MBE Advance Access published April 4, 2011

Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages

Research article

Chiara Batini,[1,2,15] Gianmarco Ferri,[3] Giovanni Destro-Bisol,[2,4] Francesca Brisighelli,[4,5,6] Donata Luiselli,[7] Paula Sánchez-Diz,[6] Jorge Rocha,[8] Tatum Simonson,[9] Antonio Brehm,[10] Valeria Montano,[1,2] Nasr Eldin Elwali,[11] Gabriella Spedini,[2,4] María Eugenia D'Amato,[12] Natalie Myres,[13] Peter Ebbesen,[14] David Comas,[1] Cristian Capelli,[5*]

[1]Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Doctor Aiguader 88, 08003, Barcelona, Spain

[2]Dipartimento di Biologia Ambientale, Sapienza Università di Roma, P.le Aldo Moro 5, 00185, Rome, Italy

[3]Dipartimento Integrato di Servizi Diagnostici e di laboratorio e di Medicina Legale, cattedra di Medicina Legale, Università di Modena e Reggio Emilia, via del Pozzo 71 41100 Modena

[4]Istituto Italiano di Antropologia, P.le Aldo Moro 5, 00185, Rome, Italy

[5]Department of Zoology, University of Oxford, South Parks Road, OX1 3PS, Oxford, UK

[6]Genomics Medicine Group, Institute of Legal Medicine, University of Santiago de Compostela, CIBER for Rare Diseases (CIBERER), c/San Francisco s/n. 15782, Santiago de Compostela, A Coruña, Spain

[7]Dipartimento di Biologia Evoluzionistica Sperimentale, Unità di Antropologia, Università di Bologna, Via Selmi, 3, 40126 Bologna, Italy

[8]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465, Porto, Portugal

[9]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA

[10]Human Genetics Laboratory, University of Madeira, Campus of Penteada, 9000-390, Funchal, Portugal

[11]Department of Basic Sciences, College of Medicine, Al Imam Mohamed Bin Saud Islamic University, P.O. Box 5701 Riyadh 11432, Kingdom of Saudi Arabia

[12]Forensic DNA Lab, Department of Biotechnology, University of the Western Cape. Private Bag X17, 7535 Bellville, South Africa.

[13]Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah 84115, USA

[14]Laboratory for Stem Cell Research, University of Aalborg, Frd. Bajers Vej 3B Aalborg Ø 9220, Denmark

[15]current address: Department of Genetics, University of Leicester, University Road, Leicester, LE1 7RH, UK

[*]corresponding author: Cristian Capelli, Department of Zoology, University of Oxford, South Parks Road, OX1 3PS, Oxford, UK; Telephone: +441865271261; Fax: +441865281843

cristian.capelli@zoo.ox.ac.uk

key words: Y chromosome; *Homo sapiens*; phylogeography; sub-Saharan Africa

running head: Early Y chromosome lineages in Africa.

Abstract

The study of Y chromosome variation has helped reconstruct demographic events associated with the spread of languages, agriculture and pastoralism in sub-Saharan Africa, but little attention has been given to the early history of the continent. In order to overcome this lack of knowledge, we carried out a phylogeographic analysis of haplogroups A and B in a broad dataset of sub-Saharan populations. These two lineages are particularly suitable for this objective because they are the two most deeply rooted branches of the Y chromosome genealogy. Their distribution is almost exclusively restricted to sub-Saharan Africa where their frequency peaks at 65% in groups of foragers. The combined high resolution SNP analysis with STR variation of their sub-clades reveals strong geographic and population structure for both haplogroups. This has allowed us to identify specific lineages related to regional pre-agricultural dynamics in different areas of sub-Saharan Africa. In addition, we observed signatures of relatively recent contact, both among Pygmies, and between them and Khoisan speaker groups from southern Africa, thus contributing to the understanding of the complex evolutionary relationships among African hunter-gatherers. Finally, by revising the phylogeography of the very early human Y chromosome lineages, we have obtained support for the role of southern Africa as a sink, rather than a source, of the first migrations of modern humans from eastern and central parts of the continent. These results open new perspectives on the early history of *Homo sapiens* in Africa, with particular attention to areas of the continent where human fossil remains and archaeological data are scant.

Introduction

In the last few decades the analysis of genetic variation in human populations has increased exponentially and has provided significant insights on the history of our species (Destro-Bisol et al. 2010; Renfrew 2010). One of the most frequently replicated results has been the support of the "Recent Out of Africa" model, initially based on mitochondrial DNA (mtDNA; Cann, Stoneking, Wilson 1987) and later gaining support from other genomic regions (Underhill et al. 2001; Rosenberg et al. 2002; Li et al. 2008). Systematic investigation of the genetic diversity in African populations focusing on mtDNA (Salas et al. 2002; Behar et al. 2008), Y chromosomes (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007) and autosomal regions (Tishkoff et al. 2009) has started to provide insights on African-specific demographic events. However, whilst mtDNA variation has been thoroughly investigated by detailed dissection of the most informative lineages (Salas et al. 2002; Gonder et al. 2007; Behar et al. 2008), and, more recently, autosomal variation has begun to be explored in detail (Tishkoff et al. 2009), such a level of resolution has been only partially applied to Y chromosome African haplogroups. Sub-Saharan African Y chromosome diversity is represented by five main haplogroups (hgs): A, B, E, J and R (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007). Hgs J and R are geographically restricted to eastern and central Africa, respectively, while hg E shows a wider continental distribution (see also Berniell-Lee et al. 2009; Cruciani et al. 2010). Despite the phylogeographic dissection of Hg E is still ongoing, it has been suggested that this clade might be linked, at least in part, with the diffusion of agriculture and pastoralism in the continent during the last 4-5 thousand years, as initially indicated by its parallel distribution to Bantu-speaking communities (Underhill et al. 2001; Henn et al. 2008). The other two lineages, A and B, represent the most basal branches within the human Y chromosome genealogy and are dispersed across different geographic areas and populations, with considerably high frequencies in hunter-gatherer populations. These hgs have been related to demographic dynamics that are independent to the recent introduction of practices for active food production mentioned above, thus suggesting an association with complex and potentially more ancient

demographic events (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007; Berniell-Lee et al. 2009).

In this work we present a detailed phylogeographic dissection of hgs A and B in a broad dataset of sub-Saharan populations, with the aim of providing new insights into the complex and poorly investigated dynamics that characterize the pre-agricultural history of sub-Saharan Africa, with special attention given to the relationships among Pygmy and Khoisan-speaking populations from southern Africa. In addition, we aim to contribute to the debate on the geographic origin of *Homo sapiens* in Africa by testing whether the male specific signals of early human origins are retained only among communities from eastern Africa (as suggested by fossil remains and mitochondrial DNA; White et al. 2003; McDougall, Brown, Fleagle 2005; Behar et al. 2008) or whether they can also be found within groups from southern Africa (as indicated by genome-wide scans and early Y chromosome analyses; Hammer et al. 2001; Semino et al. 2002; Hellenthal, Auton, Falush 2008; Tishkoff et al. 2009).

Materials and Methods

SNPs and STRs genotyping

A database of 641 chromosomes (Table S1) was generated by collecting previously published data, analysing novel samples and extending the molecular analysis of previously genotyped samples. All DNA samples were obtained from blood, buccal swabs or saliva samples and collected from unrelated healthy individuals who gave the appropriate informed consent.

