: Phylogenetic tree of the human Y chromosome. Haplogroups are indicated with letters from A to T, shown at the tips. Markers defining branches are indicated above them (from Karafet et al. 2008).

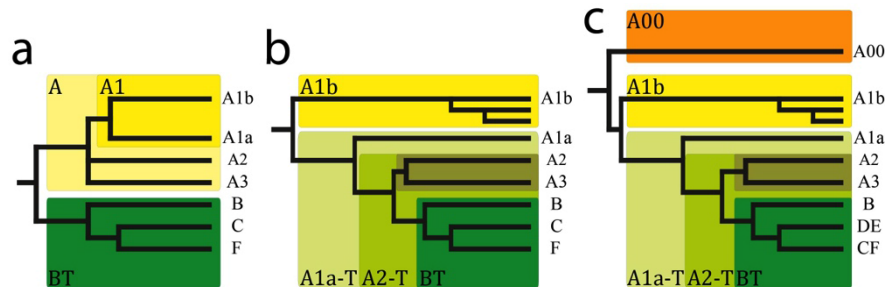


Figure 3: Changes in the basal structure of the MSY phylogeny. The basal branches are shown, as reported in Karafet et al. 2008 (a), Cruciani et al. 2011a, as modified in Scozzari et al. 2012 (b), and Mendez et al. 2013 (c).

Geographic distribution of Y chromosome haplogroups

The geographic distribution of haplogroups, seen in the light of the phylogenetic relations linking them, provides clues on the dispersal of human population over the world. In a similar fashion, the distribution of the deepest branches can be informative on the evolutionary events that involved the oldest representatives of our species.

The deepest-rooting branches of the Y phylogeny, namely A00, A1b, A1a, A2, A3 and B, are found, with some exceptions, only in the African continent, though they only represent a small fraction of the overall genetic variation in the continent (Cruciani et al. 2002, 2011a; Wood et al. 2005; King et al. 2007a; Chiaroni et al. 2009; Batini et al. 2011; Jobling et al. 2013; Mendez et al. 2013). A00 and A1b have only been reported, at very low frequencies, in small populations in Central Africa (Cruciani et al. 2011a; Mendez et al. 2013). A1a has been found at low frequencies too, having been reported in less than thirty individuals coming from a wide area spanning from Morocco to Senegal to Niger (with the exception of three Afro-Americans and an individual from Cape Verde) (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 and A3 have recently been identified as sister clades stemming from a short branch (Batini et al. 2011; Scozzari et al. 2012). A2 is mostly found in Khoisan populations from Southern Africa; a specific A2 branch has recently been observed in Central Africa pygmies (Batini et al. 2011). A3 shows a dual distribution with A3a being found in Eastern Africa only, and its sister clade A3b displaying a clear differentiation between South Africa (A3b1) and Central-Eastern Africa (A3b2). The overall A3 distribution suggests that it possibly emerged in Eastern Africa, and was later brought to the southern part of the continent during recent migrations.

Haplogroup B is mainly confined to sub-Saharan Africa (Underhill et al. 2000, 2001; Cruciani et al. 2002; Semino et al. 2002; Y Chromosome Consortium 2002; Butler 2003; Vallone and Butler 2004; Karafet et al. 2008; Gomes et al. 2010) and is divided in two main branches; the first, B1, has been so far observed in four individuals only, from southern Cameroon, Mali and Burkina Faso; the other, B2, has a wider distribution, being found in Eastern, Central and Southern Africa, with its sub-haplogroups reaching frequencies as high as 70% in some ethnic groups (Knight et al. 2003; Wood et al. 2005; Tishkoff et al. 2007; Berniell-Lee et al. 2009).

Proceeding further down the tree after haplogroup B, we find the split between macro-haplogroups DE and CT.

Haplogroup DE is characterized by the derived state of the only polymorphic *Alu* insertion in the MSY known to date (YAP, Y *Alu* Polymorphism). Haplogroup D is mainly found in Central and South-East Asia (Karafet et al. 2001), while haplogroup E is the most common haplogroup in Africa, but is also found in Europe and Middle East. Within the African continent, different E sub-haplogroups display a strong differential localization, and peak at frequencies as high as 80% in some regions or populations

(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).

Haplogroup CT is split into haplogroup C, found 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), and macro-haplogroup F, which is widely distributed over the world, although it is rarely found in sub-Saharan Africa. This macro-haplogroup contains several F branches (F1 to F4) plus other haplogroups, namely G, H, IJ and macro-haplogroup KT. Haplogroup G is mainly found in the Mediterranean basin and the Caucasian region, while H is mostly present in India (Sengupta et al. 2006). Haplogroup I is distinctive of Europe (Rootsi et al. 2004; Battaglia et al. 2009), while J is more widely distributed and is found in Europe, Middle East, India, Central Asia and Northern Africa (Hammer et al. 2000, Underhill et al. 2001; Di Giacomo et al. 2004; Chiaroni et al. 2010).

K lineages are present in India (K1), Oceania and Indonesia (K2 to K4) (Underhill et al. 2001; Y Chromosome Consortium 2002; Karafet et al. 2008). Other main haplogroups within K are also found in India (L), Indonesia (M, S), and Oceania (O, S), but also in Northern Eurasia (N), Central Asia (O, T) Africa and Middle East (T) (Capelli et al. 2001; Karafet et al. 2001; Underhill et al. 2001; 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 peculiar of the American continent (Karafet et al. 2008), while haplogroup R is one of the most important contributors to the European MSY diversity (Balaesque et al. 2010; Myres et al. 2010; Underhill et al. 2010), but is also found at very high frequency in Central-Western Africa (Cruciani et al. 2010a).

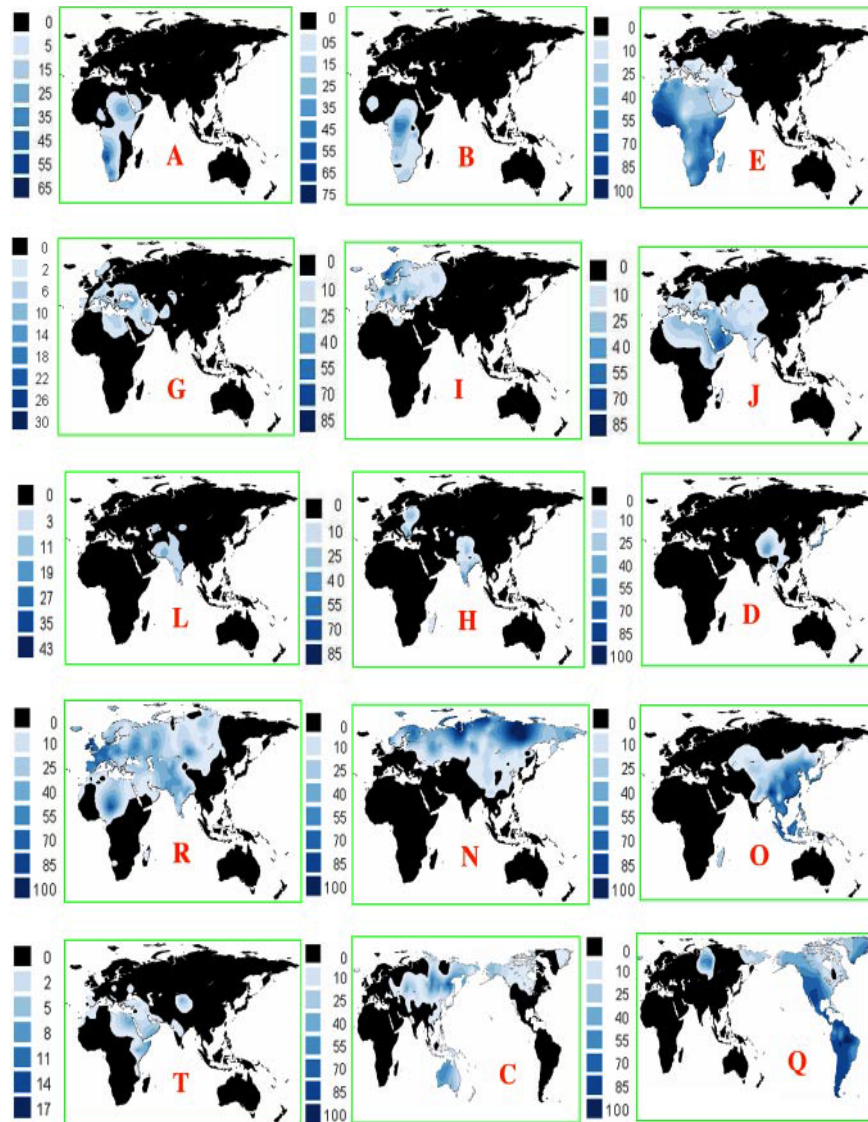


Figure 4: Distribution of the main MSY haplogroups. The panels show the geographic frequency distribution maps for the major MSY haplogroups. For each panel, the frequency scale is indicated on the left. Panel “A” collectively represent the information for haplogroups A00, A1b, A1a, A2 and A3 (from Chiaroni et al. 2009).

Time estimates for the MSY phylogeny

Together with the geographic distribution of haplogroups, time estimates associated with nodes and markers in the phylogeny are crucial to hypothesize evolutionary events. While a rooted phylogeny provides a relative chronology by itself, an absolute chronology is useful to relate the phylogeny to a wider context.

Population splits in a phylogeny can be dated using measures of genetic distance, such as F_{ST} , which tend to increase with time. However, we must assume that no gene flow occurred between populations after they split. As assumptions made about evolutionary events can heavily influence the analysis, it is convenient to model parameters like effective population size changes or migration rates jointly with the timing of population splits, comparing different models using likelihood-based approaches (Gutenkunst et al. 2009; Gronau et al. 2011; Jobling et al. 2013).

A different approach relies on alleles instead of population splits. The emergence of a new allele creates a new haplotype, which later accumulates diversity through mutation. The amount of diversity can be used to infer the age since that haplotype arose, assuming that the mutation rate is known.

Microsatellite markers can be used to this aim, as their high mutation rate leads to the quick accumulation of diversity. By assuming that the differences between alleles accumulate one by one, the number of differences represents simple genetic distance. A measure of this distance is the Average of the Squared Distance (ASD) between alleles (Goldstein et al. 1995; Slatkin 1995), which is linearly related to time (Jobling et al. 2013).

Although they are used for dating, microsatellites have two main issues: their diversity is confounded by recurrent mutation, which leads to an underestimation of diversity. Moreover, microsatellite diversity reaches a maximum (Busby et al. 2012),

making microsatellites unsuitable for the reliable estimation of ancient events (figure 5). Microsatellites also display a divergence between the mutation rate obtained by direct pedigree analysis and the effective mutation rate derived from populations with well-documented short term histories, with the latter appearing to be remarkably lower (Zhivotovsky et al. 2004, 2006). This discrepancy may be partly explained by back-and-forth changes in repeat unit number being completely observed in pedigree studies, but not in population studies, where the considerable homoplasy of alleles can be hidden (Jobling et al. 2013).

By contrast, SNSs mutation rate (Kong et al. 2012; Mendez et al. 2013) is several orders of magnitude lower than that of microsatellites, resulting in a low occurrence of recurring mutation and thus in a univocal phylogeny. As mutation alone drives diversification, it represents the unique contribution to the molecular clock, and the average number of nucleotide differences, described as the Rho (ρ) statistic (Forster et al. 1996; Saillard et al. 2000), is directly related to time (t) via the mutation rate (μ) by the equation:

$$\rho = \mu t$$

However, using SNSs for dating purposes is also somewhat limited by the intrinsic features of single nucleotide variants. A low mutation rate means that a lower amount of diversity is usually discovered in a single experiment, when compared with microsatellites. Inaccuracy in the Rho method, possibly due to the randomness of the mutational process, has been hypothesized (Cox 2009). As this randomness is averaged over time (Jobling et al. 2013), SNSs represent good markers for older events, but they behave poorly when used on shorter time scales (figure 5).

Confidence intervals for Rho can be calculated directly (Saillard et al. 2000). The increase of the number of variants considered can greatly increase the resolution of the method. In this sense, a great

impulse was given by the diffusion of NGS techniques, which vastly increased the amount of sequence surveyed, leading to the discovery of thousands of new markers (Xue et al. 2009; The 1000 Genome Project Consortium 2010; Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013). This allows investigators to overcome usual limitations of SNPs markers and use them to confidently estimate times over shorter and more recent time frames (Figure 5).

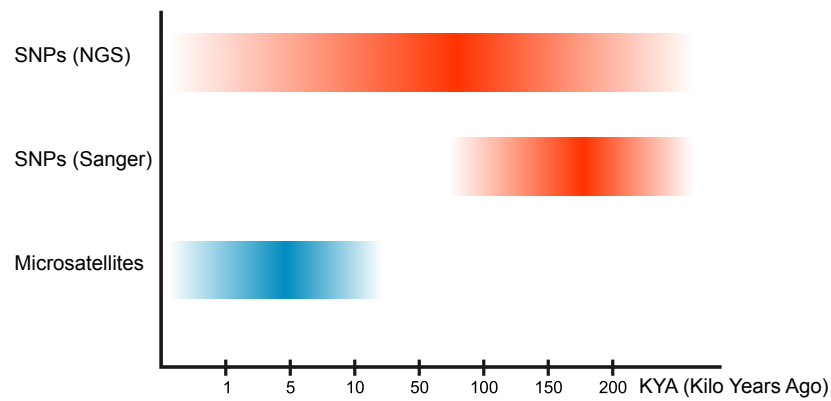


Figure 5: Dating of the MSY phylogeny using different markers. The graphic represents time intervals (in kiloyears) over which different methods produce accurate estimates. Details are given in the text.

The origin of *Homo sapiens*

The most widely accepted model for the origin of our species explains it as a recent event, which took place in the African continent less than 300 kya (Kilo Years Ago), and it is supported by archaeological, palaeoanthropological and genetic data. *Homo sapiens* would have emerged in the African continent, leaving it afterwards and dispersing over the rest of the world; it might have met archaic *Homo* species in other continents, either intermixing with them (Green et al. 2010; Reich et al. 2010) or replacing them completely.

As an alternative, the multiregional model was proposed (Wolpoff 1988). This hypothesis holds that a single human species, encompassing all archaic hominids, emerged in Africa around two millions years ago, dispersing all over the world in several migratory waves. These archaic populations would have then separately evolved into different modern *Homo sapiens* populations, through a combination of adaptation within various regions of the world and gene flow between those regions. Due to genetic (Takahata et al. 2001) and statistic (Fagundes et al. 2007) data, this model seems nowadays somewhat improbable.

The main evidences for an African origin of *Homo sapiens*, the *out of Africa* event and the routes that our species followed in its path over the world will be briefly reviewed in the following sections.

Palaeoanthropology and archaeology

When dealing with fossil human remains, the main issue is the distinction between Anatomically Modern Human (AMH) and his closest archaic relatives; the differences are often subtle (Jobling et al. 2013). Two morphological criteria are mainly used in the classification: facial retraction and neurocranial globularity (Lieberman et al. 2002); however, the morphological classification of our species and of its fossil remains is still controversial (Tattersall and Schwartz 2008; Schwartz and Tattersall 2010).

The oldest representatives of our species have been identified in the fossil remains of two individuals from the Omo Kibish sites near the Omo River, in Omo National Park in South-Western Ethiopia. These remains have been dated at 195 kya (McDougall et al. 2005; Aubert et al. 2012), pointing to Eastern Africa as the cradle of modern humans. Although they share the same dating, the two individuals show slightly different traits, with one classified as anatomically modern, and the other as “late archaic”. These differences might reflect ancient intrapopulation diversity (Trinkaus 2005).

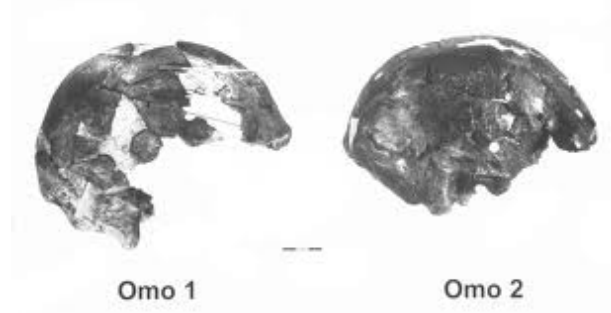


Figure 6: Skulls from Omo Kibish, Ethiopia. (From Day 1969).

Other remains from Herto (Ethiopia) have been dated slightly earlier than the Kibish fossils, around 160 kya (White et al. 2003). The Herto remains belong to three hominids, one immature and two adults, and display a mosaic of archaic and modern features, probably representing "...a population that is on the verge of anatomical modernity but not yet fully modern" (White et al. 2003). Such intermediate features are also seen as a strong evidence of AMH emergence in Africa. Though they are classified as *Homo sapiens idaltu*, they show derived morphological features that are absent in *Homo erectus* and in other older fossils (White et al. 2003).

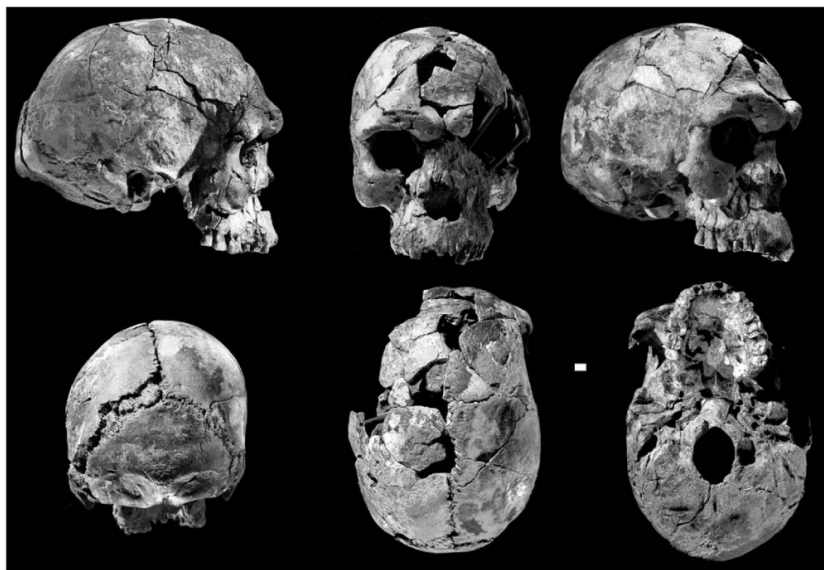


Figure 7: Fossil adult cranium from Herto, Ethiopia. Top: lateral, frontal and three-quarter views; bottom: posterior, superior and inferior views. A 1 cm bar indicates the scale of the picture (from White et al. 2003).

