On the paternal heritage of the Bantu expansion in Southeast Africa

Daine J. Rowold, David Perez-Benedico, Oliver Stojkovic, Ralph Garcia-Bertrand, Rene J. Herrera

PII: S0378-1119(16)30570-4
DOI: doi: 10.1016/j.gene.2016.07.044
Reference: GENE 41477

To appear in: *Gene*

Received date: 10 June 2016
Accepted date: 18 July 2016

Please cite this article as: Rowold, Daine J., Perez-Benedico, David, Stojkovic, Oliver, Garcia-Bertrand, Ralph, Herrera, Rene J., On the paternal heritage of the Bantu expansion in Southeast Africa, *Gene* (2016), doi: 10.1016/j.gene.2016.07.044

On the paternal heritage of the Bantu expansion in Southeast Africa

Daine J. Rowold[b], David Perez-Benedico[c], Oliver Stojkovic[d], Ralph Garcia-Bertrand[a]*, Rene J. Herrera[a]

[a]Department of Molecular Biology, Colorado College, Colorado Springs, CO 8090, USA

[b]Foundation for Applied Molecular Evolution, Gainesville, FL 32601, USA

[c]Biology Department, Miami Dade College, Miami, FL 33132, USA

[d]Institute of Forensic Medicine, School of Medicine, University of Belgrade, Belgrade, Serbia

*Corresponding author at: Department of Molecular Biology, Colorado College, 14 East Cache La Poudre Street, Colorado Springs, CO 80903-3294, USA. Tel.: + 1 719 389 6402; fax: + 1 719 389 6940

Keywords: South East Africans, forensic analysis, Y-chromosome, Y-STRs, population genetics, Bantu

Running Title: Phylogenetics among African Bantus.

Address for correspondence and reprints:
Dr. Ralph Garcia-Bertrand
Department of Molecular Biology
Colorado College
14 East Cache La Poudre Street
Colorado Springs, CO 80903-3294
Phone: (719) 389-6402
Fax: (719) 389-6940
E-mail: renejustoherrera@gmail.com

Abstract

Here we report the results of fine resolution Y chromosomal analyses (Y-SNP and Y-STR) of 267 Bantu-speaking males from three populations located in the southeast region of Africa. In an effort to determine the relative Y chromosomal affinities of these three genotyped populations, the findings are interpreted in the context of 74 geographically and ethnically targeted African reference populations representing four major ethno-linguistic groups (Afro-Asiatic, Niger Kordofanin, Khoisan and Pygmoid). In this investigation, we detected a general similarity in the Y chromosome lineages among the geographically dispersed Bantu-speaking populations suggesting a shared heritage and the shallow time depth of the Bantu Expansion. Also, micro-variations in the Bantu Y chromosomal composition across the continent highlight location-specific gene flow patterns with non-Bantu-speaking populations (Khoisan, Pygmy, Afro-Asiatic). Our Y chromosomal results also indicate that the three Bantu-speaking Southeast populations genotyped exhibit unique gene flow patterns involving Eurasian populations but fail to reveal a prevailing genetic affinity to East or Central African Bantu-speaking groups. In addition, the Y-SNP data underscores a longitudinal partitioning in sub-Sahara Africa of two R1b1 subgroups, R1b1-P25* (west) and R1b1a2-M269 (east). No evidence was observed linking the B2a haplogroup detected in the genotyped Southeast African Bantu-speaking populations to gene flow from contemporary Khoisan groups.

1. Introduction

The African continent, considered to be the epicenter of human evolution (Johanson and Edey, 1981; Leakey and Lewin, 1982; Leakey and Lewin, 1992; Newman, 1997), has been the stage of countless centuries of modern human occupation as well as the scientifically accepted site of Mitochondrial Eve (Cann et al., 1987) and Y chromosomal Adam (Karafet et al., 1987; Mendez et al., 2013). Further contributing to this already high degree of autochthonous diversity are the Afro-Asiatic speakers who journeyed from Eurasia at various times over the past millenniums and arrived at different African locations (Newman, 1997). Thus, the extant African populations are contemporary products of varying degrees of admixture (and isolation) among indigenous groups as well as these more recent Eurasian arrivals. Indeed, present day Africa is home to over one billion people (www.worldpopulationreview.com) who, not only inhabit a vast land mass and highly varied topography, but comprise an equally diverse cultural, linguistic and genetic landscape. In particular, the sub-Saharan region of Africa is believed to encompass most of the global human diversity (Cavalli-Sforza et al., 1994).