Samples were genotyped with different sets of markers (Table S1). SNP scoring was carried out using mini-sequencing multiplex reactions and direct sequencing. A total of 33 markers were selected within haplogroups A and B according to the most updated Y chromosome genealogy presented in Karafet et al. 2008. These were divided among four different Single Base Extension (SBE) assays, here referred to as MAI, MAII, MB and MB2b (see Table S2). Primers for multiplex PCR amplification were designed using Primer3Plus software (Untergasser et al. 2007) and are presented in Table S3 and Table S4. Self- and cross- compatibility among all primers pairs included in the same reaction were tested with the software Autodimer (see Web resources section). Y chromosome specificity of each primer was tested using BLASTn (Basic Local Alignment Search Tool).

The Qiagen Multiplex PCR kit and conditions specified by the producer were applied with primer concentrations ranging between 0.15 μM and 0.8 μM. PCR products (1.5 μl) were cleaned using 1.5 μl of ExoSAP-IT (USB Corporation) for 15 min at 37° C followed by 15 min at 80° C.

Minisequencing SBE primers were selected using ASPE tools in the NIST Online DNA Analysis tools Page (see Web resources section) and non-specific tails of different lengths were added to each in order to ensure complete capillary separation of SNaPshot products (supplementary Table S5 and Table S6). The multiplex minisequencing assays were performed using 1 μl of purified product in a total volume of 5 μl using 2 μl of SNaPshot reaction mix (AB) according to the SNaPshot Kit protocol. Fluorescently labelled ddNTPs in excess were inactivated and 1 μl of cleaned multiplex extension products was run on a ABI PRISM 3130 Genetic Analyzer. Allele calling was performed

using GeneMapper software (v. 3.7; Applied Biosystems Carlsbad, CA, USA).

Direct sequencing was used to screen markers P108 and P114. Primers for amplification are reported in Table S3. Amplification of MSY2 was carried out according to Bao et al. 2000.

STR genotyping was conducted using commercially available STR kits (Krenke et al. 2005; Mulero et al. 2006) as well as multiplexes developed in-house (Beleza et al. 2003). All the samples included here were genotyped for 10 STRs: DYS19, DYS389-I, DYS389-II (the allele reported in Table S1 has been obtained by subtracting the DYS389-I allele), DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439. A subset of the samples was tested for an additional 5 loci (DYS448, DYS456, DYS458, DYS635 and Y-GATA-H4). In the statistical analyses specific loci (DYS385, DYS389-II, DYS390, DYS448, DYS635) were excluded due to allelic homoplasy as reported in the NIST Y-STR Fact Sheets (see Web resources section). Following this, eight STR loci were used in both phylogeographic and intra-lineage analyses in order to maintain broad population coverage.

Network reconstruction and diversity estimation. Median-Joining networks (Bandelt, Forster, Rohl 1999) of both SNP and STR haplotypes were constructed using Network 4.5 (see Web resources). Weights were estimated using the inverse of the within-clade variances of individual STR loci. SNPs were weighted according to their hierarchical position in the genealogy identified in the present paper (see Figure S2c, d). Within-hg diversity was investigated using Arlequin 3.0 (Excoffier, Laval, Schneider 2005). The variance was estimated as the within–locus mean allele variance averaged across all loci. Confidence intervals (CI) were based on 10,000 resamplings performed across individuals. Samples showing missing data at any locus were not considered in the calculation of intra-lineage variation parameters.

Dating. The between- and within- lineage date estimates were obtained by using the model-free statistics Average Squared Distance (ASD; Goldstein et al. 1995a; Goldstein et al. 1995b). An indication of the time of lineage split can be obtained using ASD calculated between lineages (ASD=2μT; Goldstein et al. 1995a; Goldstein et al. 1995b). ASD is based on a strict single stepwise

mutation model and in the presence of multi-step mutational events the squaring process is expected to heavily influence the distance estimation, corrupting the linearity with time. In order to take into account such occurrences and avoid the impact of multi-step mutations, we calculated the expected ASD asymptotic value (E[ASD]) (Goldstein et al. 1995b) as an indication of the maximum expected ASD value per locus comparison. These values were used as locus-specific thresholds to identify and remove STR markers potentially showing between lineage multi-step mutational events. Mutation rate is a critical factor influencing the extension of ASD time-linearity. To control for this, we selected the set of eight markers among those available after multistep removal that showed the lowest mutation rate (based on the data presented on the YHRD webpage, release33; Willuweit, Roewer, International Forensic Y Chromosome User Group 2007; see also Table S7), for each inter lineage comparison. In order to compare inter- and intra-lineage estimates, we used the same number of STRs (eight) for the within-lineage estimates (see below). ASD upper-limit linearity with time can be estimated as described in Goldstein et al. 1995b. Simulations have shown that the expected values tend to overestimate the range of linearity and only provide a broad indication of the upper limit of ASD linearity with time (Goldstein et al. 1995b). We used these values as reference thresholds to ensure that all the between-lineage estimates reported in Table 1a do not cross these boundaries. The starting set of markers comprised the eight STRs used for Network analysis and diversity estimates, and were extended to eleven by including DYS456, DYS458 and YGATA-H4 loci. Due to multistep correction, different sets of STRs were used (Table S7) and the average mutation rate was estimated using locus-specific values (YHRD, release33; Willuweit, Roewer, International Forensic Y Chromosome User Group 2007). The reported 95% confidence intervals were estimated by averaging across the locus specific upper and lower mutation rate estimates (YHRD, release 33; Willuweit, Roewer, International Forensic Y Chromosome User Group 2007). Given the limitation related to ASD saturation, some potentially interesting inter-lineage comparisons were beyond the available resolution dictated by the STRs we used. For example, the ASD between A and B clades, which is expected to give an estimate of the

Time to the Most Recent Common Ancestor (TMRCA) of the entire human Y chromosome genealogy. In order to provide an independent estimate for the TMRCA of a pair of lineages, we also used a Bayesian approach as described in Walsh 2001 and implemented in the software ASHEs (Tofanelli et al. 2009). In brief, this approach calculates the likelihood distribution of the TMRCA for each haplotype-haplotype comparison across n generations. In our analysis the following parameters were used : $\lambda(1/Ne)=0.0002$ (Walsh 2001), 10,000 generations and the same set of STRs/mutation rates as for the corresponding ASD calculations. The Maximum Likelihood estimations (ML) of the number of generations to the most recent common ancestor were collected for each run and the average of these values used to obtain an indication of lineage separation. To calculate the CI, the same procedure was repeated by using the average upper and lower estimates for the locus-specific mutation rate, which was also performed for the ASD based estimates.