Although the Kibish and Herto remains provide evidence for the emergence of *Homo sapiens* in Africa, they show remarkable morphological variation (Trinkaus et al. 2005) and the persistence

of archaic traits. Due to the vastness of the African continent, and the paucity of fossil remains found to date, we can hardly say if the emergence of AMH was restricted to a limited region, or if it involved the whole continent, in a sort of “African multiregionalism” (Stringer 2003).

Considering physiological traits as development, reproduction and life span, which can sometimes be inferred from fossil remains, complicates the picture even more. Remains from Jebel Irhoud (Morocco), dated at 160 kya years ago, appear to be both anatomically and physiologically modern, while remains from the same time in Eastern Africa show more primitive traits. Above all, the teeth of the Moroccan fossils are remarkably more similar to those of *Homo sapiens* than to those of other *Homo* species, a signal that the population they belonged to already possessed the social, cultural and biological features related to a slow and prolonged development as that of modern *Homo sapiens* (Smith et al. 2007).

Archaeological remains too point at Eastern Africa as the place of origin for modern *Homo*. Remains dating to the Middle Stone Age (200 to 15 kya) display unique features when compared to other African archaeological remains, and, above all, they show both a temporal and a technological continuity, hinting a continuous occupation of the region by *Homo sapiens* (Brooks 2005).

Genetics

The worldwide distribution of genome diversity provides more evidence for an African homeland of the human species. Since the late ‘80s, mtDNA data led to hypothesize a “mitochondrial Eve” in Africa, around 200 kya (Cann et al. 1987).

As for uniparental markers, both the mtDNA and the MSY phylogenies are rooted in Africa (Behar et al. 2008; Cruciani et al. 2011a; Mendez et al. 2013), where they also display the highest diversity; the deepest branches are found in the African continent only, while the non-African clades descend from a unique node, which is in both cases a lot younger than the root (figures 8 and 9).

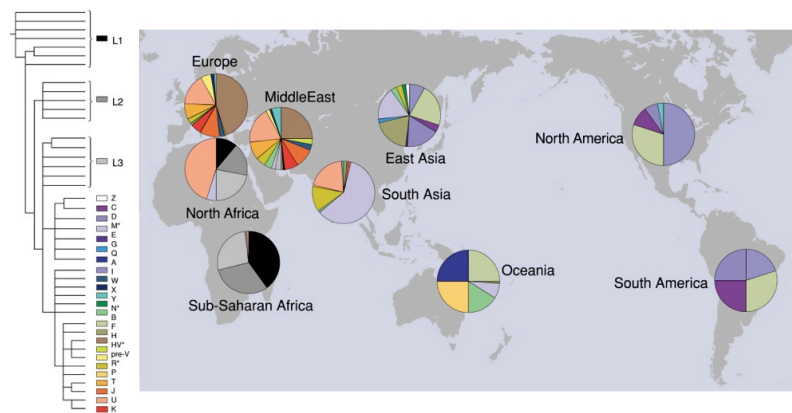


Figure 8: mtDNA haplogroups distribution. mtDNA phylogenetic tree (left) and distribution of the main mtDNA haplogroups in several world regions (right) (from Jobling et al. 2004).

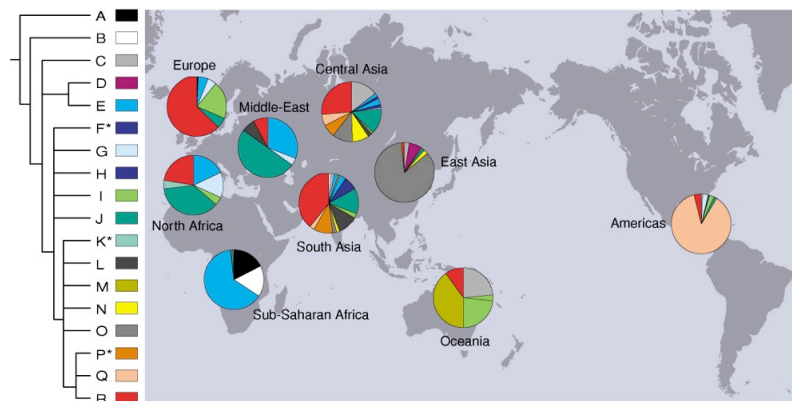


Figure 9: Y chromosome haplogroups distribution. Y chromosome phylogenetic tree (left) and distribution of the main Y haplogroups in several world regions (right) (from Jobling et al. 2004).

Autosomal diversity provides additional support. The multiregional model implies continuous gene flow among Africa, Asia and Europe for the last two millions years, resulting in a very high effective population size. As the effective population size directly relates to the amount of diversity, the multiregional model implies that we should observe a much higher diversity than that really present in human populations (Jobling et al. 2013).

As well as uniparental systems, autosomes show their higher variation in Africa. An analysis on 1327 nuclear microsatellite and insertion/deletion markers found the highest levels of within-population genetic diversity in African and African American populations (Tishkoff et al. 2009), and genetic diversity declining with distance from Africa, consistent with proposed serial founder effects resulting from the migration of modern humans out of Africa and across the globe (Rosenberg et al. 2002, 2006; Prugnolle et al. 2005; Ramachandran et al. 2005; Jakobsson et al. 2008). Whole genome data (The 1000 Genome Project Consortium 2010, 2012), also confirm this scenario, with higher variation being found in Africa. A study on morphometric cranial traits (Manica et al. 2007) also related the gradual loss of genetic diversity to the loss of phenotypic variation.

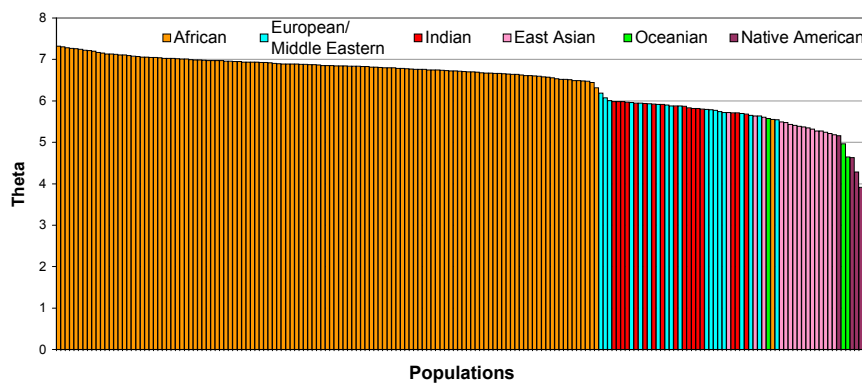


Figure 10: Genetic variation over different populations. Comparison of genetic diversity (Theta) from microsatellite allele size variance, over different populations (X-axis, each bar corresponding to a different population) (modified from Tishkoff et al. 2009).

Homo sapiens in the rest of the world: the Out of Africa

With the hypothesis of an African origin well established, it remains unclear when did the first human migratory waves leave Africa, and which route did they follow. Two main routes have been proposed: a northern route through Egypt to the Levant, (Stringer and Andrews 1988) , and a southern route through the Bab-el-Mandab strait and the Arabian Peninsula, towards the far South-Eastern Asia (Stringer 2000).

As for the timing of the exit from the African continent, two conflicting models have been put forward as well, with the first arguing that the first migrations took place as early as 120 kya, and the second supporting a much more recent exit from Africa, later than 70 kya. The two models are often referred to as “pre-Toba” and “post-Toba”, respectively, as the event taken as a landmark is the volcanic “supereruption” of the Mount Toba volcano (Sumatra) ~74,000 years ago (Chesner et al. 1991).

Palaeoanthropological and archaeological support to the first model include anatomically modern fossils dated prior to 100 kya found in Middle East (Grün et al. 2005) and tools found in Jebel Faya (Armitage et al. 2011), which dating is as old as 125 kya.

Conversely, supporters of a “late” exit argue that the pre-Toba colonization model proposes an extremely rapid, *in situ*, evolution from assemblages of technologically and typologically characteristically Middle Paleolithic form to dramatically different industries within a short time frame, with no convincingly documented occurrences of technologically intermediate or transitional industries. On the other hand, striking and detailed similarities between Asian industries and African findings dated between 65 and 55 kya are seen as an evidence for the “post-Toba” model (Mellars et al. 2013).

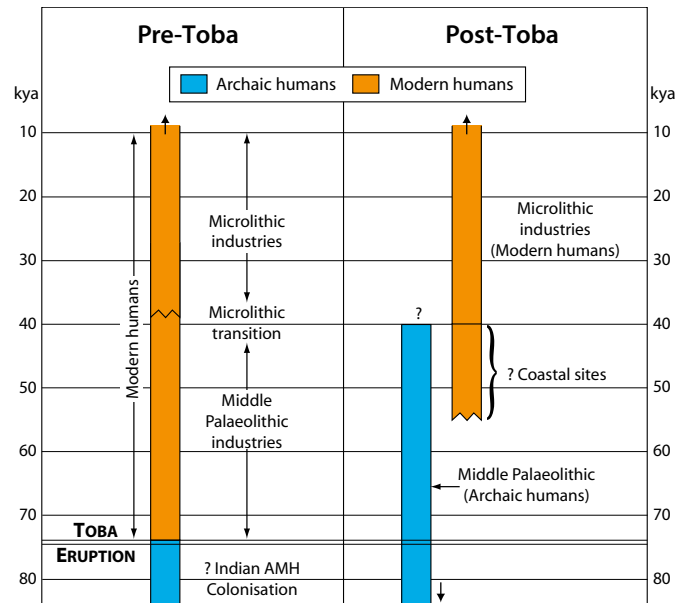


Figure 11: Comparison of the two alternative models for the initial modern human colonization of South Asia. The graphs show the inferred correlations between “archaic” and “modern” populations and their associations with Middle Paleolithic vs. microlithic technologies in the two models explained in the text. The date indicated for the initial modern human colonization in the “pre-Toba” model is a minimum number (~74 kya), with other estimates ranging up to 120 kya (from Mellars et al. 2013).

Data from the mtDNA diversity support a late, “post-Toba” exit, along the southern route (Macaulay et al. 2005; Thangaraj et al. 2005; Soares et al. 2011; Fernandes et al. 2012), not before 70 kya.

From the MSY perspective, Eastern and North-Eastern Africa are remarkably different. Some Eastern African lineages appear to be roughly as ancient as the first migrations out of the continent (haplogroups A-M32 and E-P2). The coalescence of all Eastern African haplogroups is also very close to that of the whole Y phylogeny. Lineages from North-Eastern Africa, instead, appear to be younger, as they share only terminal markers with sub-Saharan or Eastern lineages. Therefore, the MSY phylogeny is in accord

with mtDNA in pointing at Eastern Africa as the starting point for the first human migrations (Underhill and Kivisild 2007). We cannot exclude, however, that North-Eastern Africa was a corridor for more recent migrations from and to the Middle East (Underhill et al. 2001; Arredi et al. 2004; Luis et al. 2004; Semino et al. 2004).

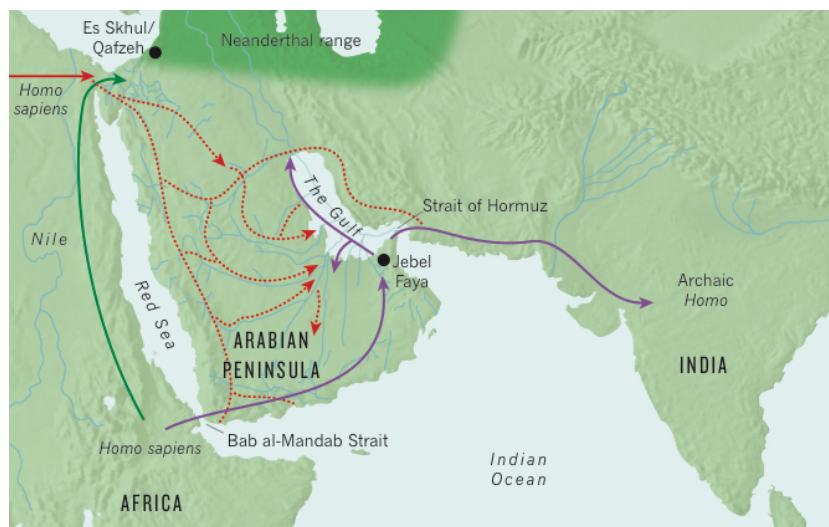


Figure 12: Earliest migrations out of Africa. The proposed routes for the earliest migrations out of Africa are shown. Earliest modern humans might have entered Arabia from the Sinai peninsula (green arrow), from East Africa through the Bab-el-Mandeb strait (purple arrow) or from North Africa (full red arrow), later dispersing over the Arabian peninsula following rivers (dotted red arrows). The positions of the earliest fossil remains out of Africa, either paleoanthropological (Es Skhul and Qafzeh) or archaeological (Jebel Faya) are shown (from Petraglia et al. 2011).

AIM OF THE STUDY

Despite the intense research carried on to date, dating estimates obtained from microsatellite and single nucleotide variation (Pritchard et al. 1999; Thomson et al. 2000; Wilder et al. 2004; Shi et al. 2010) produced a young MSY phylogeny, which provided limited information on the time horizon embracing the emergence of AMH in Africa and findings of early AMH outside Africa.

Recent updates to the MSY tree topology and new estimates for the substitution rate resulted in a much older dating for the root of the MSY tree (Cruciani et al. 2011a; Mendez et al. 2013) and nodes immediately downstream (Francalacci et al. 2013; Poznik et al. 2013). However, an unbiased search across widely divergent lineages with a low error rate in variant calling is still missing.

SNPs have always been the markers of choice to define the branches of the MSY tree (Underhill et al. 2000; Karafet et al. 2008), due to their evolutionary stability and low rate of recurrent mutations; the possibility of discovering a large number of SNPs employing next-generation sequencing technologies has also led to their re-evaluation as the optimal tool for age estimation. Recently, low-depth whole-genome sequencing studies produced thousands of MSY SNPs from a large set of males (The 1000 Genomes Project Consortium 2010, 2012). However, due to the likely abundance of false negatives and biases resulting from low-depth sequencing, a confident reconstruction and dating of the MSY phylogeny was difficult (Rocca et al. 2012). High-depth whole genome sequences have also been generated and made publicly available by Complete Genomics. MSY genotypes of 35 males from this dataset (together with a single haplogroup A3 subject as an outgroup) have been used to obtain a time-calibrated phylogeny of the MSY based on 6,662 high-confidence variants (Wei et al.

2013). However, due to the ancestry of males in the above study, deep branches of the MSY tree were underrepresented or not represented at all.

Sequencing of an array of lineages wide enough to encompass the majority of the deep variation of the MSY would allow a confident reconstruction of the MSY phylogeny. Therefore, we characterized by next generation sequencing 18 deep-rooting Y chromosomes, selected among thousands of worldwide Y chromosomes already genotyped for known markers (Cruciani et al. 2004, 2007, 2010, 2011a; Trombetta et al. 2011; Scozzari et al. 2012; present study). The 18 chromosomes were framed in a wider context of 50 more Y chromosomes, representing major branches of the entire MSY phylogeny. In this way, we were able to interpret our findings in light of the mutational pattern and branch divergence observed across the entire tree. All the 68 chromosomes were analysed by high coverage sequencing (average 50×), to avoid the occurrence of false positives and negatives, which constitute an intrinsic feature of low coverage NGS.

In this thesis, we will discuss the reconstructed MSY phylogeny, the associated time estimates produced, and its implications with the early evolutionary processes that involved our species. Causes for heterogeneity among branch lengths and the possible action of purifying selection will also be covered.

Incidentally, the selection of the 18 chromosomes required an accurate genotyping work, which independently led to a refinement of the deepest portion of the phylogeny, with the discovery of new markers and haplogroups, and the resolution of a deep trifurcation (Scozzari et al. 2012). These results will also be presented.

RESULTS

During the first part of the PhD, we focused on selecting the samples for the NGS experiment. This required a genotyping work, which also led to redefining the basal portion of the MSY phylogeny.

We then performed a high-depth resequencing (average 50×) of about 1.5 Mb of the MSY in 68 unrelated males representing major Y chromosome haplogroups. Here we will focus on the 18 deep-rooted lineages analysed in the experiment, limiting the discussion to the description of the results to single nucleotide substitutions.

Genotyping and phylogenetic mapping

Approximately 10,000 DNA samples already present in our laboratory collection were initially screened for this work, based on previous information; we selected 150 of them for further genotyping.

The Sorenson Molecular Genealogy Foundation database, containing DNA samples from approximately 40,000 individuals, was screened based on STR information; 24 samples were selected and provided to our laboratory to be added to the analysis. We also received 50 samples from the collection of Prof. Andrea Novelletto (University of Rome “Tor Vergata”).