Interwoven into this intricate human biogeographical pattern are the myriad of demic movements that transpired, in a protracted mode, throughout the last three to four millennia, collectively referred to by historians and linguists as the "Bantu Expansion" (Newman, 1997; Cavalli-Sforza et al., 1994; Vansina, 1990; Vansina, 1995). The term "Bantu" denotes a linguistic family consisting of a group of 500 or so languages belonging to the Benue Congo branch of the Niger-Kordofanin supra-group as well as the people who originated in what is today the Nigeria-Cameroon border approximately 5,000 years ago (ya) (Cavalli-Sforza et al., 1994; Vansina, 1990; Vansina, 1995; Bleek, 1862; Greenberg, 1972). Sometime between 4,000 - 3,000 ya, Bantu speakers, who practiced an agriculturally based subsistence, dispersed from their

homeland to sustain their burgeoning population (Newman, 1997; Cavalli-Sforza et al., 1994; Vansina, 1990; Vansina, 1995). A dense system of river valleys as well as their expanding farming technology enabled successful passage through the multiple climatic zones and habitats of Central Africa (Newman, 1997; Vansina, 1995).

When viewed through the one dimensional lens of history, the Bantu Expansion appears as a massive single event culminating in the movement and resettlement of the Bantu people along two major routes: 1. South from their Cameroonian homeland through grasslands and open woodlands then turning eastward across the equatorial rain forest to the rich arable lands of East Africa and 2. A southwestern course along the coastal plains to take advantage of bountiful marine fauna, which later terminated in the dry, southwest region of the continent (Newman, 1997; Vansina, 1995). However, upon closer inspection, this diaspora was an complex phenomenon made up of multiple migratory episodes of different tempos and directions occurring in a protracted timeline (Newman, 1997; Vansina, 1995). During this time, the Bantu dispersal entailed a diffusion of the Bantu language and culture as well as genetic material to the autochthonous inhabitants of the new frontier.

In general, studies on the Bantu expansion have been geographically uneven in the populations sampled and limited in scope. Several lines of investigation, including biogeography, palynology, geology, historical linguistics and archaeology have addressed different aspects of the Bantu expansion (for a lucid recent review see Bostoen et al., 2015). In addition, recently a number of studies based on genome-wide SNPs as well as Y- and mtDNA-specific markers have focus on issues such as the distribution and history of Y chromosomal haplogroups (Bantini et al 2011), and the role of linguistic and geographic factors on Y chromosomal distribution (De Filippo et al, 2011). In addition, studies based on maternally-derive inheritance have uncovered

genetic homogeneity among East and West Africa Bantu-speaking groups (Barbieri et al, 2014) with populations in spatial expansion areas exhibiting pronounced hunter-gatherer contribution (Marks et al 2015). Other investigations have reported ancient substructure of early Khoisan mDNA lineages (Barbieri et al, 2013) and genetic differentiation among Bantu-speaking population derived from Khoisan-specific gene flow (Barbieri et al., 2014). Genome-wide data, on the other hand, suggest only limited gene between the Bantu-speaking migrants and Khoisans in Southeast Africa (Gonzalez-Santos et al., 2015). Also particularly pertinent to the present report, recent studies on Y haplogroup B2a have expressed doubts about its Bantu-specific nature (Scozzari et al., 2014; Barbieri et al., 2016).