The TMRCA of a clade was estimated by calculating the ASD between all chromosomes in a lineage, and the founder haplotype which we reconstructed by combining the modal alleles at single loci (Thomas et al. 1998). ASD estimated in this way has an expected value of μT, where μ is an average effective mutation rate at the loci and T is the separation time expressed in number of generations. This approach is expected to underestimate the age of the clade if the reconstructed founder haplotype differs from the true one. The 95% confidence intervals were estimated using the software Ytime, based on a constant size demographic model (Behar et al. 2003). The locus-specific mutation rate was estimated using data from the YHRD, release 33 (Willuweit, Roewer, International Forensic Y Chromosome User Group 2007). We focused on the same set of eight STRs used in the Network analyses. For estimates within hg A1, we removed the locus DYS438 due to its multistep behaviour within this lineage and performed estimates on the remaining seven STRs. It should be also noted that many of these lineages are particularly rare and that the within-clade variation might have been only partially surveyed, a condition that may divert current estimates towards the lower bound of the real genealogical depth (see Petraglia et al. 2010). For all estimates a generation time of 31 years was used (Helgason et al. 2003). The average mutation rate

used for the dating estimates ranges between 1.6 to 2.2* 10^{-3} mutations per locus per generation depending on which set of STR markers was used (Table S7). These values are not substantially different from other estimates based on pedigree data, and are approximately 2-3 faster than the more general and non-locus specific 'evolutionary' rate (6.9 * 10^{-4} mut/locus/gen; Zhivotovsky et al. 2004; see also Ravid-Amir, Rosset 2010).

Results.

Hgs distribution and variation. We genotyped both novel and previously partially investigated samples and surveyed literature data for a total of approximately 10,000 males from more than 180 populations (Table S8), collecting data for 184 hg A and 457 hg B Y chromosomes (Table S1). Outside Africa, these clades have been sporadically found in Europe and the Americas, probably as a result of recent migrants (Semino et al. 2000; Luis et al. 2004; Capelli et al. 2006; Hammer et al. 2006; King et al. 2007). Hg A is rarely found in North, West and Central Africa while it is more frequent in the eastern and southern parts of the continent (Figure 1). Rare in both northern and western Africa, the distribution of hg B in the rest of the continent can be described by that of its two main sub-clades B2a and B2b (Figure 1). The former appears to be associated with food-producing communities and populations in contact with them, as also previously observed (Beleza et al. 2005; Berniell-Lee et al. 2009; Gomes et al. 2010) and it is present at low frequencies in all sub-Saharan areas. In contrast, B2b is mostly present in foraging communities in eastern and central Africa. The different geographic distributions of hgs A and B2b are mirrored at the population level (Figure 1 and Table S8). Little or no hg A is present in Pygmies and eastern African (EA) Khoisan speakers (for the use of the word Khoisan, issues with populations classification in southern Africa and the case of eastern Khoisan speakers see Mitchell 2010), while B2b is commonly found in these populations. On the other hand, hg A is more frequent than B among southern African (SA) Khoisan speakers (~40%), with B2b representing approximately 16% of the Y chromosome types present in these populations (Figure 1 and Table S8).

Diversity indices are shown in Table 2. Overall, hg A shows higher diversity than B and, within the latter, B2b is more variable than B2a. Network analysis based on 8-STRs haplotypes shows substantial phylogeographic patterns for A and B2b hgs (data not shown), while hg B2a reveals no clear population/geographic structure and a high level of reticulation, which is expected for lineages with a relatively short evolutionary history, associated with recent demographic expansions (Table

1b and Figure S1; see also Beleza et al. 2005; Berniell-Lee et al. 2009; Gomes et al. 2010). These results, together with the virtual absence of B2a in foraging populations, supports our decision to focus the phylogeographic analysis on hgs A and B2b only, in order to address questions related to the early history of sub-Saharan Africa. The evolutionary relationships among haplotypes within these hgs, based on both SNPs and STRs, are shown in Figure 2.

Of the A sub-clades, A1 is found only in western and central Africa, while A3b1 and A3b2 are southern and central/eastern African specific, respectively. Hg A2 is mostly represented by SA samples with only a few central African haplotypes (Figure 2a, c). Similarly, B2b1/B2b4a and B2b2 are geographically restricted to southern and eastern Africa, respectively, while B2b3, B2b4b and B2b4* (as well as the previously undescribed MSY2* lineage; Figure S2d) are specific to central Africa, albeit with few B2b4* SA haplotypes (Figure 2b, c). A prevalence of EA chromosomes is observed within B2b* together with considerable variation at the haplotype level, suggesting the possibility of yet undetected SNP-defined sub-clades within this group (Figure 2b, c). The geographically structured distribution within the B2b clade is shaped by the presence of population-specific lineages (Figure 2b, c). In fact, while B2b3, B2b4b and B2b4* are almost exclusively found among western Pygmies, B2b2 and B2b1-B2b4a are found only in eastern Pygmies and SA Khoisan speakers, respectively. Similarly, the majority of the A3b1 and A2 types are found among SA Khoisan speakers, with hg A2 also present in western Pygmies (Figure 2a, c). Pygmies and SA Khoisan speakers also share evolutionarily closely related lineages within the B2b4 clade (Figure 2b, c; see also Wood et al. 2005) .

A and B genealogies. Our extensive survey of SNP variation in hgs A and B Y chromosomes enabled us to detect genealogical incompatibilities and propose some refinements within the recently proposed topology (see Figure S2 for a comparison with the trees by Karafet et al. 2008). The PK1 marker, originally thought to be associated with the A2 lineage only, was found to cluster both A2 and A3 chromosomes. Similarly, new A2 lineages have been identified (Figure S2c). M190

had been indicated as A3b-specific (Karafet et al. 2008); however, our analysis showed that it is derived in all A3 lineages. Within hg B, P7 appears to be basal to most of the B2b lineages and, within the P7 derived chromosomes, the MSY2 marker clusters lineages defined by M211, M115/M169, M30/M129 variants (Figure S2d). The identification of two chromosomes derived at MSY2 and M30 (one of which is also derived for M129) but not for P7 suggests that this polymorphism might be prone to recurrent mutations (see Figure 2c). The physical proximity of P7 to P25 makes Y-Y gene conversion a possible explanation for this finding (see Adams et al. 2006). For simplicity, we have retained the same nomenclature as recently described in Karafet et al. 2008 (with the exception of MSY2*, see above) but renaming will be necessary as more data become available.

Discussion

Insights into the male genetic history of sub-Saharan hunter-gatherers

Pygmy groups. We dated the Eastern-Western Pygmy separation using the divergence between the B2b2 and B2b4b/B2b3 clades (Table 1a). These estimates span similar time intervals and suggest a separation time of 10-15 thousands years ago (Kya), broadly overlapping with the generally more ancient estimates provided by mtDNA and autosomal data (Destro-Bisol et al. 2004; Patin et al. 2009; Batini et al. 2011). In particular, the youngest date suggested by B2b2 vs. B2b3 (10.7 [3.5-17.1] Kya ; 10.5 [6.8-16.5] Kya) might indicate post Last Glacial Maximum (LGM; 19-26.5 Kya; Clark et al. 2009) male-mediated contacts between the two groups. This could account for the contrast between the lack of shared recent mitochondrial ancestry among Pygmy populations (Batini et al. 2011) and the quite intense post-LGM gene-flow suggested by autosomal loci (Patin et al. 2009). However, the uncertainty related to STRs choice and their time-linearity suggests that older scenarios might not be excluded (Busby G., Capelli C., personal communication). We also note that the within-clade diversity/antiquity is extremely reduced for these Pygmy-specific lineages, suggesting a bottleneck in the relatively recent demographic history of these groups, as it has been observed for other loci (see Table 1b; Weiss and von Haeseler 1998; Excoffier and Schneider 1999; Patin et al. 2009; Batini et al. 2011).