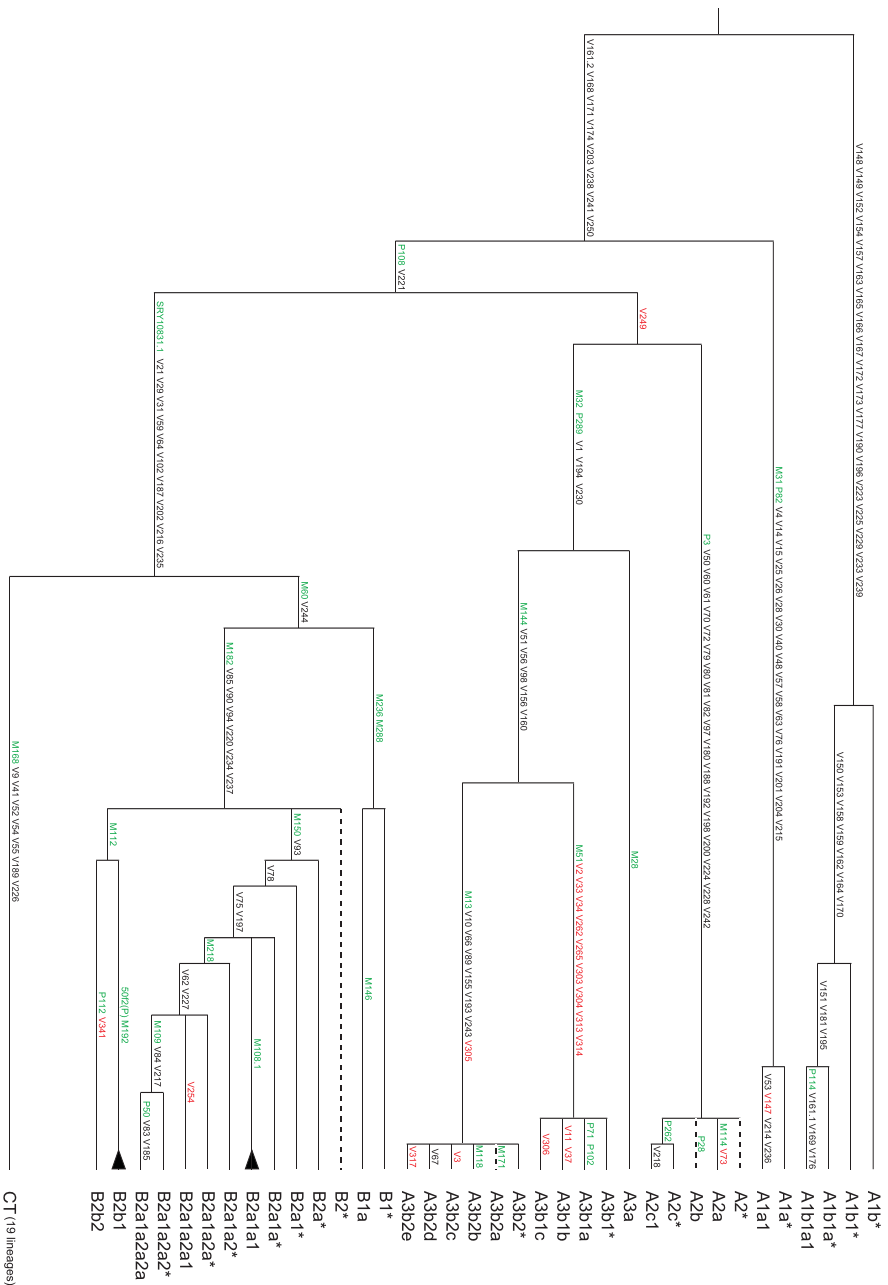
To better understand the phylogenetic relations among the chromosomes, and select a number of them for the following part of the study, we determined the allelic state at 168 previously published MSY polymorphisms; in addition, we sequenced about

90 kb for two chromosomes belonging to haplogroup A3b in order to clarify the internal structure of this clade.

The phylogenetic relations highlighted by this genotyping work are represented in Figures 13 and 14. 21 previously unreported markers, plus another already present in the dbSNP database (V306, corresponding to the SNP rs113042298), but not mapped in the phylogeny, were described. They are shown in red in figures 13 and 14 and information about them is given in Table 1.

SNP	Y-POSITION (hg19/GRCh37)	MUTATION
V2	6778215	A to C
V3	6778229	T to C
V11	6892902	A to T
V33	6894717	G to T
V34	6894718	C to T
V37	6818279	G to A
V73	16691696	A to G
V87	17947454	A to T
V147	6739492	G to A
V248	7589991	C to T
V249	25207704; 26841450; 27120952	T/T/T to G/G/G
V254	6870497	G to A
V262	6659209	C to G
V265	661164	A to G
V303	2798066 - 2798068	Del TTT
V304	2796955 - 2796957	Del GAA
V305	854573	T to C
V306	7594967	G to C
V313	7622390	G to A
V314	7642949	A to T
V317	2908553	G to A
V341	4840884	A to T

Table 1: Previously undescribed mutations. Information about the markers here described or analysed for the first time. For V249, three positions are indicated as PCR primers amplify three paralogous regions, with the T to G mutation in each of them.



(Previous page) **Figure 13: Revised topology of the deepest portion of the human MSY tree.** The mutations genotyped are indicated on the branches (green, mutations from Karafet et al. 2008; black, mutations from Cruciani et al. 2011a; red, previously undescribed mutations, see text). The internal structure of haplogroups B-M108.1 (2 branches) and B-50f2(P) (8 branches) is not shown (black triangles). Dashed lines indicate putative branches (no positive control available).

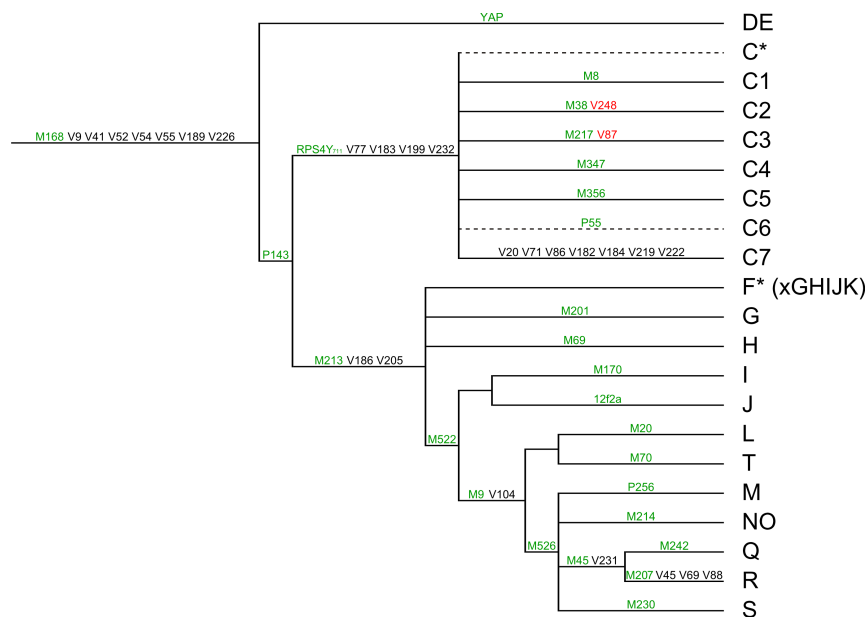


Figure 14: Structure of the macro-haplogroup CT. For details on mutations colour code see legend to Figure 13. Dashed lines indicate putative branches (no positive control available). The position of V248 (haplogroup C2) and V87 (haplogroup C3) compared to mutations that define internal branches was not determined.

Most of the mutations here analysed belong to the African portion of the MSY phylogeny, which is comprised of haplogroups A00 (absent in our study and not yet described at the time), A1b, A1a, A2, A3 and B. Through phylogenetic mapping it was possible to identify 15 new African haplogroups and to resolve one basal trifurcation (Figure 13). A new deep branch within the “out of Africa” haplogroup C was also identified (Figure 14).

Haplogroup A1b had already been identified as one of the two deepest-rooting branches of the MSY tree (Cruciani et al. 2011a). Internal structure was identified in this clade, with some chromosomes showing the ancestral allele for marker P114, which previously identified the entire haplogroup (Karafet et al. 2008). The splits inside the haplogroup, however, appear to be terminal, indicating recent differentiation. We identified a third allele (A) for the V161 polymorphism, which had been previously reported as a biallelic G>C transversion on the A1b branch (Cruciani et al. 2011a). The presence of nucleotide A at the orthologous MSY position in the chimp reference sequence (October 2010 chimp assembly, UCSC Genome Browser), along with the structure of the human MSY tree as shown in Figure 13, suggest that this triallelic polymorphism may have originated from two independent mutations, A>C (V161.1) within the A1b branch, and A>G (V161.2) at the root of macro-haplogroup A1a-T (Figure 13).

Haplogroup A1a was represented by two samples, which were typed for 22 previously described markers (Karafet et al. 2008; Cruciani et al. 2011a), plus another found in this study (V147). This analysis led to the splitting of A1a into two branches that shared the derived state for all but four of the 23 markers analysed (Figure 13).

Haplogroups A2 and A3 were confirmed to be sister clades, as previously described (Batini et al. 2011). P3 and 17 more markers previously defined haplogroup A2 (Karafet et al. 2008), with three mutations defining three terminal branches (A-M114, A-P28 and A-P262). Here, 19 mutations that had been identified for haplogroup A2 (Cruciani et al. 2011a), as well as the markers P3, M114, P28, P262, were genotyped in three A2 Y chromosomes. This analysis made it possible to identify a new branch (A-V218) within haplogroup A-P262. The newly reported V73 mutation was found to be phylogenetically equivalent to M114 (Figure 13).

Haplogroup A3 confirmed his previously known structure, containing the two sister clades A3a and A3b, with A3b further

subdivided into a Southern African (A3b1 or A-M51) and a mainly Eastern African (A3b2 or A-M13) haplogroup (Underhill et al. 2000; Cruciani et al. 2002; Semino et al. 2002; Wood et al. 2005; Hassan et al. 2008; Naidoo et al. 2010; Batini et al. 2011). Since both A3b1 and A3b2 are quite common, but are yet poorly resolved haplogroups, we performed a MSY re-sequencing analysis of about 90 kb for each of two A3b chromosomes (one A-M51* and one A-M13*) to find additional informative markers. We detected a total of 9 new mutations (V262-V317 in Table 1). A total of 41 markers (9 new mutations, 7 mutations from our database, 15 mutations identified in a previous study (Cruciani et al. 2011a), and 10 mutations defining A3 branches (Karafet et al. 2008) were analyzed in ten subjects. The phylogenetic mapping of the other mutations led to the identification of five new haplogroups, doubling the number of both A3b1 and A3b2 terminal branches. Finally, the P289 marker (Karafet et al. 2008) was positioned upstream of both A3a and A3b (Figure 13).

Haplogroup B shows well-defined and restricted geographic and ethnic distributions, with the exception of the widespread B-M109 lineage (Cruciani et al. 2002, 2011a; Semino et al. 2002; Mishmar et al. 2003; Karafet et al. 2008; Gomes et al. 2010; Naidoo et al. 2010; The 1000 Genomes Project Consortium 2010; Batini et al. 2011). Extensive sequencing of one B-M109 chromosome recently led to the identification of 17 mutations for this haplogroup (Cruciani et al. 2011a). Here, a total of 33 mutations, 31 of which previously described (Karafet et al. 2008; Cruciani et al. 2011a), and 2 (V254 and V341) identified during the mapping process, were analysed in 13 haplogroup B chromosomes. We substantially increased the resolution of the B2a clade (B-M150), with five new branches detected. The previously reported trifurcation B2a/B2b/B2c within the major clade B2 (Karafet et al. 2008) was resolved by repositioning the M112 mutation (Figure 13).

All non African-exclusive Y-clades belong to the macro-haplogroup C-T, which is defined by mutations M168, M294 and P9.1 (Underhill et al. 2001; Karafet et al. 2008) and subdivided in

the two main clades D-E and C-F (Underhill and Kivisild 2007; Karafet et al. 2008). In a previous study, we identified 25 new mutations in clade C-T, with eleven of them shared by chromosomes of haplogroups C and R, seven specific to haplogroup C and seven specific to haplogroup R (Cruciani et al. 2011a). Here, the CR shared mutations were also found to be present in one D-E sample, and positioned at the root of macro-haplogroup C-T. The seven C-specific markers were typed, along with mutations defining branches C1 to C6 (Karafet et al. 2008) in six haplogroup C chromosomes. Through this analysis we identified a chromosome from Southern Europe as a new deep branch within haplogroup C (C-V20 or C7, Figure 14). As haplogroup C is quite rare in Southern Europe (Semino et al. 2000; Battaglia et al. 2009), we then surveyed 1,965 European subjects identifying one additional haplogroup C chromosome from Southern Europe, which has also been classified as C7 (data not shown). Further studies are needed to establish whether C7 chromosomes are the relics of an ancient European gene pool or the signal of a recent geographical spread from Asia. Two previously undescribed mutations, V248 and V87, were found to be specific to haplogroups C2 and C3, respectively (Figure S1).

Three of the seven R-specific mutations (V45, V69 and V88) were previously mapped within haplogroup R (Cruciani et al. 2010a); here, we were able to position at the root of haplogroups F (V186 and V205), K (V104) and P (V231) (Figure 14) through the analysis of 12 haplogroup F samples.

Single nucleotide mutational pattern

Based on phylogenetic mapping, we selected 68 unrelated males (Table 2) for further sequencing. We performed a high-depth (50×) resequencing of five blocks, totalling 1,495,512 bases (~1.5 Mb) of

the MSY, and excluding repetitive elements from the analysis (see Methods section). We identified 2,386 positions displaying a nucleotide substitution among the 68 chromosomes under study (see Appendix 1); two positions were found to be triallelic.

Thirteen positions, which were invariant in the entire sample but different from the reference sequence, were not considered for further analyses. These positions can be interpreted as either reference-specific mutations or sequencing errors.

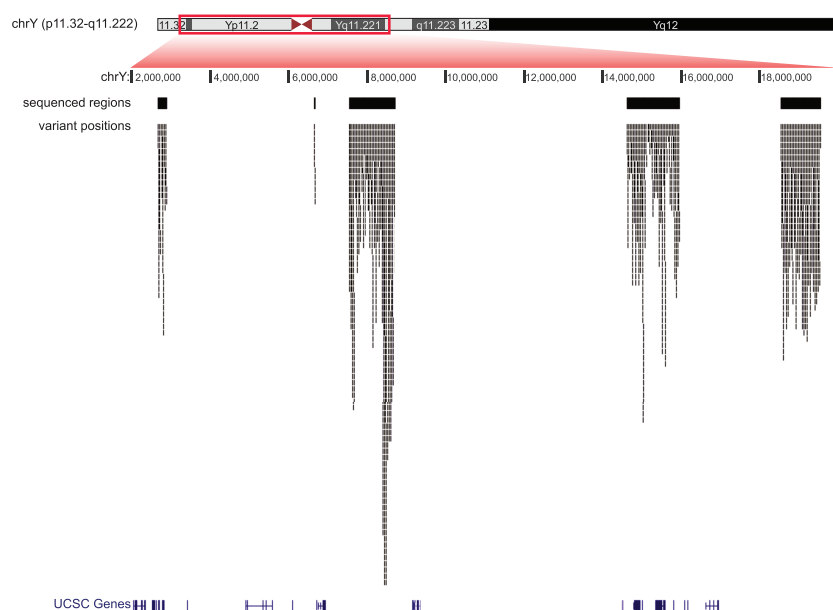


Figure 15: Regions of the Y chromosome analyzed and distribution of variants discovered. The different tracks from top to bottom report the following features: Y chromosome ideogram; Y chromosome position according to the human Y chromosome reference GRCh37/hg19; the five regions targeted for capture (black bars); variant positions discovered (thin marks); UCSC genes.

Sample	Haplogroup		Region	Country	Population	Reference
	by lineage	by mutation				
S01	A1b*	A-V148*	Central America	Colombia	General population	1
S03	A1b1*	A-V150*	Western Africa	Ghana	General population	2
S04	A1b1a*	A-V195*(R114)	Central Africa	Cameroun	General population	1
S05	A1b1a*	A-V195*(R114)	Northern America	USA	General population	1
S07	A1a*	A-M31*(R14)	Western Africa	Mali	General population	1
S08	A2a	A-M114	Southern Africa	Republic of South Africa	Kung	2,3
S09	A3b1b	A-V11	Southern Africa	Republic of South Africa	Khwe	2,3
TV20	A3b2*	A-M13*	Europe	Greece	General population	1
S10	A3b2*	A-M13*	Europe	Italy	Sardinian	1
S11	A3b2*	A-M13*	Eastern Africa	Kenya	Maasai	1
S12	A3b2c	A-V3	Eastern Africa	Ethiopia	Ethiopian Jews	1,3
S13	B1*	B-M236*	Central Africa	Cameroun	Bantleke	3
S14	B1a	B-M146	Western Africa	Burkina Faso	General population	2
S15	B2a*	B-M150*	Western Africa	Mali	General population	2
S16	B2a1a2a2*	B-M109*	Central Africa	Cameroun	Fall	2
S17	B2b1	B-P6	Southern Africa	Republic of South Africa	Kung	1,3
S18	B2b*	B-M112*(R16,P7)	Southern Africa	Republic of South Africa	Kung	1,3
S19	B2b3*	B-M30	Central Africa	Cameroun	Nganbai	3
S38	C7	C-V20	Europe	Italy	General population	2,4
18 samples	E (includes E1 and E2)	E-M40	Africa, Europe, Asia, and America	Various countries	General population	1,3,5,6,7
3 samples	I2a	E-P37,2	Europe	Various countries	General population	1
28 samples	P (includes Q and R)	P-M45	Africa, Europe, Asia, and America	Various countries	General population	1,4,8

Table 2: Subject selected for the NGS analysis. Haplogroup nomenclature according to Scozzari et al. 2012, Batini et al. 2011, and Karafet et al. 2008.

References: (1) present study; (2) Scozzari et al. 2002; (3) Cruciani et al. 2002; (4) Cruciani et al. 2011a; (5) Cruciani et al. 2004; (6) Cruciani et al. 2007; (7) Trombetta et al. 2011; (8) Cruciani et al. 2010a. All DNA samples are from blood or saliva, except four haplogroup E DNAs and one haplogroup P DNA, which are from cell culture.

Figure 15 shows the distribution of the variant positions across the five selected MSY regions. The apparent uneven density of variant positions can be explained by the different occurrence of repetitive elements, which were largely excluded from targeted sequences (see Methods section). When the occurrence of the 2,386 variants in each of the 5,274 sequenced DNA fragments (see Methods section) was considered, no evidence of any uneven distribution was obtained. The linear slope (0.001591) closely matched the overall rate of occurrence ($2,386/1,495,512 = 0.001595$) and all points but two fell within the 0.999 confidence interval estimated according to the Poisson distribution (Figure 16). Of the 2,386 positions, 12.1% were inferred to be located at ancestral CpG dinucleotides (see Methods), a proportion similar to that reported by previous studies (Kong et al. 2012).

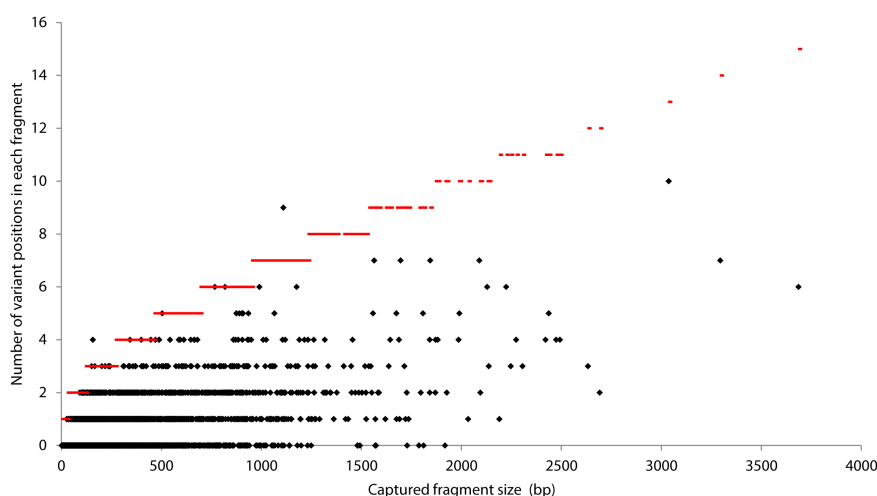


Figure 16: Scatterplot of number of variable positions as a function of captured fragment length. Red bars indicate the upper boundary of the 99.9% C.I. for the expected number of mutations (using Poisson distributions with means equal to the expected number of variants given the fragments length).