Although the major routes and times have been established by the Bantu linguistic patterning (Cavalli-Sforza et al., 1994; Vansina, 1995; Bleek, 1862) and other forms of cultural and genetic evidence throughout sub-Saharan Africa (Cavalli-Sforza et al., 1994; Vansina, 1995; Diamond, 2003; Phillipson, 2005; Beleza et al., 2005), many questions remain unanswered. One of these concerns the genetic heritage of Bantu-speaking populations at the Southeast fringe of the expansion wave, believed to represent the most recently (ending about 300 ya) established Bantu-speaking settlements (Vansina, 1995). In particular, the relative genetic contributions of Bantu speakers from the eastern versus the western expansion routes to these southeast populations have not been established. Also unknown is the retention of the original non-Bantu gene pool and the impact of subsequent gene flow with endemic populations such as the Khoisan and the Eurasians merchants actively trading along the Southeast African coast. Thus, albeit a number of studies reporting on Southeast Bantu-speaking populations including Y-STR (Carvalho et al., 2010), mtDNA (Salas et al., 2002) and genome-wide SNPs markers (Sikora et al., 2011), the region has not been comprehensively investigated.

In an effort to alleviate the lacuna of Y-specific genetic information of Bantu populations from Southeast Africa, we have analyzed the Y chromosomal haplogroup and Y-STR haplotype distributions of three populations in the context of 74 geographically and ethically selected reference groups (Knight et al., 2003; Luis et al., 2004; Wood et al., 2005; Tishkoff et al., 2007; Berniell-Lee et al., 2009; Balamurugan and Duncan, 2012) representing the many different regions of Africa including Bantu and non-Bantu-speaking groups from throughout the continent (Figure 1). Considering the history of active trade involving Southeast Africa with Arab and Portuguese merchants, we theorized that Eurasian genetic signals resulting from admixture would be detected. In addition, we aim to provide insight into the origin of the B2a haplogroup, abundant in the genotyped Southeast African Bantu-speaking populations and theorized to derive from Khoisan or Bantu-speaking source populations.

2. Materials and methods

2.1. Sample collection and DNA extraction

A total of 267 buccal cell swab samples were collected with informed consent and IRB approval from male individuals of North Mozambique, Central Mozambique, and South Mozambique (Maputo). The paternal ancestral information was recorded for at least two generations for each donor. Supplementary Table 1 provide the population designations and number of individuals in each of the three genotyped groups. Isolation of the genomic DNA was performed as previously described (Rowold et al., 2014; Chennakrishnaiah et al., 2013).

2.2. Y-SNP and Y-STR genotyping

In order to assess Y-haplogroup diversity, 114 bi-allelic makers (Figure 2) were hierarchically genotyped using standard methods as previously reported (Luis et al., 2004;

Martinez et al., 2005; Hammer and Horai, 1995). Nomenclature of the Y-SNP haplogroups is in accordance with designations in ISOGG v10.34 based on information from several sources (Y Chromosome Consortium, 2002; Underhill et al., 2010; Myres et al., 2011). Y-STR haplotype analysis was performed using the AmpFlSTR® Yfiler™ system (Applied Biosystems, Foster City, CA) as per manufacturer's specified instructions. This multiplex system was employed to examine length polymorphisms at twelve loci: DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439. PCR products were separated by capillary electrophoresis on an ABI Prism 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) and the allelic categories of the separated Y-STR fragments were determined by the GeneMapper® v.3.2 software. In all analyses, the size of the DYS389I allele was subtracted from that of the DYS389II fragment.

2.3. Accession number

The genotypes of individuals have been successfully submitted and are now included in the YHRD database (https://yhrd.org) under accession number YA004028. All the Y-specific haplotypes reported in this article are accessible in the YHRD database.