Pygmies and San. We identified evolutionary links between western Pygmies and San in both A and B clades, developing the initial findings presented in Wood et al. (2005). Hg A2, found among SA Khoisan speakers at 25-45% (Wood et al. 2005; Table S9), was detected for the first time in the present work at non trivial frequency (5%) among the Baka Pygmies from Cameroon and Gabon. On the other hand, B2b4 was present at 6-7% among Khoisan speakers but reached 45-67% in both Biaka and Baka Pygmies (Wood et al. 2005; Table S9). We dated the TMRCA among the western Pygmies and San specific sub-clades of these two haplogroups to between 3 and 4 Kya (CI 1.9-4.6 Kya and 2.2-5.4 Kya for A2; 2.3.-5.8 Kya and 2.8-6.9 Kya for B2b4; Table 1a). It should be pointed

out that the large number of mutations specific to the Khoisan A2 lineage (see Figure 2a, 2c) is probably the result of the SNP discovery process, which included Khoisan but not Western Pygmies A2 chromosomes (Wood et al. 2005), thus making the use of STRs for dating the most obvious choice. Evidence for a Pygmy/San link has also been provided by recent genome-wide studies. In the work presented by Hellenthal, Auton, Falush (2008) the first genetic link to emerge among human populations was indeed between the San and western Pygmies. Furthermore, a shared ancestry between San and eastern Pygmies has been observed recently, and more generally, it has been seen between western Pygmies and the Hadza from Tanzania (Tishkoff et al. 2009), even though this has been interpreted as the result of a possibly more ancient common genetic background than the one suggested by our results. Intriguingly, the genetic link seems to be paralleled by the sharing of cultural traits such as those found in the rock art geometric designs produced by Pygmies from the Ituri forest and the Khoe-speaking groups from southern Africa (Smith 1995; Smith 1997; Smith and Ouzman 2004; Smith 2006). According to this model, Khoe-speaking pastoralists would have moved from an area in Central-South Africa bringing pastoralism into southern Africa before the Bantu dispersion in the region, having previously experienced cultural and genetic exchanges with central and eastern African populations (Henn et al. 2008; Rocha 2010).

Genetic evidence for the peopling of sub-Saharan Africa before the diffusion of agriculture

West Africa. Haplogroup A in western Africa is represented only by the A1a lineage. The variation within this clade dates back to 10.5 [4.2-23.7] Kya, and to 8 [3.1-19.4] Kya when only western African haplotypes are considered (see Table 1b), which is in agreement with the archaeological and linguistic evidence related to the peopling of this region. The Ounanian culture has in fact been recorded in Mali as far back as 9-10 Kya (Clark 1980; Raimbault 1990; Mac Donald 1998) and the lithic and ceramic assemblages from Ounjougou date back to 12 Kya (Huysecom et al. 2004; Huysecom et al. 2009). Similarly, the origin of the early Niger-Congo Atlantic branch has been

placed at least 8 Kya (Ehret 2000; Blench 2006). The detection of a specific genetic signal associated with early human presence in this area is of interest given the homogeneity between western and central African populations that has been observed so far for genome-wide analysis (Cruciani et al. 2002; Wood et al. 2005; Tishkoff et al. 2007; Li et al. 2008; Tishkoff et al. 2009).

South Africa. We dated variation in SA hgs A2 and A3b1 to 6.2 [2.2-14.1] Kya and 10.2 [4.4-23] Kya, respectively (Table 1b). These dates do not extend beyond the LGM which contrasts with the early human presence in southern Africa suggested by fossil and archaeological remains (McBrearty and Brooks 2000; White et al. 2003; Lewin and Foley 2004; McDougall, Brown, Fleagle 2005; Marean et al. 2007). This could be possibly due to our partial population coverage, as suggested by extensive population surveys (Quintana-Murci et al. 2010; Marks S., Capelli C., unpublished data), as well as to past lineage extinctions (see Petraglia et al. 2010) that followed the significant demographic changes during the Marine Isotope Stage 3 (25-60 Kya) and the LGM (Mitchell 2008). Moreover, the possible limitation of available STRs in exploring events dating further back in time may also have had an effect (Busby G., Capelli C., personal communication). It is also worth considering the possibility that A2 and A3b1 retain signatures of two independent pre-Bantu dispersal events in the region. This scenario is also supported by the different geographic distribution of these two clades: A3b1 is present across all of southern Africa while A2 is almost exclusively associated with populations in south-western Africa, or those originally from this area (Table S1 and Table S9; see also Table 2 in De Filippo et al. 2010 and unpublished data from Lesotho and additional South African populations, where A3b1 but not A2 chromosomes were found - Marks S., Capelli C., personal communication). The A2 distribution broadly overlaps that of Khoe-speakers and could potentially represent a genetic signature of the contacts/migrations of the Khoe-speaking pastoralist societies from northern Botswana, southern Angola and western Zambia area, approximately 2 Kya (see also above; Mitchell and Whitelaw 2005).

South-East Africa. Hg B2b4* chromosomes were present in the Mozambican samples, a lineage

which is mainly shared with Baka Pygmies from Cameroon. The low frequency of these chromosomes in the southern and eastern African populations, together with the lack of appropriate evidence of a link among early inhabitants of these regions with western Pygmies, leaves the issue difficult to disentangle and calls for more detailed and focused investigation. In this sense, a scenario worth exploring could be based on the presence of this lineage in pre-Bantu populations already settled in the regions, which could either have been absorbed by the incoming agro-pastoralist groups (Sikora et al. 2010), or reflect the broader network of contacts around central-southern Africa (see above).

East Africa. The sub-clade A3b2 is present at high frequencies in Eastern African populations, in particular among Nilo-Saharan speakers. Based on the analysis of this lineage in Uganda, Gomes et al. 2010 proposed its association with this linguistic phylum. Our estimates of A3b2 antiquity (9 Kya; CI 3.7-20.2 Kya) do not refute this hypothesis, as they are broadly in agreement with the initial date for the spread of Nilo-Saharan phylum approximately between 12 and 18 Kya (Ehret 2000; Blench 2006).

B2a as a marker of the Bantu expansion? Although B2a has not been investigated with the same resolution as the A and B2b hgs, our data support its association with Bantu-speaking populations, as previously reported (see Table S1; Beleza et al. 2005; Berniell-Lee et al. 2009). Within-clade variation suggests a more recent origin for B2a than B2b, while network analysis did not reveal population specific or geographically localized STR-based clusters (Figure S1). However, the relatively deep within-clade dating (6.1 [2.2-14] Kya) suggests a scenario possibly pre-dating the diffusion of Bantu languages, in line with what has been observed for some sub-clades of hg E (Montano V., Destro-Bisol G., Comas D., personal communication). Deeper phylogenetic resolution within the B2a clade, coupled with additional population sampling, may help to clarify the demographic dynamics associated with its dispersal.