Six protein-coding genes (*RPS4Y1*, *ZFY*, *USP9Y*, *DDX3Y*, *UTY* and *TMSB4Y*) were covered in our capture design. 19 variant positions were located within codons (Table 3). Overall, the number of variant positions we found in coding regions (19/15,397 bases) was proportionally lower than that residing in non-coding regions (2,367/1,480,115), though not significantly ($P = 0.30$, Fisher exact test).

Mutations in coding regions are more likely to be negatively selected when compared to neutral non-coding mutations. Therefore, mutations in coding regions observed today can be expected to be younger than neutral non-coding mutations. To prioritize the detection of older mutations, we restricted our analysis on mutations in older branches, setting a threshold of 30 kya (kilo-years ago). No evidence indicating any enrichment of non-coding variants was obtained. In our data, 10 of the substitutions were predicted to produce amino acid changes, of which 8 were private (i.e. found in a subject only) and 2 were shared between at least two subjects. This compares with 3 private and 6 shared synonymous variants. A similar imbalance was present after removing four long terminal branches, each of them represented by a single individual (S07, S08, S09 and S38, see Figure 17), in which some of the private mutations can be relatively old (7 private vs. 2 shared and 3 private vs. 5 shared for non-synonymous and synonymous mutations, respectively). Though these differences were not nominally significant (Fisher exact test, $P = 0.07$, and $P = 0.15$, respectively), they suggested that purifying selection might have caused the slight underrepresentation of missense variants we observed in shared lineages. Two private previously unreported variants (chrY:14834046 G>A, *USP9Y* R84Q; chrY:14870505 G>T, *USP9Y* V567L) were predicted to be conserved damaging (PolyPhen2 scores 0.927 and 0.955, respectively), but also conserved (PhyloP scores 0.974 and 0.983, respectively). Both fall in the *USP9Y* gene, a member of the peptidase C19 family.

Y-position (GRCh37/hg19)	Gene	Reference allele	Alternative allele	Class of variant	Amino acid change	PhyloP score	PhyloP prediction	PolyPhen2 score	PolyPhen2 prediction	Shared
2710013	<i>RPS4Y</i>	A	G	nonsynonymous	K10R					no
2734854	<i>RPS4Y</i>	C	T	synonymous	S237S					no
14832620	<i>USP9Y</i>	G	T	nonsynonymous	E65D	0.972962	Conserved	0.011	Benign	yes
14834046	<i>USP9Y</i>	G	A	nonsynonymous	R84Q	0.974416	Conserved	0.955	Probably damaging	no
14851554	<i>USP9Y</i>	T	C	synonymous	D471D					yes
14870505	<i>USP9Y</i>	G	T	nonsynonymous	V567L	0.983019	Conserved	0.927	Probably damaging	no
14888783	<i>USP9Y</i>	A	C	synonymous	L876L					yes
14898546	<i>USP9Y</i>	C	A	nonsynonymous	P1125H	0.974416	Conserved	0.0	Benign	no
14902417	<i>USP9Y</i>	C	T	synonymous	C1213C					no
14952467	<i>USP9Y</i>	A	G	synonymous	Q2005Q					no
15026544	<i>DDX3Y</i>	G	T	nonsynonymous	A64S	0.108136	Non-conserved	0.0	Benign	no
15027863	<i>DDX3Y</i>	G	A	synonymous	E227E					yes
15028176	<i>DDX3Y</i>	G	C	nonsynonymous	S254T	0.975056	Conserved	0.0	Benign	yes
15028931	<i>DDX3Y</i>	T	C	synonymous	Y390Y					no
15435514	<i>UTY</i>	C	T	nonsynonymous	R451H	0.838203	Non-conserved	0.0060	Benign	no
15467824	<i>UTY</i>	A	G	synonymous	S114S					yes
15591193	<i>UTY</i>	G	A	nonsynonymous	L53F	0.781742	Non-conserved	0.028	Benign	no
15591445	<i>UTY</i>	T	C	nonsynonymous	E34G	0.041684	Non-conserved	0.0010	Benign	no
15591537	<i>UTY</i>	G	C	synonymous	S3S					yes

Table 3: Mutations in protein coding regions found in the present study. We consider a mutation to be "shared" if the alternative allele is found in two or more individuals.

MSY phylogeny

We used the 2,386 variable positions to reconstruct a maximum parsimony tree using two independent methods (see Methods). These methods yielded trees with identical topologies, with substitutions at the same positions in each branch, and indicated recurrent mutational events in four positions, for a total of 2,392 distinct mutational events (including double hits at the two triallelic positions). The proportion of recurrent events (4/2,386 positions, 0.2%) was significantly lower ($P = 2.2 \times 10^{-16}$, Fisher exact test) than that reported in a recent comparable study (Wei et al. 2013) (172/5,865 mutations, 2.9%), a discordance that can be attributed to differences between the two studies in both the regions analysed and in the strategies adopted to infer the ancestral states (see Methods). We were able to determine ancestral and derived state for 2,356 mutational events. The overall transition/transversion ratio ($1,513/879 = 1.72$) was within the range of genome-wide estimates for *de novo* events (Campbell et al. 2012; Kong et al. 2012; Michaelson et al. 2012) with an excess of G>A and C>T compared to the opposite changes (Table 4).

Ancestral allele	Derived allele				Total
	A	C	G	T	
A		106	329	87	522
C	151		106	384	641
G	450	115		113	678
T	83	328	104		515
Total	684	549	539	584	2356

Table 4: Summary of the 2,356 mutational events for which ancestral/derived alleles were unequivocally determined.

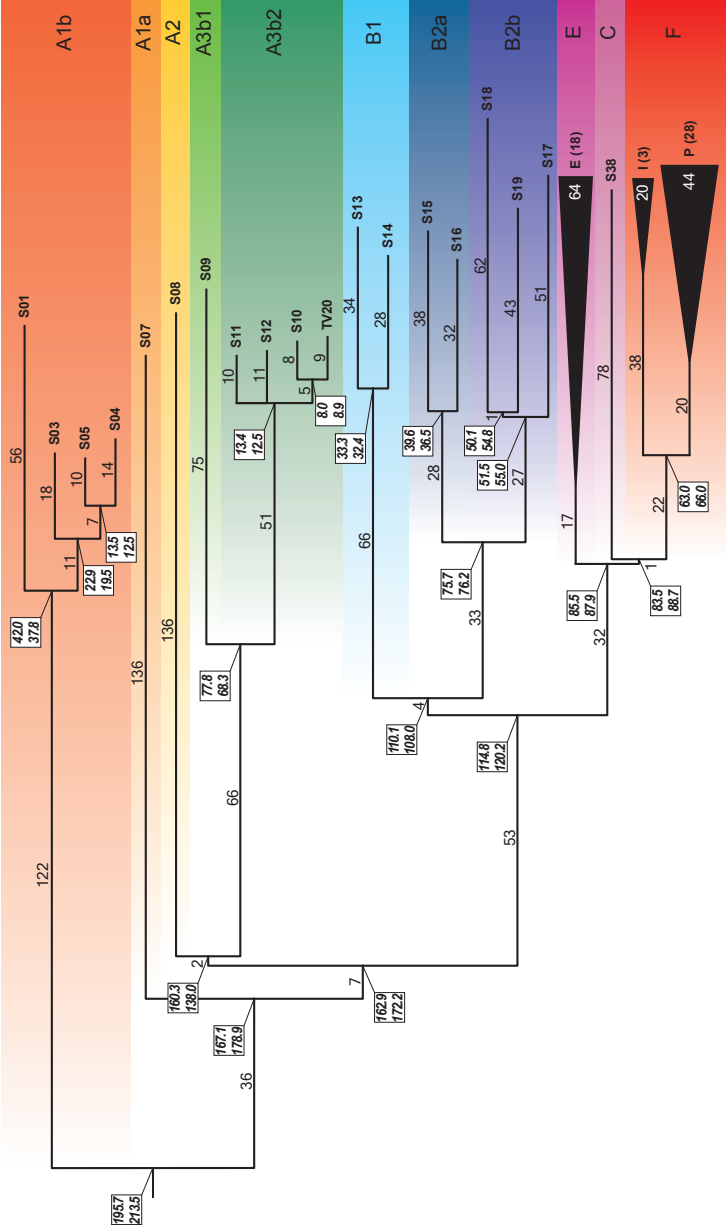


Figure 17: Maximum parsimony tree obtained with 2,386 variable positions. The number of mutational events defining each branch is reported above or near it. For the collapsed haplogroups E, I and P, the average number of mutations is shown. Dating estimates are reported in boxes near each node (upper and lower values obtained with BEAST and the rho method, respectively). Coloured belts indicate major haplogroups according to current nomenclature (Karafet et al. 2008; Scozzari et al. 2012).

Figure 17 shows a condensed version of the maximum parsimony tree, with emphasis on the deeply rooted African lineages. Chromosomes previously known to belong to different major haplogroups partitioned into distinct clades in the tree, with the same phyletic relationships reported in previous studies (Karafet et al. 2008; Batini et al. 2011; Cruciani et al. 2011a; Scozzari et al. 2012; Francalacci et al. 2013; Poznik et al. 2013) and highlighted during the typing carried on during the sample selection (see “Genotyping and phylogenetic mapping” subsection). The polyphyletic nature of “haplogroup” A was confirmed (Cruciani et al. 2011a), with A1b being the most deeply rooted clade in our set, followed by A1a. Markers for each of four A1b lineages were discovered. Within A2-F, a major bifurcation grouped A2 and A3 together, which stemmed from a short branch (branch 11 in Figure 18), previously defined by markers PK1 (Batini et al. 2011) and V249 (Scozzari et al. 2012 and “Genotyping and phylogenetic mapping” subsection). Within A3b2, a small clade grouped together two European A-M13* subjects (S10 and TV20) which differed by 17 mutations, and separated them from two African sub-haplogroups. Such loose affinity between the two European A-M13 chromosomes denotes a more remote relatedness than that recently reported for seven A-M13 chromosomes from Sardinia (Francalacci et al. 2013).

The other samples, belonging to haplogroup B-F, shared a long branch (branch 21), with B as a monophyletic clade sister to E-F. Haplogroup B was confirmed to consist of two deep clades, corresponding to B1 and B2, with the latter in turn consisting of B2a and B2b. Compared to the previous topology (Scozzari et al. 2012), we found markers for each of the paragroups B1*, B2a* and B2b*. While a remarkable advancement in the phylogenetic structure of haplogroup B2b was obtained by a recent study (Poznik et al. 2013), we detected a new haplogroup-defining node for B2a, which is deeper than that reported in previous studies. The remaining haplogroups (E, C, and F) were arranged according to the previously known topology (Karafet et al. 2008). In particular a

single mutation (branch 37 in Figure 18), which is phylogenetically equivalent to P143, defines a sister clade of E comprising haplogroups F and C, the latter of which has never been covered in other large-scale resequencing studies (Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013).

A remarkable aspect of our tree was that the relative lengths of major branches (in number of mutations) differed greatly from previously reported values. Some striking examples include branches 1, 9 and 23 (defining A1b, A1a and B1, respectively), in which the number of mutations increased by at least 6 times compared to previous studies (Karafet et al. 2008; Scozzari et al. 2012). A notable increase in length was also observed for branches 21 and 35 (defining BF and EF, respectively). Conversely, basal branches of haplogroup P did not show the same increase in length as compared to previously known markers. This can be partially attributed to the asymmetry involved when using a sequence that is mainly derived from haplogroup P DNA as a reference, and to the intense sequencing and search for mutations carried out on haplogroup P subjects (Underhill et al. 2010; Myres et al. 2011).

In order to further confirm the increase in branch length compared to previous studies, we considered dbSNP (build 135) as an alternative source of variants. Only 407 of the 2,386 variant positions (17.1%) here detected were reported in dbSNP (see Appendix 1 and Figure 19). dbSNP polymorphisms were underrepresented in deep-rooting African-specific branches of the phylogeny and in haplogroup C. A paucity of known SNPs was evident for 17 of the 18 terminal branches of African-specific haplogroups. Conversely, dbSNP markers almost saturated branches leading to haplogroups E and F.

These results dramatically modified a major feature of the tree, i.e. the proportions of mutations that mark the phylogeny for the periods prior to and after the exit out of Africa. Haplogroup E-F is informative for this event because it includes, among others, all the lineages found out of Africa (Underhill and Kivisild 2007).

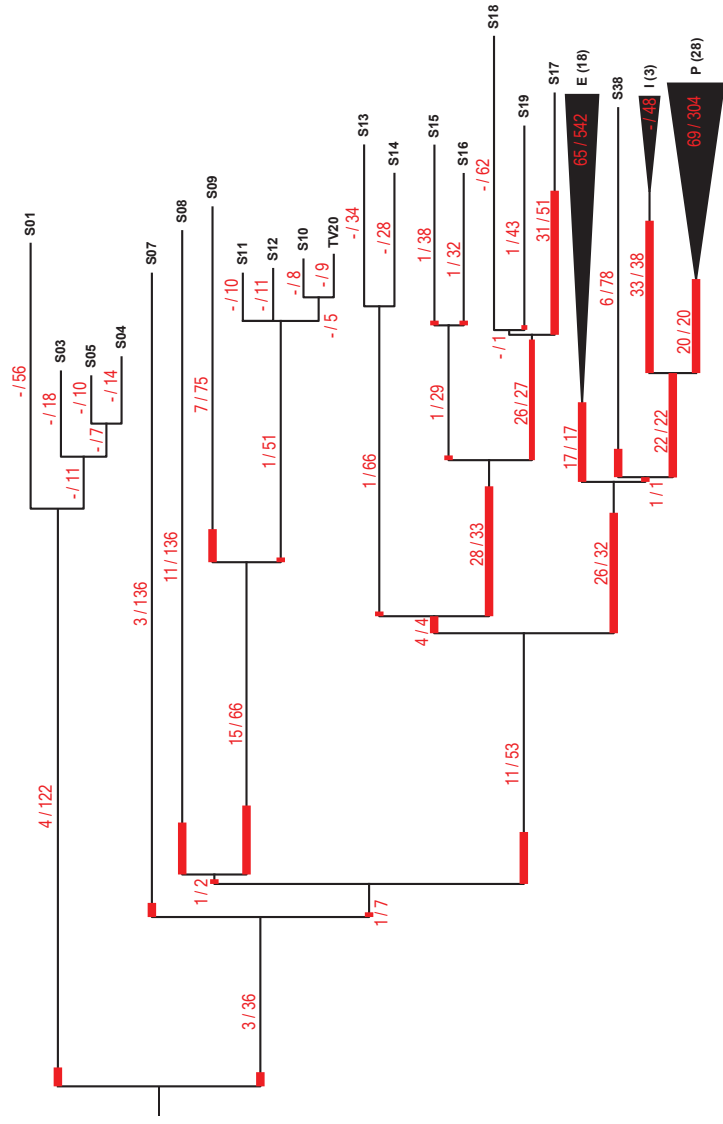


Figure 19: Comparison between markers here described and dbSNP database. Same tree as in Figure 17, showing the number of previously known SNPs (dbSNP, build 135) in each branch (in red, before the slash) as compared to the number of SNPs described in this work (in red, after the slash). This fraction is also represented by the red bar on each branch.

The previous tree (Karafet et al. 2008) depicted a close proximity between the root and the MRCA of E-F. Conversely, in our tree, 128 mutations separated the root from the node defining E-F, whereas 84.2 mutations per lineage (on average) were downstream of haplogroup E-F.

While our results corrected an evident under-detection of variants for deep-rooted branches, a short length of haplogroups A1b, A1a and A2-A3 (158, 172 and 177 average mutations from the root, respectively) compared to the rest of the tree (haplogroup B-F, 211 average mutations from the root) was nevertheless apparent. This added to the findings of a recent study (Wei et al. 2013) of a short A3 branch. We tested whether mutation rate heterogeneity among branches of the entire tree could explain the data better than a strict clock model. We compared the distributions of tree log(likelihoods) generated by BEAST (Drummond and Rambaut 2007) under both models and found that the difference between the harmonic means was 2.3, corresponding to positive evidence (Nylander et al. 2004) in favour of rate heterogeneity. Thus, comparing our findings to two recent large screenings (Francalacci et al. 2013; Poznik et al. 2013), deep branches of the Y phylogeny reveal an appreciable heterogeneity in the accumulation of mutations. In particular, in our tree, A1b was by far the shortest branch (158 average mutations from the root). When the length of A1b was compared with the rest of the tree (A1a-F, 207 average mutations from the root), the difference turned out to be statistically significant ($\chi^2 = 6.72$, $P = 0.0095$). The corresponding tests for A1a (vs. A2-F) and A2-A3 (vs. B-F) produced nominally significant P values, which however did not resist the Bonferroni correction for multiple tests.

Besides rate heterogeneity, large structural rearrangements could be responsible for the reduction in the number of countable positions. We then investigated whether large deletions were present in our dataset, but on the short branches (A1b, A1a and A2-A3) we only detected a 6.2 kb deletion (0.42% of the total

sequence) shared by all A1b chromosomes, unlikely to cause any imbalance in countable positions.

Another hint of structural rearrangements could be the clustering of variants, caused by misalignment of the NGS reads with paralogous sequences. However, when counting the instances of two variants within 100 bp and the instances of two, three and four variants within 1000 bp, we had no evidence of excess (Table 5).