2.4. Statistical analyses

Forty six of the 74 reference populations (Supplementary Table 1, Figure 1) along with the three study groups were utilized to determine broad phylogeographic patterns based on major Y-SNP haplogroups and sub-haplogroups. Two correspondence analysis (CA) plots based on 43 Y-SNP markers were generated as previously described (Gayden et al., 2007). The first CA included the 46 reference populations in addition to the three genotyped populations from

Mozambique for a total of 49 groups. Only 46 reference populations were utilized for the CA since the rest (28) did not provide haplogroup data of sufficient resolution. In addition, a second CA was performed on the 23 populations forming a tight cluster in the upper left quadrant of the first CA. Contour maps for haplogroups A, B2a, B2b, and E were generated using Surfer® software version 12 (Golden software Inc., Colds Spring Harbor, NY, USA, www.goldensoftware.com). Molecular variance analyses (AMOVA) were conducted using the Arlequin software v3.5 (Excoffier et al., 2003) to evaluate the geographical and linguistic partitioning of the Y Chromosome haplogroup variation. Twenty-nine populations were employed to generate pair wise genetic distance (R_{ST} values) assessments based on the ten Y-STR loci (DYS19, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439) in common among populations, excluding DYS385a/b, using the Arlquin v3.5 software package (Excoffier et al., 2003). For this analysis North and Central Mozambique are combined due to the limited number of individuals in each group. Y-STR haplotype data was not available for the remaining 47 reference populations. Pairwise population comparisons were evaluated at a significance level of 0.005 and 1,000 permutations (Kayser et al., 2003) and the Bonferroni correction was applied in order to account for potential type I errors, ($\alpha = 0.005/378 = < 0.000014$). R_{ST} distances were used to construct a multidimensional scaling (MDS) plot using the XLSTAT software from Addinsoft Corp (www.xlstat.com). Genetic relationships among haplotypes on A, B2a, B2b and E2 haplogroups were assessed by generating Median Joining (MJ) networks at the 10 loci level in accordance with previous studies (Chennakrishnaiah et al., 2013; Martinez et al., 2007). Network calculations of the Y-STR loci were inversely weighted to the size variance (Regueiro et al., 2013; Underhill et al., 2000; Cruciani et al., 2002).

3. Results

3.1. Y-SNP haplogroup diversity in Southeast Africa

The frequency distributions of the three genotyped populations (Central Mozambique, Maputo and North Mozambique) are shown in Figure 2 and Supplementary Table 2. E is the most frequent haplogroup for all three populations (78%, 93% and 60%, respectively), followed by haplogroup B (13%, 16% and 30%, respectively). B2a, a haplogroup previously associated with the Bantu dispersal (Tishkoff et al., 2007; Berniell-Lee et al., 2009) , is detected in 9%, 16% and 10% of Central Mozambique, Maputo and North Mozambique individuals, respectively, whereas B2b, a haplogroup traditionally correlated to Pygmy populations (Tishkoff et al., 2007; Berniell-Lee et al., 2009) is represented by 0.4%, 4% and 20% of these three populations. For Central Mozambique, R constitutes the remaining haplogroup (7%) and for the North Mozambique, K (10%) completes the distribution. In contrast, the Maputo population is more diverse as R (3%), A (1%), K (1%), T (1%) and J (0.4%) are observed. However, the increase in the detected haplogroup diversity may due, at least in part, to much larger sample size of the Maputo (N=234) versus those of Central (N=23) and North Mozambique (N=10).

The partitioning of the E sub-haplogroups for the three genotyped populations (Figure 2) indicates that in the Central Mozambique, Maputo and North Mozambique distributions, the majority of the E haplogroup members contain the M180 polymorphism, a marker traditionally associated with the Bantu dispersal (74%, 66% and 50% of total samples, respectively). M35 is observed at low frequencies (4% and 1% in Central Mozambique and Maputo, respectively) but is not detected in North Mozambique.