The emergence of modern humans

Whereas the dissection of single Y-chromosomal clades or sub-clades has helped define the relationships between specific populations/groups, as well as reconstruct the demographic impact of migratory and cultural events, a wider and exhaustive phylogeographic analysis may indicate areas of the African continent where the extant human Y chromosome diversity first originated. Haplogroups A and B are ideal candidates for this task, given their distribution in Africa and the fact that they represent the earliest lineages to branch off within the Y chromosome genealogy. Previous analysis of the Y chromosome variation pointed to a SA/EA origin following the identification of hg A3b and, to a lower extent, B types in populations from these areas (Hammer et al. 2001; Semino et al. 2002). However, our results clearly indicate that A3b branched later within hg A, making it uninformative on the origin of the early human Y lineages. Hg A is divided into two branches: A1, represented by western and central African types, and A2-A3, containing southern and eastern African chromosomes, with a few from central Africa. Hg A2 is mostly composed of southern Africa types; however, an early branch in A2 is found in central Africa. Within hg A3, A3b1, the southern Africa clade, is a sister clade to A3b2, common in eastern Africa, while A3a is only found among eastern Africans (Figure 2c). In hg B, B2a and B2b are two sister clades, while B*(xB2) aggregates a number of chromosomes from central Africa which were ancestral for the set of SNPs we tested. B2a has a very wide distribution and is mainly present in Bantu-speaking populations. Within hg B2b, B2b* contains samples from eastern, south-eastern and central Africa, with P6-derived chromosomes from South Africa and P7 types mainly from hunter-gatherer populations from central, eastern and southern Africa (see Figure 2c). These results seem indicate that southern Africa was an early destination of ancient human migrations from other regions other than the original source, which fails to support the hypothesis presented in a recent large-scale study of autosomal loci (Tishkoff et al. 2009). With respect to the roles of eastern and central Africa, the dataset presented here, while tentatively pointing towards a wide-scale preservation of ancient lineages in central Africa, is still compatible with a primary role for eastern Africa, in agreement

with hypotheses generated from both mtDNA analysis and the study of the earliest *Homo sapiens* fossil remains (White et al. 2003; McDougall, Brown, Fleagle 2005; Behar et al. 2008).

Concluding remarks

Detailed phylogeographic analysis of human Y chromosome hgs A and B, combined with a large population survey and extensive sub-lineages characterisation, has allowed us to gain new insights into the processes which shaped the pre-agricultural peopling of the African continent. Our results provide a male specific perspective on some key aspects of the genetic history of sub-Saharan Africa and form the basis for future research.

We have shown evidence for further complexity in the evolutionary relationships among African hunter-gatherers. Phylogeographic analyses of mtDNA point to an ancient separation among ancestral populations, with limited or no subsequent gene flow after the split (see Salas et al. 2002, Destro-Bisol et al. 2004, Batini et al. 2007, Quintana-Murci et al. 2008, Behar et al. 2008, Batini et al. 2011). Conversely, the analysis of autosomal loci suggests a common, and possibly more recent, genetic background (see Tishkoff et al. 2009), with contrasting evidence concerning the reciprocal relationships among Pygmies and San (see Li et al. 2008, Hellenthal, Auton, Falush 2008, Tishkoff et al. 2009), although this lacks a well defined temporal context. Our extensive phylogeographic and dating approach has provided evidence for relatively recent contact both among Pygmies and between them and San groups from southern Africa. Our current estimates for the coalescent time between Eastern and Western Pygmy specific Y chromosome clades (10-15 Kya) are compatible with post-LGM contact among the two groups, with evidence for recent bottlenecks in the demographic histories of the two groups (see also Patin et al. 2009, Batini et al. 2011). Otherwise, the very recent common ancestry detected among Western Pygmies and San (3-4 Kya) suggests that this could be the signature of Khoe-speaking pastoralist mediated contact among the two groups, rather than resulting from retention of ancient traits.

Lastly, the peopling of sub-Saharan Africa has been studied from linguistic, archaeological and genetic perspectives in the last decade, but its most ancient period is not yet well understood (see Campbell, Tishkoff 2010, Scheinfeldt, Soi, Tishkoff 2010). We have highlighted some signatures of pre-agricultural peopling undetected by previous research work. In fact, West, East and South

African populations show specific clades whose TMRCAs are compatible with a differentiation pre-dating the arrival of Bantu-speaking people and farming in the area. Intriguingly, even B2a, which has been mainly found in Bantu-speaking communities, has been dated (6 [2-14] Kya) before the supposed time of diffusion of Bantu languages. A novel link among Pygmy hunter-gatherers from west-central Africa and farmers from Mozambique has been identified, pointing to a shared genetic legacy between these two geographically separate and anthropologically distinct population groups (see also Sikora et al. 2010).

Finally, our study contributes to the debate on the geographical origin of *Homo sapiens* in sub-Saharan Africa, providing evidence for the retention of early Y chromosome lineages in East and Central but not in Southern Africa. However, we note that the current absence of significant palaeo-anthropological investigation, together with the possibility of different fossil preservation conditions in central Africa, makes the extremely long human fossil record in eastern Africa inconclusive in solving this issue. The screening of Y-chromosomal variation at an increased level of resolution, combined with additional sampling from these regions, is expected to further elucidate the early steps of *Homo sapiens* in Africa.

Supplementary Material

Supplementary material includes one pdf file with two figures and five tables, and four supplementary excel tables (S1, S7, S8 and S9).

Acknowledgments

We would like to thank: Sergio Tofanelli and Davide Merlitti for giving access to early versions of the ASHEs software; Jim Wilson, Fabio Verginelli and Renato Mariani-Costantini for providing samples and unpublished data; Peter Mitchell for helpful discussions on the African archaeological record; Marco Giorgi and Isabel Mendizabal for providing scripts used during data analysis; Mònica Vallés, Stéphanie Plaza, and Roger Anglada (UPF) and Milena Alu'(Università di Modena e Reggio Emilia) for technical support. Finally we would like to express our gratitude to all the people that have made this work possible by donating their DNA.

The research presented was supported by the Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain (CGL2007-61016), and Direcció General de Recerca, Generalitat de Catalunya (2009SGR1101). GDB and GS were supported by the University of Rome "La Sapienza" (grant prot. C26A09EA9C/2009). JR was supported by the FCT (grant PTDC/BIA-BDE/68999/2006). PSD is supported by the Isidro Parga Pondal program [Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica-INCITE (2006–2010) from Xunta de Galicia,Spain]. CC is a RCUK Academic Fellow.

CB, CC designed the research. CB, GF, DC, CC conceived and designed the experiments. GDB, DL, JR, LJ, AB, VM, NEE, GS, MEDA, NM, PE, DC provided the samples and part of the genotypings. CB, GF, FB, PSD performed the experiments. CB and CC analyzed the data. CB and CC wrote the paper with the contribution of GDB and DC. All co-authors have reviewed the manuscript prior to submission.

Web resources

Autodimer: http://cstl.nist.gov/

NIST Online DNA Analysis tools page: http://yellow.nist.gov:8444/dnaAnalysis/index.do

Y-STR Fact Sheets: http://www.cstl.nist.gov/strbase/ystr_fact.htm

Network 4.5: www.fluxus-engineering.com

ASHEs: http://ashes.codeplex.com/

References

Adams SM, King TE, Bosch E, Jobling MA. 2006. The case of the unreliable SNP: Recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. Forensic Sci. Int. 159:14-20.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37-48.

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, van der Veen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D. 2011. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol. Biol. Evol. 28:1099-1110.