Tree branch	Number of variants	Instances of two variants within 100 bp	Instances of two variants within 1000 bp	Instances of three variants within 1000 bp	Instances of four variants within 1000 bp
1	94	2	6	1	0
2	56	1	4	0	0
3	11	0	0	0	0
4	18	0	2	0	0
5	7	0	0	0	0
6	10	0	0	0	0
7	14	0	0	0	0
8	28	0	0	0	0
9	136	2	6	0	0
10	7	0	0	0	0
11	2	0	0	0	0
12	136	2	10	1	0
13	66	1	1	0	0
14	75	1	2	0	0
15	51	0	0	0	0
16	10	0	0	0	0
17	11	0	0	0	0
18	5	0	0	0	0
19	8	0	0	0	0
20	9	0	0	0	0
21	53	0	1	0	0
22	4	0	0	0	0
23	66	0	1	0	0
24	34	0	0	0	0
25	28	0	0	0	0
26	33	0	2	0	0
27	28	0	0	0	0
28	38	1	2	0	0
29	32	0	0	0	0
30	27	0	0	0	0
31	1	0	0	0	0
32	62	0	2	0	0
33	43	0	1	0	0
34	51	0	1	0	0
35	32	0	1	0	0
36	17	0	1	0	0
37	1	0	0	0	0
38	78	1	3	0	0
39	22	0	0	0	0
40	38	0	0	0	0
41	20	0	0	0	0
1/8	36	1	6	0	0
36-internal	542	5	9	2	1
40-internal	48	0	0	0	0
41-internal	304	2	6	0	0
Total	2392	19	67	4	1

Table 5: Clusters of variants observed in the present study. Tree nomenclature as in Figure 18. Tree branch 1/8 indicates mutations with unknown ancestral state (see Appendix 1 and Tree construction section in Material and Methods).

Dating and phylogeography

We used sequence data and two independent methods to estimate the age of the nodes in the tree. In both methods, we used a substitution rate obtained by adjusting the rate of autosomal *de novo* mutations from recent genome-wide screens to the MSY, as independently worked out in a recent study (Mendez et al. 2013) with minor differences (see Methods). The results are shown as boxes in Figure 17. The two methods produced highly concordant values (Figure 20). Hereafter we refer to the results obtained with BEAST (Drummond and Rambaut 2007), which averages the influence of many parameters over the entire tree, also accounting for rate heterogeneity among branches. The consensus tree showing the node ages with associated confidence intervals is reported in Figure 21.

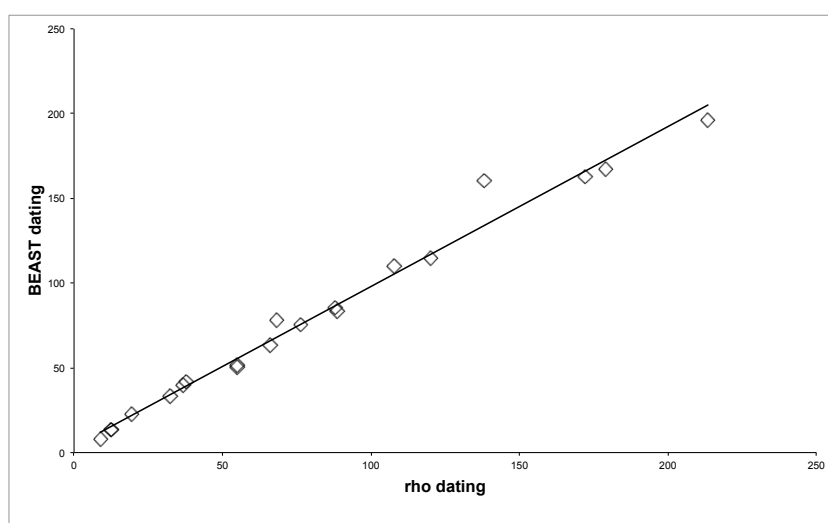


Figure 20: Comparison between two dating methods. Scatter plot of ages (kya) for the 20 nodes shown in Figure 17 estimated with rho (X-axis) and BEAST (Y-axis).

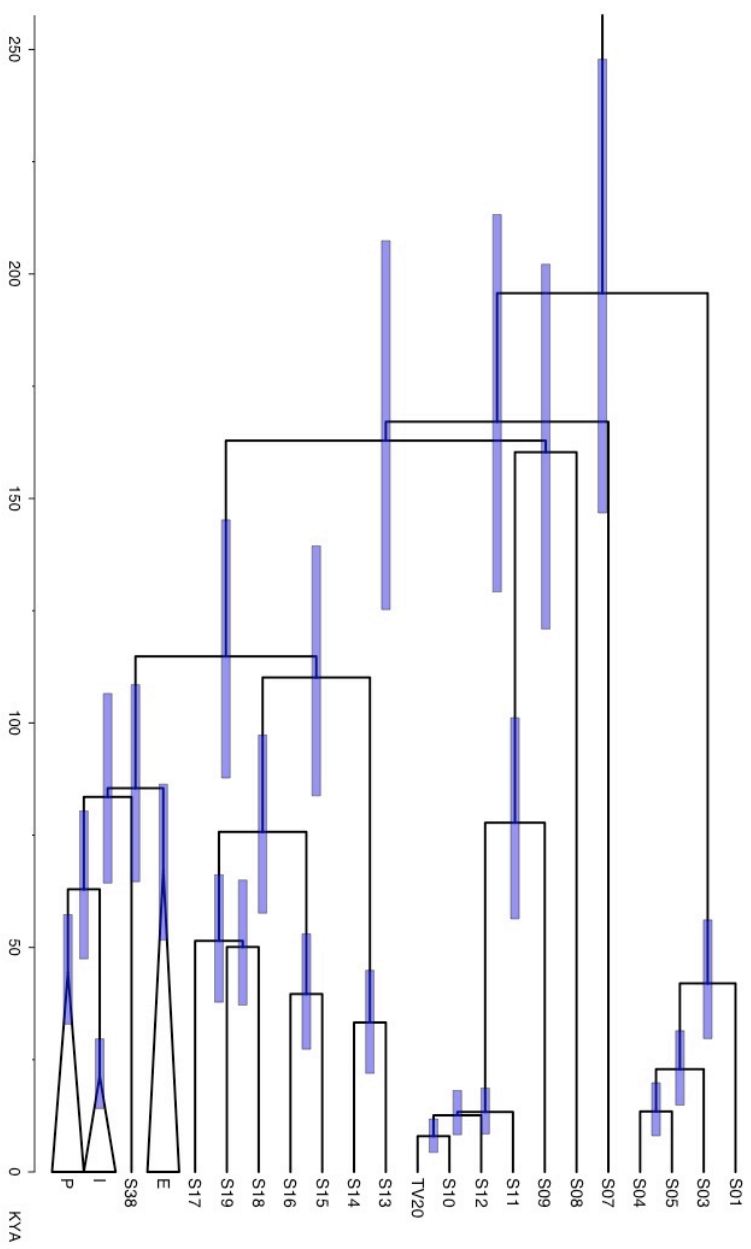


Figure 21: Y chromosome tree obtained with BEAST. The estimated age of nodes is shown, with branch tips aligned to present day. X-axis reports the age from present in thousand years (Kya). Blue bars indicate the 95% HPD. Note that the software resolves the multifurcation of branches 16-18 (Figure 18) into two consecutive bifurcations.

The TMRCA of the samples here examined [equivalent to haplogroup A0-T in the nomenclature of Mendez et al. (2013)] was estimated at 196 kya (95% C.I. 147-248 kya), in agreement with the value obtained by Mendez et al. (2013) with a different method. Our estimate was much older than the previous one based on a similar topology but a different substitution rate (Cruciani et al. 2011a). Three nodes basal to A1a-F, A2-F and A2-A3 clustered in the narrow interval between 167 and 160 kya. The node basal to A2-F coincides with the MRCA in the datasets by Francalacci et al. (2013) and Poznik et al. (2013) and our estimated date (162.9 kya; 95% C.I. 125-207 kya) is comparable to both studies (185 and 138 kya, respectively). We observed no more nodes until 115 kya, a date which marks the separation of African-specific haplogroups from the rest of the phylogeny. An age of as much as 110 kya was estimated for haplogroup B, corresponding to the split between chromosomes currently found only in Central-Western Africa (B1) and chromosomes spread all over sub-Saharan Africa (B2) (Figure 3 and Table 6). Such an old date could not be highlighted in recent large-scale resequencing studies (Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013), due to the lack of B1 representatives.

In the time frame between 85.5 and 75.7 kya, four splits were observed: (1) the node within haplogroup A3b, which separates southern (A3b1) from eastern (A3b2) African lineages; (2) the node within haplogroup B2, separating clades B2a and B2b which are frequently observed among present day African food-producers and hunter-gatherers, respectively; and (3) two nodes that are highly informative for the exit out of Africa which are basal to E-F and C-F, respectively. In fact, haplogroup E has representatives both within and out of Africa, whereas haplogroup C-F (83.5 kya, C.I. 64.3-106.6 kya) encompasses chromosomes found virtually only outside of Africa.

We used two discrete phylogeographic analyses (Lemey et al. 2009; Yu et al. 2010) to associate each node of the tree to each of four broad geographic regions, with emphasis on sub-Saharan Africa. First, we used a Bayesian analysis (Figure 22), in which,

when starting with an even prior, the posterior probabilities (0.44-0.45) favoured a Central-Western African placement for the four deepest nodes in the tree, i.e. from 196 to 160 kya. Southern and eastern African locations were favoured for the nodes defining haplogroups A3b and A3b2, respectively. The emergence of new diversity out of Africa was captured in this analysis by a shift in location assignment along the branch leading to E-F, with all nodes downstream assigned to non-sub-Saharan African locations with high confidence (Figure 22). Finally, a further shift in location assignments was observed within haplogroup A3b2, from Eastern Africa to non-sub-Saharan Africa. Second, we used a maximum parsimony approach (Yu et al. 2010), which similarly predicted a 100% probability of a central-western African location for the two deepest nodes. In this analysis, however, the oldest node unambiguously assigned to non-sub-Saharan African locations was the MRCA of haplogroup C-F (Figure 23).

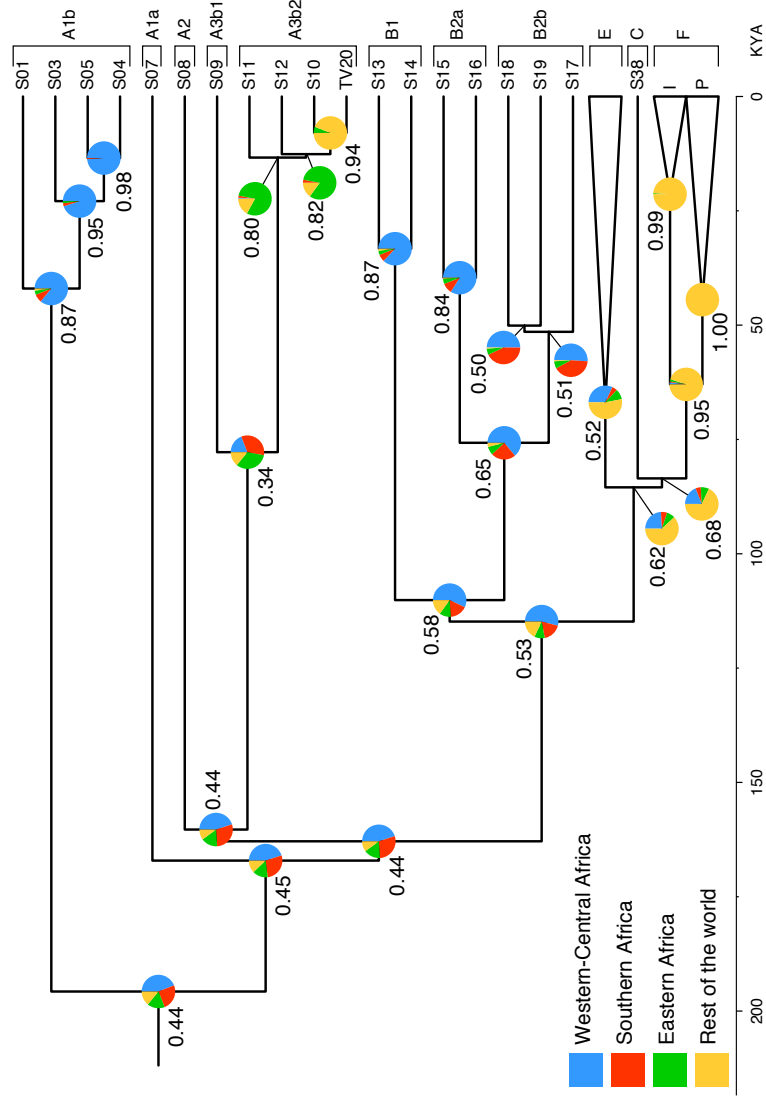


Figure 22: Results of the phylogeographic analysis using a Bayesian method as implemented in BEAST. Pie charts reflect the probability of the respective geographic areas, as coded in legend, with the greater probability explicitly indicated.

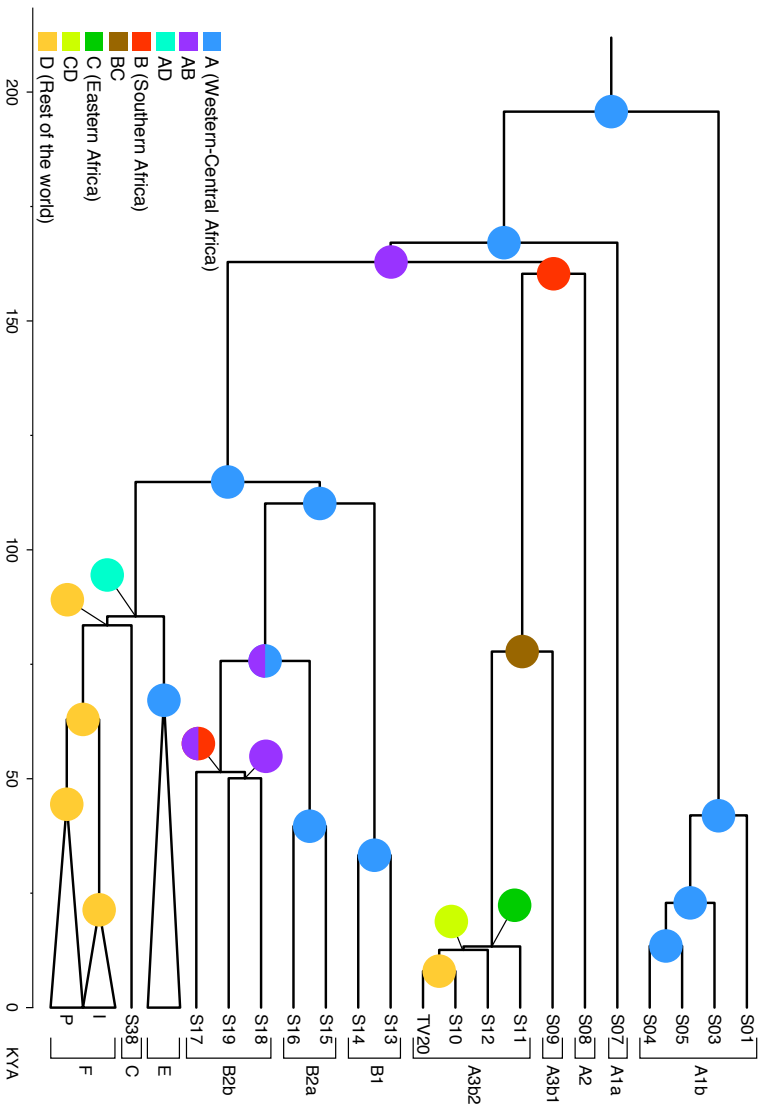


Figure 23: Results of the phylogeographic analysis using a parsimony-based method (S-DIVA) as implemented in RASP. Pie charts reflect the probability of the respective geographic areas, as coded in legend, allowing for single or combined distributions.

Distribution of the main Y haplogroups in Africa

The distribution of the main deep lineages of the Y phylogeny in 73 populations from different regions of the African continent is reported in Table 6.

Both A1b and A1a are extremely rare lineages. Haplogroup A1b was also found in one subject from Ghana (ethnic affiliation and haplogroup frequency in the population not reported) (Scozzari et al 2012) and in 1/66 subjects from the Bahamas (Simms et al. 2011).

Haplogroup A1a was originally reported in 1/44 samples of unspecified ethnic affiliation from Mali (Underhill et al. 2000). This haplogroup was also found in Guinea Bissau in 8/282 subjects from seven different ethnic groups (Rosa et al. 2007), in Senegambia [2/39 Mandenka (Wood et al. 2005)] and Mali [1/55 Dogon (Wood et al. 2005)].

Haplogroup A3a was originally reported in 1/88 subjects of unspecified ethnic affiliation from Ethiopia (Underhill et al. 2000). This haplogroup was also found in 1/20 "south Semitic" from Ethiopia (Wood et al. 2005) and 5/78 Dawro from Ethiopia (Batini et al. 2011).

Haplogroup B1 was also reported in 1/44 subjects of unspecified ethnic affiliation from Mali (Underhill et al. 2000,) and in 2/66 subjects from the Bahamas (Simms et al. 2011).

(Next pages) **Table 6: Absolute frequencies of major MSY haplogroups in 73 African populations.** References: (a) Cruciani et al. 2002; (b) Cruciani et al. 2004; (c) Cruciani et al. 2007; (d) Cruciani et al. 2010a; (e) Cruciani et al. 2011a; (f) Cruciani et al. 2011b; (g) Present study.