3.2. Inter-population diversity and population relationships

Four Y-SNP haplogroups, A, B2b, B2a and E, are represented by frequency contour maps of Africa (Figures 3, 4, 5 and Supplementary Figure 1, respectively). The contour map of A haplogroup (Figure 3), which includes the root of the Y chromosomal phylogeny[6], reveals a frequency pattern with a relatively high concentration in the southwest corner of the continent reflecting the presence of the four Khoisan populations in that region: Nama (64% A), Tsumkwe San (64% A), !Kung (48% A), and Dama (12% A). In addition, an East African area is characterized by frequencies reaching as high as 28% in the Maasai (Nilotic). The second most ancient haplogroup B2b (Figure 4) is conventionally associated with the Pygmy people and exhibits the highest density in the central region of Africa (up to 67% in Baka 2). The next highest concentration of B2b is in East Africa (42% in Mbuti, a Pygmy-speaking people) followed by a frequency of 33% in a southwest Khoisan group (Tsumkwe San). Moreover, it is moderately abundant (20%) in the North Mozambique Bantu population of Southeast Africa as well. The highest contour density of the B2a marker (Figure 5), previously considered a signature of the Bantu Expansion by the orthodoxy (Underhill et al., 2000; Cruciani et al., 2002), is observed in the geographic region encompassing the western African homeland of Cameroon (33% in Ngumba 1). It is also detected at relatively high frequency in the Bantu-speaking populations of Southeast Africa (up to 16% in Mozambique) and South Africa (up to 18% in Sotho-Tswana). However, B2a is also present at polymorphic levels (≥5%) in Pygmies as well as in several Afro-Asiatic populations (Chadic, Cushitic, Gur, Nilotic) in various regions of Africa and the second highest (31%) concentration is in the Chadic-speaking Uldeme of Central Africa. In contrast, B2a appears to be absent in Southwest Africa and nearly so in West Africa. The contour map of the E (Supplementary Figure 1) reveals high levels (100%) of this haplogroup in

some West, Central and East African populations representing several Niger-Congo language family members (Fon, Kwa and Bantu). Specifically, two Bantu-speaking populations (out of a total of 42) are 100% haplogroup E, Bakaka in Central Africa and Nande in East Africa. Three other populations with 100% haplogroup E are Fon, Ewe and Ga of West Africa. This haplogroup is absent in only two of the Central African groups (Uldeme and the Gur).

A CA, generated from haplogroup frequency distribution of 49 populations (the three genotyped in this study and 46 reference populations) representing all five regions and all language branches (Supplementary Table 1) is presented in Supplementary Figure 2a. This CA plot features three diffuse major assemblies. The upper left quadrant contains all of the Khoisan, Pygmy and Bantu-speaking groups as well as some of the Atlantic (Mandinka, Wolof), Dogon (Dogon), Kwa (Ewe, Fant and Ga), Fon (Fon) and Nilotic (Luo) -speaking groups. In this conglomerate, Pygmy and Khoisan populations form loose upper and lower aggregates around a compact nucleus of 23 populations, the majority of which are Bantu speakers. All study regions of sub-Saharan Africa (west, central, east and south) are represented in this tight upper left quadrant. The upper right quadrant exhibits a loose assembly of two Nilo-Saharan (Alur and Maasai) and seven Afro-Asiatic populations representing Cushitic (Oromo, S. Cushitic), Semitic (Amhara, Tunisian, South Semitic and Egyptian Arabs) and Erythraic (Egyptian) dialects of East Africa and/or Egypt. The lower left quadrant includes four Central African groups, three of which are Chadic and the fourth, Gur speakers. It is noteworthy that these four populations are relatively close on the geographical map (Figure 1) as well.

The second CA plot (Supplementary Figure 2b) is a new analysis utilizing only the populations located within the crowded cluster in the upper left quadrant. This was done to visualize the individual populations within the compact cluster and allow assessment of their

partitioning. Most (15 or 65%) of the 23 groups in this second CA are Bantu speakers. Members of West African language families (Atlantic, Dogon, Fon and Kwa) as well as a lone Central African Pygmy population (Bakola 2) constitute the remaining eight populations. Within this predominantly Bantu-speaking cluster, little geographic substructuring is evident.

The multidimensional scaling (MDS) plot (Kruskal stress index = 0.098) (Supplementary Figure 3) was generated from pair-wise R_{ST} distances (Supplementary Table 3) based on Y-STR frequency data (Supplementary Table 4) of the genotyped groups (Maputo, Central and North Mozambique) and 27 reference African populations listed in Supplementary Table 1. A diffuse assembly in the upper and lower right quadrants contains the Bantu-speaking groups. The Mozambique populations are proximal to Central (Eshira and Ndumu) as well as to East (Sukuma and Turu) African Bantu-speaking populations. Peripheral to this predominantly Bantu cluster are Nilotic (Datog) and Khoisan (Sandawe) groups from East Africa, while Pygmy (Bakola 1 and Baka 1, Central Africa), Khoisan (Hadza, East Africa) and Cushitic (Burunge, East Africa) groups are more distantly scattered across the upper and lower left quadrants.