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F. 2007. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol. Phyl. Evol. 43:635-644.

Bao W, Zhu S, Pandya A, Zerjal T, Xu J, Shu Q, Du R, Yang H, Tyler-Smith C. 2000. MSY2: A slowly evolving minisatellite on the human Y chromosome which provides a useful polymorphic marker in Chinese populations. Gene 244:29-33.

Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D et al. (12 co-authors). 2003. Multiple origins of Ashkenazi Levites: Y chromosome evidence for both near eastern and European ancestries. Am. J. Hum. Genet. 73:768-779.

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H,

Tzur S, Comas D et al. (15 co-authors). 2008. The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82:1130-1140.

Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A. 2005. The genetic legacy of western Bantu migrations. Hum. Genet. 117:366-375.

Beleza S, Alves C, Gonzalez-Neira A, Lareu M, Amorim A, Carracedo A, Gusmao L. 2003. Extending STR markers in Y chromosome haplotypes. Int. J. Legal Med. 117:27-33.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D. 2009. Genetic and demographic implications of the Bantu expansion: Insights from human paternal lineages. Mol. Biol. Evol. 26:1581-1589.

Blench R. 2006. Archaeology, language, and the African past. Lanham, MD ; Oxford: AltaMira Press.

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

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31-36.

Capelli C, Redhead N, Romano V, Cali F, Lefranc G, Delague V, Megarbane A, Felice AE, Pascali VL, Neophytou PI et al. (18 co-authors). 2006. Population structure in the Mediterranean basin: A Y chromosome perspective. Ann. Hum. Genet. 70:207-225.

Clark JD. 1980. Human populations and cultural adaptations in the Sahara and the Nile during prehistoric times. In: William MAJ, Faure H, editors. The Sahara and the Nile: quaternary environments and prehistoric occupation on Northern Africa. Rotterdam: Balkema. p. 527-582.

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM. 2009. The last glacial maximum. Science 325:710-714.

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.

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V et al. (16 co-authors). 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.

De Filippo C, Heyn P, Barham L, Stoneking M, Pakendorf B. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. Am. J. Phys. Anthropol. 141: 382-394.

Destro-Bisol G, Jobling MA, Rocha J, Novembre J, Richards MB, Mulligan C, Batini C, Manni F. 2010. Molecular anthropology in the genomic era. J. Anthropol. Sci. 88:93-112.

Destro-Bisol G, Coia V, Boschi I, Verginelli F, Caglia A, Pascali V, Spedini G, Calafell F. 2004. The analysis of variation of mtDNA hypervariable region 1 suggests that eastern and western pygmies diverged before the Bantu expansion. Am. Nat. 163:212-226.

Ehret C. 2000. Language and history. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 272-297.

Excoffier L, Schneider S. 1999. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. Proc. Natl. Acad. Sci. U. S. A. 96:10597-10602.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47-50.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995a. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. U. S. A. 92:6723-6727.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995b. An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471.

Gomes V, Sanchez-Diz P, Amorim A, Carracedo A, Gusmao L. 2010. Digging deeper into east African human Y chromosome lineages. Hum. Genet. 127:603-613.

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. 24:757-768.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol. Biol. Evol. 18:1189-1203.

Hammer MF, Chamberlain VF, Kearney VF, Stover D, Zhang G, Karafet T, Walsh B, Redd AJ. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. Forensic Sci. Int. 164:45-55.

Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefansson K. 2003. A population wide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: Evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. Am. J. Hum. Genet. 72:1370-1388.

Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a copying model. PLoS Genet. 4:e1000078.

Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc. Natl. Acad. Sci. U. S. A. 105:10693-10698.

Huysecom E, Ozainne S, Raeli F, Ballouche A, Rasse M, Stokes S. 2004. Ounjougou (mali): A history of Holocene settlement at the southern edge of the Sahara. Antiquity 78:579-593.

Huysecom E, Rasse M, Lespez L, Neumann K, Fahmy A, Ballouche A, Ozainne S, Maggetti M, Tribolo C, Soriano S. 2009. The emergence of pottery in Africa during the tenth millennium cal BC: New evidence from Ounjougou (Mali). Antiquity 83:905-917.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 18:830-838.

King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA. 2007. Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. Eur. J. Hum. Genet. 15:288-293.

Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, Gross AM, Gornall T, Frappier JR, Eisenberg AJ et al. (14 co-authors). 2005. Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. Forensic Sci. Int. 151:111-124.

Lewin R, Foley R. 2004. Principles of human evolution. UK: Wiley-Blackwell.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100-1104.

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ. 2004. The Levant versus the horn of Africa: Evidence for bidirectional corridors of human migrations. Am. J. Hum. Genet. 74:532-544.

Mac Donald KC. 1998. Archaeology, language and the peopling of west Africa: A consideration of the evidence. In: Blench R, Spriggs M, editors. Archaeology and Language II: correlating archaeological and linguistic hypotheses. London: Routledge. p. 33-66.

Marean CW, Bar-Matthews M, Bernatchez J, Fisher E, Goldberg P, Herries AIR, Jacobs Z, Jerardino A, Karkanas P, Minichillo T. 2007. Early human use of marine resources and pigment in South Africa during the middle pleistocene. Nature 449:905-908.

McBrearty S, Brooks AS. 2000. The revolution that wasn't: A new interpretation of the origin of modern human behavior. J. Hum. Evol. 39:453-563.

McDougall I, Brown FH, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733-736.

Mitchell P. 2010. Genetics and southern African prehistory: An archaeological view. J. Anthropol. Sci. 88:73-92.

Mitchell P. 2008. Developing the archaeology of marine isotope stage 3. South African Archaeological Society Goodwin Series 10:52-65.

Mitchell P, Whitelaw G. 2005. The archaeology of southernmost Africa from c. 2000 bp to the early 1800s: A review of recent research. Journal of African History 46:209-241.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK. 2006. Development and validation of the AmpFlSTR yfiler PCR amplification kit: A male specific, single amplification 17 Y-STR multiplex system. J. Forensic Sci. 51:64-75.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM et al. (15 co-authors). 2009. Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet. 5:e1000448.

Petraglia MD, Haslam M, Fuller DQ, Boivin N, Clarkson C. Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. Ann. Hum. Biol. 37:288-

311.

Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, van Helden PD, Hoal EG, Behar DM. 2010. Strong maternal Khoisan contribution to the south African coloured population: A case of gender-biased admixture. Am. J. Hum. Genet. 86:611-620.

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S et al. (23 co-authors). 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *P*roc. Natl. Acad. Sci. USA. 105:1596-1601.

Raimbault M. 1990. Pour une approche du néolithique du sahara malien. Trav.Du LAPMO :67-82.

Ravid-Amir O, Rosset S. 2010. Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. Bioinformatics. 26:i440-i445.

Renfrew C. 2010. Archaeogenetics - towards a 'new synthesis'? Curr. Biol. 20:R162-R165.

Rocha J. 2010. Bantu-Khoisan interactions at the edge of the Bantu expansions: Insights from southern Angola. J. Anthropol. Sci. 88:5-8.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science 298:2381-2385.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo

A. 2002. The making of the African mtDNA landscape. Am. J. Hum. Genet. 71:1082-1111.

Scheinfeldt LB, Soi S, Tishkoff SA. 2010. Colloquium paper: Working toward a synthesis of archaeological, linguistic, and genetic data for inferring african population history. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl 2:8931-8938.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am. J. Hum. Genet. 70:265-268.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S et al. (17 co-authors). 2000. The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. Science 290:1155-1159.

Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. 2010. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur. J. Hum. Genet. 19: 84-88.

Smith BW. 2006. Reading rock art and writing genetic history: Regionalism, ethnicity and the rock art of southern africa. In: Soodyall H, editor. The prehistory of Africa: tracing the lineage of modern man. Johannesburg and Cape Town: Jonathan Ball Publishers. p. 76-96.

Smith BW. 1997. Zambia's ancient rock art: The painting of kasama. Oxford: Nuffield Press for the National Heritage Conservation Commission of Zambia.

Smith BW. 1995. Rock art in south-central africa. Department of Archaeology, University of Cambridge. PhD Thesis.

Smith BW, Ouzman S. 2004. Taking stock: Identifying khoekhoen herder rock art in southern Africa. Curr. Anthropol. 45:499-526.

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB. 1998. Origins of old testament priests. Nature 394:138-140.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U et al. (12 co-authors). 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol. Biol. Evol. 24:2180-2195.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O et al. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. Science 324:1035-1044.

Tofanelli S, Bertoncini S, Castri L, Luiselli D, Calafell F, Donati G, Paoli G. 2009. On the origins and admixture of malagasy: New evidence from high-resolution analyses of paternal and maternal lineages. Mol. Biol. Evol. 26:2109-2124.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. 65:43-62.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 35:W71-4.

Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics 158:897-912.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. Genetics 149:1539-1546.

White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, Howell FC. 2003. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. Nature 423:742-747.

Willuweit S, Roewer L, International Forensic Y Chromosome User Group. 2007. Y chromosome haplotype reference database (YHRD): Update. Forensic. Sci. Int. Genet. 1:83-87.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H et al. (11 co-authors). 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: Evidence for sex-biased demographic processes. Eur. J. Hum. Genet. 13:867-876.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G et al. (18 co-authors). 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.

Figures legends

Figure 1. Frequencies of haplogroups A (yellow), B2a (light blue) and B2b (dark blue) in Africa. NFPR, northern Food-Producers; WFPR, western Food-Producers; WPYG, western Pygmies; CFPR, central Food-Producers; EPYG, eastern Pygmies; EKHO, eastern Khoisan speakers; EFPR, eastern Food-Producers; SKHO, southern Khoisan speakers; SFPR, southern Food-Producers. For details on specific populations included in these groups, please refer to the column "Group code" in Table S8.

Figure 2. Evolutionary relationships among A and B chromosomes. a) haplogroup A network, combined STRs and SNPs haplotypes; b) haplogroup B2b network, combined STRs and SNPs haplotypes; c) haplogroups A and B, SNP-based haplotypes. Haplotypes are coloured according to the key in the figure, and circle size is proportional to number of haplotypes, with the smallest representing n=1. STR loci used in the present analysis are: DYS19, DYS389-I, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439. The star represents the root of the Y chromosome tree as inferred from Karafet et al. 2008. For population group abbreviations refer to legend of Figure 1 and Table S8. !: back-mutation.

Tables

Table 1. Between-(a) and within-(b) lineages Time to the Most Recent Common Ancestor (TMRCA) estimates based on Average Squared Distance (ASD) and Maximum Likelihood (ML); generation time has been considered as 31 years (Helgason et al. 2003). Loci showing multi-step mutational behaviour were removed and mutation rate per locus has been estimated as in the YHRD, release 33 (Willuweit, Roewer, International Forensic Y Chromosome User Group 2007; see Table S7 for details). N, number of chromosomes included in the calculation; BP, before present; CI, confidence intervals. For the clades indicated with (*) only 7 STRs have been used for dating (see Methods section for details). For population group abbreviations refer to legend of Figure 1 and Table S8.

a. TMRCAs among lineages	**N**	**Years BP [95% CI]**
B2b2 vs B2b3 [ASD]	7 vs 7	10695 [3534-17143]
B2b2 vs B2b3 [ML]		10478 [6882-16523]
B2b2 vs B2b4b [ASD]	7 vs 6	14322 [9300-22909]
B2b2 vs B2b4b [ML]		15221 [10013-23932]
A2 SKHO vs WPYG [ASD]	5 vs 2	2883 [1891-4619]
A2 SKHO vs WPYG [ML]		3379 [2201-5363]
B2b4 SKHO vs WPYG [ASD]	3 vs 3	3627 [2356-5766]
B2b4 SKHO vs WPYG [ML]		4371 [2821-6913]
b. TMRCAs within lineages [ASD with modal]		
A1-M31*	19	10540 [4185-23684]
A1-M31 West Africa only*	12	8091 [3100-19437]
A2-South	15	6200 [2232-14198]
A3b1	22	10261 [4464-23095]
A3b2	93	9083 [3720-20274]
B2a	233	6107 [2263-14012]
B2b3	10	1984 [372-6510]
B2b4b	11	713 [31-3906]
B2b2	7	3131 [868-8990]

Table 2. Diversity indices for hg A and B, including sub-haplogroups B2a and B2b, based on 8 STRs (DYS19, DYS389I, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439). Only samples with all the 8 STRs available were included. N, number of chromosomes included in the calculation; k, number of different haplotypes; sd, standard deviation; CI, confidence interval.

Haplogroup	N	k/N	Haplotype Diversity (sd)	Variance (CI 2.5-97.5%)
A	180	0.589	0.988 (0.003)	1.099 (0.955-1.217)
B	443	0.400	0.987 (0.002)	0.562 (0.523-0.594)
B2a	233	0.373	0.965 (0.005)	0.294 (0.264-0.328)
B2b	184	0.451	0.980 (0.003)	0.743 (0.689-0.784)