Region/population	Country	N	Haplogroup								References			
			A1b	A1a	A2	A3a	A3b1	A3b2	B1	B2a		B2b	CDEF	
43	Bamileke	Cameroon (South)	52							2			50	a,e,f,g
44	Bakaka	Cameroon (South)	12										12	a
45	Ewondo	Cameroon (South)	32							3			29	a,d,e,f,g
46	Bakola Pygmies	Cameroon (South)	36	3							nt	nt	nt	e
47	Biaka Pygmies	CAR	33							7	4	4	22	a,b,d,e,g
	<i>Eastern Africa</i>													
48	Mbuti Pygmies	DRC	13							2	4	4	7	a,b,d,e,g
49	Twa Pygmies	Burundi	7							4			3	d,e,g
50	Tutsi	Burundi	9										9	d,e,g
51	Hutu	Burundi	14							2			12	d,e,g
52	Cunama	Eritrea	20						2	3			15	d,e,g
53	Nara	Eritrea	15				2		1				12	d,e,g
54	Tigrai	Eritrea	28						3				25	d,e,g
55	Tigre	Eritrea	5										5	d,e,g
56	Afar Djibuti	Djibuti	25										25	d,e,g
57	Somali Djibuti	Djibuti	40										40	d,e,g
58	Somali Somalia	Somalia	23										23	b,c,d,e,g
59	Tigrai	Ethiopia	5										5	b,c,d,e,g
60	Amhara	Ethiopia	83						17				66	b,c,d,e,g
61	Gurage	Ethiopia	7										7	b,c,d,e,g
62	Ethiopian Jews	Ethiopia	22						9				13	a
63	Oromo	Ethiopia	62				1		12				49	b,c,d,e,g

Region/population	Country	N	Haplogroup										References
			A1b	A1a	A2	A3a	A3b1	A3b2	B1	B2a	B2b	CDEF	
64	Wolayta	12						2				10	b,c,d,e,g
65	Borana	9										9	b,c,d,e,g
66	Nilotic	65						11		3	4	47	b,c,d,e,g
64	Wolayta	12						2				10	b,c,d,e,g
65	Borana	9										9	b,c,d,e,g
66	Nilotic	65						11		3	4	47	b,c,d,e,g
67	Kikuyu	9								1		8	b,c,d,e,g
68	Luhya	51									1	50	b,c,d,e,g
69	Other Bantu	13						1				12	b,c,d,e,g
	<i>Southern Africa</i>												
70	Kung	64			5		18				5	36	a
71	Khwe	26					3					23	a
72	San	7			3						4		b,c,d,e,g
73	Bantu	8								1		7	b,c,d,e,g

DISCUSSION

In the present study, we applied next generation sequencing coupled with sequence capture to obtain a large number of variable positions that both test and improve the MSY phylogeny. We focused on two aspects of the experimental design to obtain high quality data. First, we selected segments with little or no homology with the X chromosome, to reduce the rate of alignment and variant calling errors attributable to the presence of gametologous sequences in the captured material. Second, we designed a high depth sequencing experiment, allowing reliable SNP calling also for below-average enriched segments.

Mutational pattern and evidence for selection

One main question related to our approach is whether the set of DNA segments subjected to capture faithfully summarized the pattern and tempo of mutations in the entire MSY. For example, the exclusion of largely represented paralogous regions prone to ectopic gene conversion (Rozen et al. 2003; Bosch et al. 2004; Rosser et al. 2009; Cruciani et al. 2010b; Trombetta et al. 2010) might have prevented the identification of a non-trivial proportion of variants. However, excluding paralogous regions likely caused the remaining amount of DNA (that captured here) to behave in a more similar manner to the autosomes. Questions raised by recent studies (Awadalla et al. 2010; Roach et al. 2010; The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012; Michaelson et al. 2012) challenge the transfer of the *de novo* autosomal mutation rate to MSY evolutionary studies.

However, many aspects of the pool of variants here discovered resemble the pattern observed among genome-wide *de novo* events, including the transition/transversion ratio, the proportion of mutations at CpG dinucleotides, and the shift from strongly bound to weakly bound base pairs. This justifies the use of an autosome-derived *de novo* mutation rate, also if one considers that convergence is being observed between evolutionarily derived estimates and pedigree derived estimates (Scally and Durbin 2012). In deriving our rate, we took into account the effect of the transmission through the male germline only, but could not account for other unknown specificities of the Y chromosome.

We noticed an enrichment of putatively physiologically relevant mutations along terminal branches. The action of purifying selection in the MSY was initially suggested based on the non-synonymous/synonymous diversity in 16 single-copy X-degenerate genes (Rozen et al. 2009). When performing the same calculations on our data, we obtained an approximate K_a/K_s ratio of 0.45, in line with effective purifying selection. As an alternative approach, we pooled our results together with those contained in the studies by Rozen (Rozen et al. 2009) and Wei (Wei et al. 2013) and counted the number of shared and private mutant alleles across the males represented in the three studies (Table 7). In this way, we found a significant excess of private missense variants ($P = 0.01$, Fisher exact test).

Among our variants, we found two potentially damaging mutations in *USP9Y*, a gene possibly involved in spermatogenesis. Both of these were found in very young branches (less than 10.5 kya). It should be noted, however, that the complete deletion (Luddi et al. 2009) and additional missense mutations (Rozen et al. 2009; Wei et al. 2013) in this gene were reported as not only heritable, but also compatible with the spread to a large number of males (Rozen et al. 2009), making it unlikely that *USP9Y* alone could be a cause of severe infertility (Tyler-Smith and Krausz 2009).

Y-position (GRCh37/hg19)	Gene	Reference allele	Alternative allele	Class	Amino acid change	PhyloP prediction	PolyPhen2 prediction	Shared	References
2655180	<i>SRY</i>	G	A	synonymous	S155S			yes	(3)
2710013	<i>RPS4Y1</i>	A	G	nonsynonymous	K10R			no	(1)
2712132	<i>RPS4Y1</i>	G	A	synonymous	S32S			yes	(3)
2722727	<i>RPS4Y1</i>	C	T	synonymous	Y149Y			no	(3)
2734854	<i>RPS4Y1</i>	C	T	synonymous	S237S			yes	(1), (3)
2829444	<i>ZFY</i>	A	G	nonsynonymous	M105V	Conserved	Possibly damaging	no	(3)
2846005	<i>ZFY</i>	A	G	nonsynonymous	N292S	Non-conserved	Benign	no	(2)
6736154	<i>AMELY</i>	C	T	nonsynonymous	R166Q	Non-conserved	NA	no	(3)
6736443	<i>AMELY</i>	C	T	nonsynonymous	V70M	Non-conserved	Possibly damaging	no	(3)
6736812	<i>AMELY</i>	G	T	synonymous	L36L			no	(3)
6958165	<i>TBL1Y</i>	C	A	synonymous	G494G			yes	(3)
14832620	<i>USP9Y</i>	G	T	nonsynonymous	E65D	Conserved	Benign	yes	(1), (2), (3)
14834046	<i>USP9Y</i>	G	A	nonsynonymous	R84Q	Conserved	Probably damaging	no	(1)
14838700	<i>USP9Y</i>	C	T	nonsynonymous	R211C	Non-conserved	Benign	yes	(2), (3)
14851554	<i>USP9Y</i>	T	C	synonymous	D471D			yes	(1), (3)
14870505	<i>USP9Y</i>	G	T	nonsynonymous	V567L	Conserved	Probably damaging	no	(1)
14888783	<i>USP9Y</i>	A	C	synonymous	L876L			yes	(1), (3)
14889974	<i>USP9Y</i>	C	T	synonymous	L887L			no	(3)
14898163	<i>USP9Y</i>	G	A	nonsynonymous	A1060T	Conserved	Benign	yes	(3)
14898546	<i>USP9Y</i>	C	A	nonsynonymous	P1125H	Conserved	Benign	no	(1)
14902414	<i>USP9Y</i>	G	A	synonymous	E1212E			no	(3)
14902417	<i>USP9Y</i>	C	T	synonymous	C1213C			no	(1)
14924869	<i>USP9Y</i>	C	T	synonymous	P1497P			yes	(3)
14952467	<i>USP9Y</i>	A	G	synonymous	Q2005Q			no	(1)
14954280	<i>USP9Y</i>	T	C	synonymous	P2109P			yes	(2), (3)
14959237	<i>USP9Y</i>	C	G	nonsynonymous	S2350C	Conserved	Probably damaging	no	(2)
14968331	<i>USP9Y</i>	G	A	synonymous	S2377S			no	(3)
15024924	<i>DDX3Y</i>	G	A	nonsynonymous	D163N	Non-conserved	Probably damaging	no	(3)
15026544	<i>DDX3Y</i>	G	T	nonsynonymous	A64S	Non-conserved	Benign	no	(1)
15027863	<i>DDX3Y</i>	G	A	synonymous	E227E			yes	(1)
15028176	<i>DDX3Y</i>	G	C	nonsynonymous	S254T	Conserved	Benign	yes	(1), (3)
15028931	<i>DDX3Y</i>	T	C	synonymous	Y390Y			no	(1)
15346546	<i>UTY</i>	C	A	nonsynonymous	M1056I			no	(2)
15435237	<i>UTY</i>	T	C	nonsynonymous	M762V			no	(2)
15435514	<i>UTY</i>	C	T	nonsynonymous	R451H	Non-conserved	Benign	no	(1)
15467824	<i>UTY</i>	A	G	synonymous	S114S			yes	(1), (2), (3)
15591193	<i>UTY</i>	G	A	nonsynonymous	L53F	Non-conserved	Benign	no	(1)
15591445	<i>UTY</i>	T	C	nonsynonymous	E34G	Non-conserved	Benign	no	(1)
15591537	<i>UTY</i>	G	C	synonymous	S3S			yes	(1), (2), (3)
16936081	<i>NLGN4Y</i>	C	T	synonymous	T45T			no	(3)
16942397	<i>NLGN4Y</i>	T	C	synonymous	T365T			no	(3)
21154426	<i>CD24</i>	G	A	nonsynonymous	A57V			no	(2)
21751440	<i>TXLNG2P</i>	A	C	nonsynonymous	N106T			no	(3)
21751449	<i>TXLNG2P</i>	C	T	nonsynonymous	S109L			yes	(3)
21867887	<i>KDM5D</i>	C	A	nonsynonymous	Q713H	Non-conserved	Benign	no	(2)
21868068	<i>KDM5D</i>	C	T	nonsynonymous	R653Q	Non-conserved	Benign	no	(3)
21868726	<i>KDM5D</i>	T	C	synonymous	G516G			yes	(3)
21869264	<i>KDM5D</i>	G	A	synonymous	P431P			no	(3)
21869923	<i>KDM5D</i>	C	T	nonsynonymous	S274N	Non-conserved	Benign	no	(3)
21869928	<i>KDM5D</i>	T	C	synonymous	S272S			no	(3)
21897321	<i>KDM5D</i>	T	C	nonsynonymous	T239A	Non-conserved	Benign	no	(2)
22921768	<i>RPS4Y2</i>	G	A	synonymous	S32S			yes	(2), (3)
22942897	<i>RPS4Y2</i>	T	C	synonymous	A257A			yes	(3)

Table 7: Mutations in protein coding regions. We consider a mutation to be “shared” if the alternative allele is independently found in different studies, or in different individuals from a single study. PhyloP and Polyphen2 prediction represent the interpretation of the respective scores (not shown). References: (1) present study; (2) Wei et al. 2013; (3) Rozen et al. 2009.

In summary, we cannot exclude a role of purifying selection in shaping the MSY diversity to some extent, particularly when considering that the absence of recombination leads to the removal of all markers on a selected-against MSY haplotype. Nevertheless, the features of our mutational pattern strongly suggest that in the MSY a large proportion of newly arisen alleles have survived in the phylogeny.

Based on all the above arguments and the fact that only 1% of the sequence here screened is coding, we confidently applied genetic dating equating the substitution rate to the mutation rate over all positions, in the same way as other authors (Mendez et al. 2013; Wei et al. 2013).

Implications of the new MSY chronology

For a long time, the MSY tree has suffered from a lower level of resolution than that of the mtDNA phylogeny. However, with the advent of new technologies now allowing for high-throughput identification of Y chromosome SNSs, these markers can now be used to characterize the MSY tree at a greater level of resolution as well as to improve age estimates. Because the accumulation of SNSs in the MSY over the course of human evolutionary history would not have plateaued, these markers provide a nearly unlimited resource for refining and dating the phylogeny, given that an appropriately long sequence is evaluated. By contrast, diversity at microsatellite loci, not only is confounded by recurrent mutations, but may also reach a maximum (Busby et al. 2012). This limits their use in resolving the phylogeny and has often given rise to unreliable dating results, especially for deep-rooted lineages. The accuracy of SNS-based dating methods relies on the equally efficient discovery of new SNSs across all lineages. Our results show that a single targeted next-generation run can produce

a highly reliable and informative phylogeny with a uniformly intense search for markers across all lineages included in the study.

Knowledge of the deepest branches in the MSY tree has long been incomplete and the phyletic relationships between lineages have often been reordered, including the placement of the root (Batini et al. 2011; Cruciani et al. 2011a; Scozzari et al. 2012; Mendez et al. 2013). The representation of deep lineages at low frequencies and often from small remote populations has also made their study difficult. Yet, deep-rooted lineages are particularly informative in the reconstruction of the scenario of an ancient population structuring in Africa. This is currently considered to be the source of global patterns of genome-wide diversity under the generally held view of an exit out of Africa originating from the eastern portion of the continent (Campbell and Tishkoff 2010; Henn et al. 2012; Scally and Durbin 2012).

We will now discuss the implications of the structure and timings of the MSY phylogeny reconstructed in our experiment. These considerations are made in light of the present day distribution of Y chromosome haplogroups in Africa, which is depicted in Figure 24.

The subjects analysed in this work were intentionally selected to represent a wide range of diversity and antiquity among MSY lineages, in order to resolve and date the deepest branches in parallel with more widely studied lineages. The resulting tree, with a TMRCA of 196 kya, displays extraordinary deep ancestry for most of the early branches within Africa, a fact that has not been fully acknowledged in previous works.

The first two splits in our tree, dated at 196 kya and 167 kya, separate branches (A1b and A1a) that are currently found at low frequency in Central-Western Africa (Figure 24 and Table 6), but have not been detected from elsewhere in the African continent. This geographical confinement of deep lineages is at odds with the mainly Eastern African position of sites providing fossils of comparable ages (McDougall et al. 2005) [though it must be kept

into consideration that the assignment of fossils to AMH or to archaic forms of *Homo* is still a controversial question (Tattersall and Schwartz 2008; Schwartz and Tattersall 2010)].

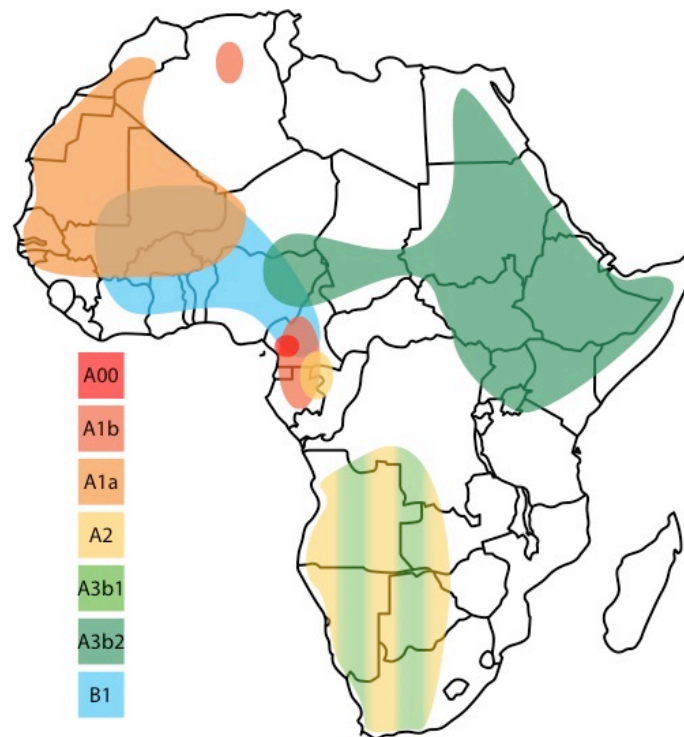


Figure 24: Geographic distribution of deep rooting haplogroups in the African continent. Map of Africa showing the present-day home ranges of the MSY haplogroups discussed in the text. Haplogroup A00 (Mendez et al. 2013) is also shown. Colour intensity does not reflect haplogroup frequencies in the corresponding populations. Haplogroup B2, ubiquitous in sub-Saharan Africa, was omitted. Redrawn from Chiaroni et al. 2009, with modifications and updates based on haplogroup frequencies reported in Table 6 and in Mendez et al. 2013.

Two hypotheses can be put forward to attempt a reconciliation of this geographic discordance between fossils and genetics. First, Eastern Africa is the original homeland of A1b and A1a, which might have then relocated to Central-Western Africa, and gone

extinct (possibly together with other yet unknown deep rooted branches) in Eastern Africa. Second, A1b and A1a actually originated in Central-Western Africa, where they are still observed today, but fossil record of ancient populations there was lost. The finding of the oldest lineage recorded so far (A00, 338 kya) in Cameroon (Mendez et al. 2013), adds to our phylogeographic results in suggesting that deep MSY lineages might have resided in Central-Western Africa earlier than 160 kya.

In discussing the implications of these findings, we note that the tens of thousands of years separating some of the consecutive branching events dramatically reduce the power of the Bayesian phylogeographic inference (Figure 21), resulting in decreasing statistical support from the tips towards the root of the tree. It should also be considered that the demography of past populations, characterized by small effective sizes with intense drift, and, possibly, subsequent expansions, may have caused lineages to wander over vast geographic regions, also in response to climate pressures (Burroghs 2005; Castañeda et al. 2009), with the potential for generating an altered phylogeographic signal. Finally, we note that the sampling scheme here used is geographically uneven, and is constrained by current knowledge on the distribution of extremely rare deep lineages (Table 6). This calls for a more even sampling coverage of MSY diversity in Africa, which should be also compared with the conclusions of recent autosomal genetic and craniometric data (Ramachandran et al. 2005; Manica et al. 2007; Tishkoff et al. 2009; Pagani et al. 2012; Schlebusch et al. 2012).

Our data place the TMRCA of haplogroup B at 110 kya, a date which is unexpectedly old considering the branch lengths in the previously known phylogeny (Karafet et al. 2008). The current distribution of chromosomes and dating of the two B subclades (Figure 23) also testify early dispersals followed by partial isolation. In particular, haplogroup B2a-M150 has been associated with the expansion of Bantu-speakers (Beleza et al. 2005; Berniell-Lee et al. 2009), and it was previously dated at 6.0 kya on the basis

of its STR diversity (Batini et al. 2011). In our analysis, it turned out to be a very ancient lineage (40 kya), long predating the alleged timing of the Bantu expansion. Beyond the disparity of the microsatellite- and SNP-derived ages, these data indicate that only a small subset of the overall B2a diversity became incorporated into the male gene pool of Bantu speakers. As for B2b, it has been reported that a highly divergent subset of these chromosomes is found South African Khoe speakers, with a STR-based TMRCA of 69.9 kya (Tishkoff et al. 2007), suggesting that we possibly did not sample the most divergent lineages of this clade.

Two main routes for the AMH dispersal out of Africa are still widely debated: the northern route through Egypt to the Levant, where AMH fossils dated prior to 100 kya have been found (Grün et al. 2005), and the southern route through the Bab-el-Mandab strait to the Arabian peninsula at 125 kya as argued by Armitage et al. (2011) based on archaeological records. As far as genetic evidence is concerned, mtDNA data (Soares et al. 2011; Fernandes et al. 2012) favors this latter route, but not before ~70 kya.

In our phylogeographic analyses, the nodes basal to haplogroups E-F and C-F provide information regarding the exit out of Africa. On purely phylogeographic grounds, as E-F is basal to chromosomes found both inside and outside Africa (haplogroup E), while C-F is the ancestor of all non-African haplogroups, the exit from the continent should have taken place in a time frame corresponding to the branch connecting these two nodes (segment 37 in Figure 18). Moreover, these are the oldest nodes for which a non-sub-Saharan ancestral state received statistical support, with a level of probability two times higher than any of the alternatives in the Bayesian analysis (Figure 21).

Three scenarios are compatible with our phylogeographic analyses (Figures 21 and 22), dating results (Figures 17 and 21), the known geographic distribution of patrilineages (Underhill and Kivisild 2007; Chiaroni et al. 2009), also keeping into account the possibility that lineages were driven to extinction or to exceedingly

low frequencies by drift. In the first one (Figure 25, panel A), the exit of a precursor of haplogroup E-F occurred anytime between 114.8 and 85.5 kya (overall window 145-65 kya, corresponding to the length of branch 35 and C.I.s of its defining nodes, see Figures 18 and 21), followed by the diversification of E-F in Eurasia. This scenario requires the re-entry of a single lineage (haplogroup E) in Africa, as originally proposed (Hammer et al. 1998). In the second scenario, the node basal to E-F originated in Africa and the exit of a precursor of C-F took place between 85.5 and 83.5 kya (overall window 108-64 kya, corresponding to the length of branch 37 and C.I.s of its defining nodes, see Figures 18 and 21), together or separately from E, and followed by the extinction of the early C-F in Africa (Figure 25, panel B). In the third scenario, three or more lineages left Africa after 83.5 kya; this would require the not remote possibility that multiple lineages went extinct or are yet to be found in Africa (Figure 25, panel C).

One of the implications of the first scenario is that the AMH occupation of the Middle East and/or the Arabian Peninsula before 100 kya could no longer be regarded as a “temporary excursion” (Scally and Durbin 2012) but rather the seeding event for the MSY diversity found today in Eurasia. The dates from the second and third scenario are similar to estimates from mtDNA variation that date the exit of matrilineages based on the topology and TMRCA of haplogroup L3 (Atkinson et al. 2009; Soares et al. 2011; Fu et al. 2013a). Neither MSY scenario excludes an out-of-Africa exit before a major event marking AMH occupation further East, i.e. the Toba eruption ~74 kya (Chesner et al. 1991; Mellars et al. 2013). Moreover, they all open up the possibility of a temporal gap in which an intermediate bottlenecked population existed in the Middle East/Arabian peninsula, and whose genetic signature is now visible in the genome pool of Eurasians. They also fit with the finding of deep-rooted Eurasian Y haplogroups in the southern Arabian Peninsula (Abu-Amero et al. 2009) and Lebanon (Zalloua et al. 2008).

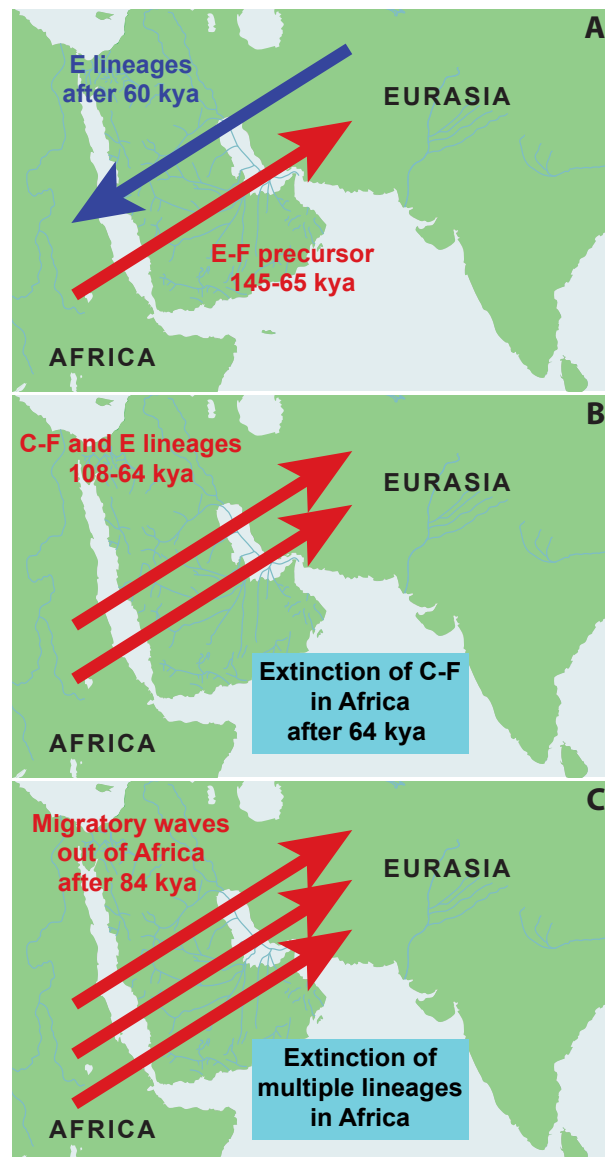


Figure 25: Models for the Out of Africa. Models for the earliest migrations of *Homo sapiens* out of the African continent. Each panel represents one of the three alternative models proposed in the text. (modified from Petraglia et al. 2011).

In summary, inferences regarding the ancestral relationships and timing of movements of human populations in the exit out of Africa based on extant MSY diversity remain rather imprecise, due to the reduced topological structure at branches leading to haplogroups E-F and C-F through the time window 145-64 kya (Figure 17). The array of new markers here generated strongly prompts the typing of haplogroup D (not represented among our males) as well as rare African and non-African carriers of E*, C* (Weale et al. 2003; Zalloua et al. 2008; Abu-Amro et al. 2009) or even older paragroups, in search for lineages that could modify the topology of the MSY tree with new informative nodes. Reconciling archaeological and genetic dates for the uniparental systems is also linked to the finding of rare old lineages for the mtDNA, as recently suggested (Rose et al. 2011). Strong evidence may also derive from ancient DNA collected in appropriate archaeological layers of Eastern Africa and South-Western Eurasia.

One remarkable aspect of our results regards the statistically significant low number of mutations on the branch corresponding to haplogroup A1b. The reason for this is yet to be clarified. In a genome-wide analysis (Conrad et al. 2011), a fewer *de novo* mutations were found in a Yoruban trio compared to a non-African trio, suggesting the need for more extensive analyses to assess possible population-specific effects. As far as selection is concerned, the low number of coding variants here observed prevented the testing of their differential occurrence across branches, and we cannot exclude different selective pressures acting on different lineages. Two studies (Lohmueller et al. 2008; Fu et al. 2013b) showed a higher number of deleterious variants in Europeans compared to Africans, which is most likely due to the combined effects of a long lasting bottleneck and re-expansion of Europeans. The reduction of the male effective population size, which is most likely to be associated with the exit out of Africa, may have provided enough opportunity for deleterious variants to appear and increase in frequency. In this case, the net effect would

be an extra load of mildly deleterious mutations that elongated the branches that were involved in the bottleneck.

Further developments

Approximately 80% of the markers reported here are novel and open new perspectives for the refinement of the phylogeny through the study of additional subjects, including A00 and possibly undiscovered deep lineages. In fact, the resolution here attained will enable a search for highly specific lineages with PCR-based approaches, which will be eventually also applicable to ancient DNA and help shed light on the possible historical continuity between individuals of the past and current populations.

We endorse the need for a reference sequence incorporating ancestral alleles at all known variable positions (Wei et al. 2013), which would greatly facilitate further works in this field.

Estimates of the substitution rate for the MSY (Xue et al. 2009; Francalacci et al. 2013; Mendez et al. 2013; Poznik et al. 2013; Wei et al. 2013 and present work) currently suffer from an appreciable uncertainty. Major improvements in dating may derive from taking into account the complexity of the mutational process (Michaelson et al. 2012). For example, the possibility that clusters of mutations may hit the MSY seriously challenges the concept of linear accumulation with time. We highly recommend that, in future, mutation rates be worked according to the local features of the MSY sequences, and that they then be used on appropriately partitioned datasets as previously suggested (Fu et al. 2013a). We see deep-rooted pedigrees (Xue et al. 2009) as the material that should be chosen in order to work out robust estimates of these rates, given the low chances of observing mutational events in such a small portion of the genome in a single generation.

MATERIALS AND METHODS

Samples

Human Y chromosomes to be sequenced were selected on the basis of their SNP/STR genotype which had been determined in the present or previous studies (Cruciani et al. 2004, 2007, 2010a, 2011a, 2011b; Trombetta et al. 2011; Scozzari et al. 2012). Most samples were chosen in order to represent as many deep branches of the Y phylogeny as possible. In the vast majority of cases, DNA was prepared from fresh venous blood, with no cell culturing.

Genotyping

To obtain a refined MSY tree, as well as to select the Y chromosome for the NGS experiment, we determined the allelic state at 168 markers (22 newly described mutations and 146 previously reported mutations [Cruciani et al. 2011a]) in 224 Y chromosomes which were representative of different Y haplogroups. Fifty-nine mutations reported by Karafet (Karafet et al. 2008) were also analysed (Figures 13 and 14). In order to identify new mutations, we also analysed by Sanger sequencing about 90 kb in each of two unrelated Y chromosomes belonging to haplogroups A3b1 (89.8 kb) and A3b2 (89.3 kb).

We designed polymerase chain reaction (PCR) and sequence primers on the basis of the Y-chromosome sequence reported in the February 2009 assembly of the UCSC Genome Browser (<http://genome.ucsc.edu/>) using Primer3 software (<http://primer3.ut.ee>). Sequencing templates were obtained through PCR in a 50-ml reaction containing 50 ng of genomic DNA, 200 mM each deoxyribonucleotide (dNTP), 2.5 mM MgCl₂, 1 unit of Taq polymerase, and 10 pmoles of each primer. A touch-down

PCR program was used with an annealing temperature that decreased from 62°C to 55°C over 14 cycles, followed by 30 cycles with an annealing temperature of 55°C.

Following DNA amplification, PCR products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany). Cycle sequencing was performed using the BigDye Terminator Cycle Sequencing Kit with Amplitaq DNA polymerase (Applied Biosystems, Foster City, CA) and an internal or PCR primer. Cycle sequencing products were purified by ethanol precipitation and run on an ABI Prism 3730XL DNA sequencer (Applied Biosystems). Chromatograms were aligned and analyzed for mutations using Sequencher 4.8 (Gene Codes Corporation, Ann Arbor, MI).

DNA library preparation

About 3 µg of genomic DNA was sheared using a Covaris ultrasonicator to obtain DNA fragments that were mainly distributed between 200 bp and 300 bp. Purified fragments were end-repaired and ligated to paired-end sequencing adapters containing short sample-specific sequence tags to allow multiple samples to be simultaneously analysed during the sequencing step. The samples were then amplified by LM-PCR (Ligation-Mediated PCR) in order to selectively enrich those DNA fragments that have adapter molecules on both ends.

Selection and targeting of MSY unique regions

We selected five regions (Figure 15 and Table 8) of the X-degenerate portion of the MSY, which showed a low degree of similarity with X gametologous sequences, for a total of 3,768,982 bp.

A custom sequence capture array, manufactured by Roche Nimblegen, was used for the target enrichment of the indexed

genomic library. A set of unique and overlapping probes was designed to capture unique sequences at the five MSY regions under study. The Sequence Search and Alignment by Hashing Algorithm (SSAHA) was used to assess probe uniqueness (Ning et al. 2001). Probe tiling of the target regions excluded most of the repetitive interspersed elements. The capture probe set covered a total of 1,495,512 bases of the target region, distributed into 5,274 fragments.

Region number	Start position (GRCh37/hg19)	End position (GRCh37/hg19)	Size (bp)
1	2689001	2910620	221620
2	6655519	6673610	18092
3	7540722	8740721	1200000
4	14629842	15962456	1332615
5	18553379	19550033	996655

Table 8: Genomic coordinates for the five X-degenerate regions of the MSY selected for the present study.

DNA sequencing and alignment

The captured library was loaded onto an Illumina HiSeq 2000 platform to produce a 50× mean depth sequence for the 1.5Mb targeted region.

The raw sequencing output was processed in order to discard low quality reads, adapter contamination and repeated reads. Clean data were then sorted using the subject-specific identifiers (see previous section). For each subject, the sequencing reads were aligned to the human Y chromosome reference sequence (GRCh37/hg19) using the Burrows-Wheeler Aligner (BWA)

software (Li and Durbin 2009), to generate an alignment file (.sam, Sequence Alignment/Map) (Li et al. 2009a). Library preparation, targeting, sequencing and alignment were performed by BGI-Tech (Shenzhen, China).

SNP calling and filtering

Candidate variant nucleotide positions (compared to the human reference sequence) were identified by using the SOAPsnp software (Li et al. 2009b), with haploid specific parameters for variant calling. Among the candidate mutations from the SOAPsnp analysis, we only considered those found in the 1.5 Mb target region and which fulfilled all of the following criteria: 1) quality score of consensus (QS) ≥ 90 ; 2) depth $\geq 4\times$; 3) difference between the depth and the total number of reads for the two best bases ≤ 4 . This latter criterion was adopted to identify false SNP calls due to misalignments in proximity of insertions/deletions. The filtering was then refined by visual inspection of the .sam files using the Integrative Genomics Viewer (IGV) software (Thorvaldsdóttir et al. 2013). In particular, we inspected variant calls falling in these categories: $90 \leq \text{QS} \leq 98$; $4\times \leq \text{depth} \leq 10\times$; distance from the closest SNP in the same sample ≤ 20 bases; depth for the second best base ≥ 3 . One well-known complication associated with estimating sequence divergence is that mapping quality for a read depends on the number of differences between the read and the reference. We then inspected the alignments of all subjects, in sliding windows of 25 kb (using IGV), searching for extreme variations in sequence coverage among samples, which may be indicative of structural rearrangements and the inability to detect variants. We also used GASVPro (Sindi et al. 2012) to identify structural variants from paired-end mapping data. Since such rearrangements can also lead to the unscheduled capture of paralogous divergent sequences, we checked for clustering of variants in short stretches of DNA on each tree branch after constructing the tree.

SNP data quality control

To assess the accuracy of our set of filtered variants, we performed a series of quality controls using both resequencing and literature data.

In order to test for false positives, a total of 80 SNPs (here described for the first time) were tested and confirmed using either Sanger resequencing or RFLP assays. Also, one individual (S38 in Table 2) had also been sequenced in a previous study (Cruciani et al. 2011a), allowing the comparison of 39,859 bp of overlapping MSY sequence. Neither false positives nor false negatives were discovered.

We also compared our variants with those reported in a recent high-coverage resequencing study of the MSY (Wei et al. 2013). In order to check nucleotide positions with a high probability of displaying variation in both studies, we selected SNPs reported by Wei et al. (2013) to define haplogroups DR and DE. Among 53 SNPs falling within the MSY regions here sequenced, a 100% concordance was observed between the two data sets. Additionally, we compared the allelic states reported in sample NA21313 (Wei et al. 2013) with those we observed in sample NA21367 (S11 in Table 2). Since these two individuals belong to the same population and share the same terminal MSY haplogroup (A-M13), we expected them to share most of their alleles at variable positions. All but one of the 76 alternative alleles in sample NA21313, which were in regions analysed in the present study, were also observed in sample NA21367.

Finally we confirmed the allele state at all positions described by Mendez et al. (2013) that define haplogroups A0-T and A0 and that reside within the DNA segments captured by our experimental protocol.

Reconstruction of ancestral allele states

For each variable position, the filtered data consisted in the listing of subjects carrying an alternative base, i.e. a base call that differed from the reference sequence (GRCh37/hg19). We then determined the ancestral and derived state at each position in the entire phylogeny using the following procedure, which takes into account that A1b (A0) is one of the deepest branches of the MSY tree (Cruciani et al. 2011a; Scozzari et al. 2012): alternative bases in one or more (but not all) A1b subjects (S01, S03, S04, and S05) were considered derived; alternative bases in one or more (but not all) of the remaining subjects were considered ancestral if shared with subject S01 and derived otherwise; the ancestral state at the 158 positions remaining after these steps was determined by comparison with the orthologous positions of the chimpanzee (CSAC 2.1.4). This led to 94 and 28 mutations that were unambiguously attributed to branches 1 and 8, respectively (Figure 18); the ancestral state at 36 positions remained undetermined.

We chose this method to avoid uncertainties associated with the straightforward application of the human-chimp comparison. In fact, mutations occurring in the chimpanzee lineage may result in the erroneous assignment of human alleles as derived alleles (estimated genome-wide rate 0.5%). This problem is particularly acute for the Y chromosome, for which an enhanced divergence between the two species has been reported (Chimpanzee Sequencing and Analysis Consortium 2005).

We also used the ancestral allele information to count the number of mutations occurring at ancestral CpG dinucleotides.

Annotation of variants in coding regions

The program wANNOVAR (Chang and Wang 2012) was used to identify and note exonic variants found in this and other (Rozen et al. 2009; Wei et al. 2013) works. The program also returns conservation levels (PhyloP scores) and predicted functional

importance (PolyPhen) scores. The UCSC known gene and Ensembl gene definitions were used.

Tree construction

A contingency table of alternative bases by subject (rows) and chromosome position (columns) was converted into .rdf and .meg files, to be handled with the programs Network (Bandelt et al. 1999) and MEGA (Tamura et al. 2011), respectively. Network was used to obtain a median joining network for rho calculations, a complete listing of mutated positions along each branch and a precise count of inferred recurrent mutations at the same position. This tiny subset of positions (4 recurrent mutations, see Appendix 1) was re-checked and confirmed in the original alignment files. MEGA was used to obtain a maximum parsimony tree. Note that both methods ignore the information on the ancestral vs. derived state for the particular allele observed in each subject, as they only consider state changes. The position of the root in the MP tree was determined by partitioning the 36 mutations with unknown ancestral state (see Appendix 1) proportionally to those that had been unambiguously assigned to branches 1 and 8, respectively (28 and 8 assigned, respectively).

Mutation rate

In order to model the substitution process at the surveyed positions, we took into account the careful measurements of the genome-wide *de novo* mutation rates recently obtained from parent-child transmissions and deep-rooted pedigrees (Awadalla et al. 2010; Roach et al. 2010; The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012; Michaelson et al. 2012). Remarkable findings in this field include the effect of the sex of the transmitting parent (summarized by the alpha ratio) and paternal age at conception.