Figure 1

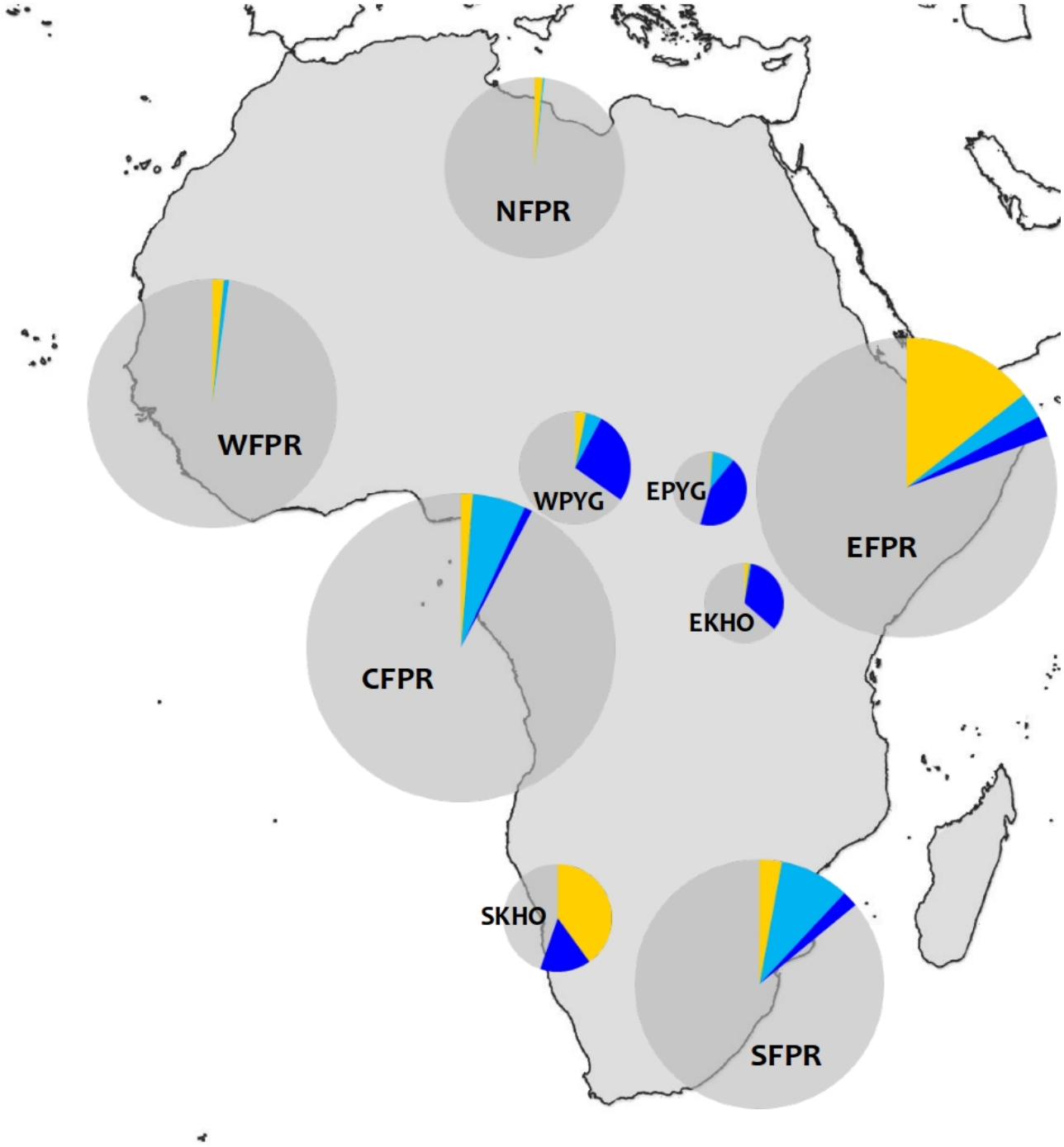
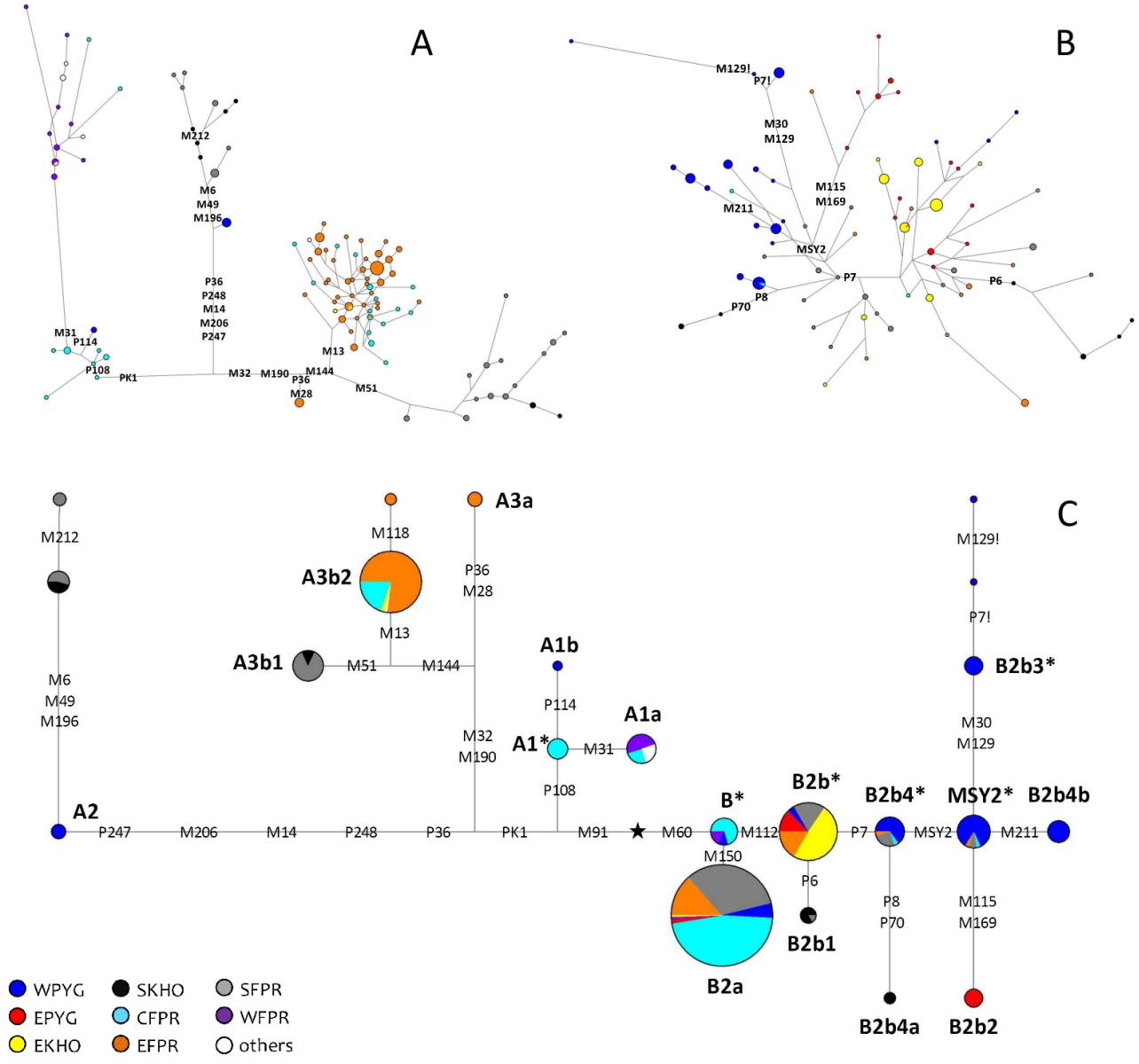


Figure 2.


A
M212
M6
M49
M196
P36
P248
M14
M206
P247
M31
P114
P108
PK1
M32
M190
M144
M13
M51
M28
B
M129!
P7!
M30
M129
M115
M169
M211
MSY2
P7
P6
P8
P70
C
M212
M118
A3a
A3b2
P36
M28
M13
A3b1
M51
M144
A1b
P114
A1a
A1*
M31
M32
M190
P108
M129!
P7!
B2b3*
M30
M129
A2
P247
M206
M14
P248
P36
PK1
M91
M60
B*
M112
B2b*
B2b4*
P7
MSY2*
MSY2
M211
B2b4b
M150
P6
P8
P70
M115
M169
B2a
B2b1
B2b4a
B2b2
WPYG
EPYG
EKHO
SKHO
CFPR
EFPR
SFPR
WFPR
others