We used the repeatedly confirmed genome-wide value of 1.2×10^{-8} /position/gamete/generation to infer an MSY-specific value. To this aim, we considered a 4:1 alpha ratio (Campbell et al. 2012; Kong et al. 2012) and the strict patrilineal inheritance of the MSY. We obtained the value of 0.64×10^{-9} /position/gamete/year (assuming 30 years/generation), i.e. a very similar value to that of Mendez et al. (2013) who incorporated the regression on paternal age in the estimation. Here we notice that, in the multigenerational timescale for the MSY, the *de novo* mutation rate per year is less sensitive to paternal age compared to the rate per generation. In fact, fathering at an old age implies a higher mutation rate but also corresponds to a longer inter-generation interval for the Y chromosome. Based on the data reported in Figure 2 of Kong et al. (2012), we were able to calculate that a shift of the average paternal age from 20 to 30 years corresponds to an increase of 40% in the rate per generation (44 vs. 64 total mutations under the linear model) but only a moderate decrease of the rate per year (2.2 vs. 2.13 total mutations). We obtained the 95% C.I. for our mutation rate ($0.47 - 0.82 \times 10^{-9}$) with 10,000 simulations of Poisson-distributed mutational events in the paternal and maternal gametes, with averages from Kong et al. (2012). In summary, our rate is somewhat lower than that calculated by directly examining the MSY transmission in a single deep-rooted pedigree (Xue et al. 2009), but its estimation is indirectly based on a much larger number of mutational events. Also, our value is similar to that obtained in a comparative genomics perspective (about 0.70×10^{-9}) by applying the same calculations to the divergence between humans and great apes with the alpha ratio reported therein (Scally et al. 2012). Two recent papers have recalculated an evolutionary effective rate by calibrating the age of a limited number of nodes on known MSY founding events. Of the two values, one based on the Neolithic population expansion in Sardinia [0.65×10^{-9} (Francalacci et al. 2013)] turned out to be very similar to that used here, while the one based on the initial colonization of the

Americas [0.82×10^{-9} (Poznik et al. 2013)] is at the upper limit of our C.I.

Time estimates and phylogeographic analyses

We applied two independent methods for dating the tree nodes. The first is based on the rho statistic, i.e. the average number of differing sites between a set of sequences and a specified common ancestor (which needs not be among the sampled sequences) (Forster et al. 1996). This statistic is linearly related to time and mutation rate ($\rho = \mu \times t$) (Jobling et al. 2004), assuming constancy of the rate across the tree branches. The statistic and associated confidence interval were computed with the software Network (Bandelt et al. 1999). The mutation rate was as reported above, corresponding to a substitution every 1,044 years over the 1.5 Mb sequence here scored.

We also applied a Bayesian estimation of node ages, through BEAST (Drummond and Rambaut 2007) using the entire set of 2,386 variable positions. This makes it possible to consider complex models, including different substitution matrices, a relaxed clock for heterogeneous rates across the tree branches, and different dynamics of population growth. We used a GTR model for nucleotide substitutions under a lognormal relaxed clock for rate heterogeneity across branches. The initial tree was that obtained by maximum parsimony (see above). The prior for the substitution rate was given a normal distribution centred on the mutation rate value reported above. An expansion model was used for the population size in order to appropriately account for the faster recent growth (Gignoux et al. 2011; Keinan and Clark 2012). Rather flat priors were used, i.e. lognormal[10,3] for the current population size, exp[0.2] for the ancestral/current population size ratio and uniform[0, 0.00133] for the growth rate/year in the expansion phase, using the upper limit reported in the literature for pre-agricultural cultures (Ammerman and Cavalli-Sforza 1984; Boone 2002; Hamilton et al. 2009). We used two runs of 20

million steps each, sampled every 10,000 steps. The initial 20% of each run was discarded as burn-in and the outputs combined and analysed with Tracer.

The heterogeneity of substitution rates in different tree branches was tested by repeating the same runs under a strict clock model with identical priors, and comparing the tree likelihoods by means of Bayes' factors (Nylander et al. 2004). A test to compare rates between selected branches of the tree was performed by using a χ^2 test as reported in eq. 10 in Kumar and Filipski (2001), taking into account only non-recurrent mutations as recommended.

In order to make inferences on the most likely locations for ancestors corresponding to nodes in the tree, the 68 subjects were assigned to four geographic macroregions, i.e. Central-Western Africa, Southern Africa, Eastern Africa and rest of the world (comprising Northern Africa and other continents). A discrete phylogeographic model was examined by both Bayesian search and maximum parsimony. A run of BEAST using geographic categories as a discrete trait (Lemey et al. 2009) was performed, with the same population parameters as above. The maximum parsimony approach implemented in the program RASP (Yu et al. 2010) was applied to the maximum parsimony tree of Figure 2, allowing ancestral ranges to include no more than two of the four geographic macroregions. For each of the two phylogeographic analyses, the inference on ancestral locations for each node was represented as a pie chart and overlaid on the BEAST tree.

REFERENCES

- Abu-Amro KK et al. (2009) Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. *BMC Genet.* 10: 59-68.
- Ammerman AJ and Cavalli-Sforza LL (1984) *The Neolithic transition and the genetics of populations in Europe*. Princeton University Press, Princeton.
- Armitage SJ et al. (2011) The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science* 331: 453-456.
- Arredi B et al. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am. J. Hum. Genet.* 75: 338-345.
- Atkinson QD et al. (2009) Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc. Roy. Soc. B: Biol. Sci.* 276: 367-373.
- Aubert M et al. (2012) Confirmation of a late middle Pleistocene age for the Omo Kibish 1 cranium by direct uranium-series dating. *J. Hum. Evol.* 63: 704-710.
- Awadalla P et al. (2010) Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* 87: 316-324.
- Balaresque P et al. (2010) A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* 8: e1000285. doi:10.1371/journal.pbio.1000285.

- 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.
- Battaglia V et al. (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur. J. Hum. Genet.* 17: 820-830.
- Batzler MA and Deininger PL (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3: 370-379.
- Behar DM et al. (2008) The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* 82:1130-1140.
- Beleza S et al. (2005) The genetic legacy of western Bantu migrations. *Hum. Genet.* 117: 366-375.
- 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.
- Boone JL (2002) Subsistence strategies and early human population history: an evolutionary ecological perspective *World Archaeol.* 34: 6-25.
- Bosch E et al. (2004) Dynamics of a human interparalog gene conversion hotspot. *Genome Res.* 14: 835-844.
- Brooks AS (2005) The Middle Stone Age of Eastern Africa: an anti-revolutionary perspective. In *Rethinking the human revolution: new behavioural and biological perspectives on*

the origins and dispersal of modern humans. Conference 7-11 September, Cambridge.

Burroughs WJ (2005) *Climate change in prehistory*. Cambridge University Press, Cambridge, U.K.

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

Butler JM (2003) Recent developments in Y-short tandem repeat and Y-single nucleotide polymorphism analysis. *Forensic Sci. Rev.* 15: 91-111.

Campbell CD et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44: 1277-1281.

Campbell MC and Tishkoff SA (2010) The evolution of human genetic and phenotypic variation in Africa. *Curr. Biol.* 20: R166-R173.

Cann RL et al. (1987) Mitochondrial DNA and human evolution. *Nature* 325: 31-36.

Capelli C et al. (2001) A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am. J. Hum. Genet.* 68: 432-443.

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

- Castañeda IS et al. (2009) Wet phases in the Sahara/Sahel region and human migration patterns in North Africa. *Proc. Natl. Acad. Sci. USA* 106: 20159-20163.
- Chang X and Wang K (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.* 49: 433-436.
- Chesner CA et al. (1991) Eruptive history of Earth's largest Quaternary caldera (Toba, Indonesia) clarified. *Geology* 19: 200-203.
- 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.
- Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69-87.
- Conrad DF et al. (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43: 712-715.
- Cox MP (2009) Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. *Hum. Biol.* 80: 335-357.
- 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. (2010a) 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. (2010b) About the X-to-Y Gene Conversion Rate. *Am. J. Hum. Genet.* 86: 495-497.

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

- Cruciani F et al. (2011b) Strong intra- and inter-continental differentiation revealed by Y chromosome SNPs M269, U106 and U152. *Forensic Sci. Int. Genet.* 5: e49-e52.
- Day MH (1969) Omo human skeletal remains. *Nature* 222: 1135-1138.
- 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 Rienzo A et al. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91: 3166-3170.
- Drummond AJ and Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7: 214.
- Fagundes NJR et al. (2007) Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* 104: 17614-17619.
- Fernandes V et al. (2012) The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *Am. J. Hum. Genet.* 90: 347-355.
- Feuk L et al. (2006) Structural variation in the human genome. *Nat. Rev. Genet.* 7: 85-97.
- Francalacci P et al. (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* 341: 565-569.

- 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.
- Fu Q et al. (2013a) A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* 23: 553-559.
- Fu W et al. (2013b) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493: 216-220 and Erratum in 495:270.
- Gignoux CR et al. (2011) Rapid, global demographic expansions after the origins of agriculture. *Proc. Natl. Acad. Sci. USA* 108: 6044-6049.
- 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: 467-72.
- Graves AMJ (2006) Sex chromosome specialization and degradation in mammals. *Cell* 124: 901-914.
- Green RE et al. (2010) A draft sequence of the Neandertal genome. *Science* 328, 710-722.

- Gronau I et al. (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43: 1031-1034.
- Grün et al. (2005) U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J. Hum. Evol.* 49: 316-334.
- Group NCCM (1992) A comprehensive genetic linkage map of the human genome. *Science* 258: 67-86.
- Gusmão L et al. (2005) Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* 26: 520-528.
- Gutenkunst RN et al. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695. doi: 10.1371/journal.pgen.1000695.
- 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 (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 and Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet.* 56: 651-662.

- 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 et al. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15: 427-441.
- 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 Y-chromosome 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.
- 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) The great human expansion. *Proc. Natl. Acad. Sci. USA* 109: 17758-17764.

- 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.
- Jakobsson M et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.
- Jobling MA (2008) Copy number variation on the human Y chromosome. *Cytogenet. Genome Res.* 123: 253-262.
- Jobling MA and Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker come of age. *Nat. Rev. Genet.* 4: 568-612.
- Jobling MA et al. (1996) Recurrent duplication and deletion polymorphism on the long arm of Y chromosome in normal males. *Hum. Mol. Genet.* 5: 1767-1775.
- Jobling MA et al. (2004) Human evolutionary genetics (first edition). Garland Science, New York.
- Jobling MA et al. (2013) Human evolutionary genetics (second edition). Garland Science, New York.
- 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.
- 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.
- Keinan A and Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336: 740-743.
- 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 diversity and the antiquity 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.
- Kumar S and Filipski AJ (2001) Molecular clock: Testing. In *Enciclopedia of Life Sciences*. Macmillan, London.
- Lahn BT and Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 46: 331-343.
- Lahn BT et al. (2001) The human Y chromosome, in the light of evolution. *Nat. Rev. Genet.* 2: 207-216.
- Lemey P et al. (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5: e1000520.
- Levinson G and Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4: 203-221.
- Li H and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.
- Li H et al. (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.
- Li R et al. (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19: 1124-1132.
- Lieberman DE (2002) The evolution and development of cranial form in *Homo sapiens*. *Proc. Natl. Acad. Sci. USA.* 99: 1134-1139.

- Lohmueller KE et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451: 994-997.
- Luddi A et al. (2009) Spermatogenesis in a man with complete deletion of *USP9Y*. *N. Engl. J. Med.* 360: 881-885.
- Luis JR et al. (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migration. *Am. J. Hum. Genet.* 74: 532-544.
- Macaulay V et al. (2005) Single rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308: 1034-1046.
- Manica A et al. (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448: 346-348.
- 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.
- McDougall I. et al. (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433: 733-736.
- Mellars et al. (2013) Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc. Natl. Acad. Sci. USA* 110:10699-10704.

- 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.
- Michaelson JJ et al. (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151: 1431-1442.
- Mishmar D et al. (2003) Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* 100: 171-176.
- 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.
- Mona S et al. (2007) Patterns of Y-chromosome diversity intersect with the Trans-New Guinea hypothesis. *Mol. Biol. Evol.* 24: 2546-2555.
- Morton NE (1991) Parameters of the human genome. *Proc. Natl. Acad. Sci. USA* 88: 7474-7476.
- 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.
- Myres NM et al. (2011) A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* 19: 95-101.
- Naidoo et al. (2010) Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investig. Genet.* 1: 6.

- Ning Z et al. (2001) SSAHA: A fast search method for large DNA databases. *Genome Res.* 11: 1725-1729.
- Nylander JAA et al. (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53: 47-67.
- 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.
- Pagani L et al. (2012) Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* 91: 83-96.
- Petraglia MD (2011) Trailblazers across Arabia. *Nature* 470: 50-51.
- Poznik GD et al. (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 341: 562-565.
- Pritchard JK et al. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16: 1791-1798.
- Prugnolle F et al. (2005) Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15: R159-160.
- Ramachandran S et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102: 15942-15947.
- Reich D et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468: 1053-1060.

- Rice WR (1987) Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* 116: 161-167.
- Roach JC et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328: 636-639.
- Rocca RA et al. (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 Genomes project data: an online community approach. *PLoS ONE* 7: e41634.
- 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.
- Rose JI et al. (2011) The Nubian Complex of Dhofar, Oman: an African middle stone age industry in Southern Arabia. *PLoS ONE* 6: e28239.
- Rosenberg et al. (2002) Genetic structure of human populations. *Science* 298: 2381-2385. *PLoS Genet.* 2: e215.
- Rosenberg et al. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India.

- Ross MT et al. (2005) The DNA sequences of the human X chromosome. *Nature* 17: 325-337.
- Rosser HZ 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 chromosome. *Nature* 423: 873-876.
- Rozen S et al. (2009) Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. *Am. J. Hum. Genet.* 85: 923-928.
- 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.
- Scally A and Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* 13: 745-753 and erratum in 13: 824.
- Scally A et al. (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483: 169-175.
- Schlebusch CM et al. (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338: 374-379.

- Schwartz JH and Tattersall I (2010) Fossil evidence for the origin of *Homo sapiens*. *Am. J. Phys. Anthropol.* 143 Suppl 51: 94-121.
- 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 et al. (2012) Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. *PLoS ONE* 7:e49170. doi:10.1371/journal.pone.0049170.
- 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. (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.

- 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.
- Shi W et al. (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol. Biol. Evol.* 27: 385-393.
- Simms et al. (2011) Paternal lineages signal distinct genetic contributions from British Loyalists and continental Africans among different Bahamian islands. *Am. J. Phys. Anthropol.* 146: 594-608
- Sindi S et al. (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.* 13: R22.
- 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.
- Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.
- Smith TM et al. (2007) Earliest evidence of modern human life history in North African early *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* 104: 6128-6133.
- Soares P et al. (2011) The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* 29: 915-927.

- Stringer C (2000) Coasting out of Africa. *Nature* 405: 24-27.
- Stringer C (2003) Out of Ethiopia. *Nature* 423: 692-695.
- Stringer CB and Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science* 239: 1263-1268.
- Takahata N et al. (2001) Testing multiregionality of modern human origins. *Mol. Biol. Evol.* 18: 172-183.
- 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.
- Tattersall I and Schwartz JH (2008) The morphological distinctiveness of *Homo sapiens* and its recognition in the fossil record: clarifying the problem. *Evol. Anthropol.* 17: 49-54.
- Thangaraj K et al. (2005) Reconstructing the origin of Andaman Islanders. *Science* 308: 996.
- The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.
- The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56-65.
- Thomson R et al. (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. USA* 97: 7360-7365.

- Thorvaldsdóttir H et al. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14: 178-192.
- Tishkoff SA et al. (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* 24: 2180-2195.
- Tishkoff SA et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324: 1035-1044.
- Trinkaus E (2005) Early modern humans. *Annu. Rev. Anthropol.* 34: 207-230.
- 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.
- Tyler-Smith C and Krausz C (2009) The will-o'-the-wisp of genetics – hunting for the azoospermia factor gene. *N. Engl. J. Med.* 360: 925-927.
- 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: 358-351.
- 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.
- Weale ME et al. (2003) Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. *Genetics* 165: 229-234.
- 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.
- White TD et al. (2003) Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* 423: 742-747.

- Whitfield LS et al. (1995) Sequence variation of the human Y chromosome. *Nature* 378: 379-380.
- Wilder JA et al. (2004) Genetic evidence for unequal effective population sizes of human females and males. *Mol. Biol. Evol.* 21: 2047-2057.
- Wolpoff MH et al. (1988) Modern human origins. *Science* 241: 772-774.
- 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 deep-rooting 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.
- Yu Y et al. (2010) S-DIVA (Statistical Dispersal-Vicariance Analysis): a tool for inferring biogeographic histories. *Mol. Phylogenet. Evol.* 56: 848-850.
- Zalloua PA et al. (2008) Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am. J. Hum. Genet.* 82: 873-882.
- Zhivotovsky LA et al. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* 74: 50-61.

Zhivotovsky LA et al. (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol. Biol. Evol.* 23: 2268-2270.

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.* 255: 428-435.

Appendices

APPENDIX 1.

List of 2,386 variable positions and inferred mutational events defining the branches of the tree shown in Figure 17.

The list will be available on the Genome Research website, in the Advance Online Articles section (<http://genome.cshlp.org/content/early/recent>) as Supplemental Table S2 of the paper “An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa” by Scozzari et al., published online as accepted preprint on January 6, 2014.

Variant positions are also deposited in dbSNP (handle: HUMGEN, ssid ss778077189-ss778079576).

APPENDIX 2.

List and genomic coordinates of the 5,274 fragments of the capture probe set covering the 5 selected MSY regions.

The list will be available on the Genome Research website, Advance Online Articles section (<http://genome.cshlp.org/content/early/recent>) as Supplemental Table S9 of the paper “An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa” by Scozzari et al., published online as accepted preprint on January 6, 2014.

Recent publications

1. Scozzari R, Massaia A, Trombetta B, Bellusci G, Myres NM, Novelletto A, Cruciani F. (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals ancient genetic events in Africa. *Genome Res.* 24: 535-544.
2. 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(11): e49170. doi:10.1371/journal.pone.0049170.
3. Cruciani F, Trombetta B, Massaia A, Destro-Bisol G, Sellitto D, Scozzari R. (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.
4. Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, Watson E, Beraud Colomb E, Dugoujon JM, Moral P, Scozzari R. (2010) Reply to Lancaster. *Eur. J. Hum. Genet.* 18: 1186-1187.
5. Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, Watson E, Beraud Colomb E, Dugoujon JM, Moral P, Scozzari R. (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.