Seven populations comprise the network for haplogroup A (Supplementary Figure 4). Four Central African Bantu speakers (Duma, Ndumu, Nzebi and Tsogo of Gabon), one Central African Pygmy (Baka of Gabon), one East African Bantu speaker (Rwanda) and the Maputo of Southeast Africa form a phylogeny with several short branches (≤ 3 mutational steps) emanating from an invisible node and terminating in singletons representing Gabon individuals. A fourth limb splits further into two long boughs, one bearing Y-STR profiles of lone males from Duma and Rwanda, while the other subdivides again leading to either a branch with two Maputo singletons or a lone Baka node. The only haplotype sharing observed is within the Baka population.

The B2a network (Figure 6) was generated from Y-STR profiles of 20 Bantu-speaking populations, 17 Central African and the three Southeast African groups (Central Mozambique, North Mozambique and Maputo). The projection exhibits a configuration with six edges radiating from a central node shared by three groups from Gabon (Benga, Eshira and Eviya). Three short spokes terminating in either Ngumba or Benga singletons are one mutational step from the center. The remaining limbs are more complex, displaying multigroup nodes, several secondary radiations and longer edge lengths. The Southeast Bantu haplotypes are found in terminal clusters on two of these major branches. One of these assemblies, containing only members of the Central Mozambique, North Mozambique and Maputo populations, is approximately 18 mutational steps from the haplotype at the central node. The second southeast cluster is much closer (about six mutational steps from the ancestral haplotype) and includes a branch featuring four groups from Gabon (Duma, Nzebi, Punu and Shake).

The network analysis of B2b Y-STR haplotypes, shown in Supplementary Figure 5, connects eight populations: two Pygmy (Baka and Bakola of Central Africa), four Central Africa Bantu speakers and two of the three genotyped Southeast African groups (Maputo and North Mozambique) in non-star-like pattern. Multiple Baka B2b Y-STR profiles are dispersed throughout the network and six of the other populations exhibit at least one haplotype that is three or less mutational steps from a Baka node. The one exception is the Shake (Bantu speaker) of Central Africa. Haplotype sharing occurs only within the Pygmy populations.

The E2 network (Supplementary Figure 6) exhibits a non-star-shaped configuration consisting of males from 17 Central African Bantu-speaking populations and Maputo. All but one of the six Maputo individuals occupy multi-branched side limbs about twelve mutational steps from the largest multi population node. This cluster encompasses Bantu speakers from

eight Central African groups (Duma, Eviya, Eshira, Galoa, Kota, Makina, Ngumba, Obamba). However, there is very little inter-population sharing of haplotypes in this region of the network.

The results of the AMOVA analysis presented in Supplementary Table 5 indicate that that Y-STR haplotype variation is significant on all three levels (within populations, among populations within groups and among groups) for both linguistic and geographic partitioning. However, the variation among linguistic groups is substantially greater and more highly significant than that based on geographic partitioning (27%, $p < 0.00001$ versus 4%, $p < 0.01$, respectively).

4. **Discussion**

4.1. The Bantu

Due to the protracted series of human dispersals known as the Bantu Expansion (Newman, 1997; Cavalli-Sforza et al., 1994; Vansina, 1990; Vansina, 1995), the Bantu exhibits the widest geographical range of all of the major ethno-linguistic groups represented in this study (Khoisan, Pygmy, Afro-Asiatic, Bantu, and various other Niger Kordofanian speakers as designated in Supplementary Table 1). This contention is supported by the geographical distribution of Y haplogroups broadly and traditionally associated with three major ethno-linguistic groups (A: Khoisan, B2b: Pygmies, and B2a and E: Niger-Kordofanian) as depicted in the contour maps (Figures 3, 4, 6 and Supplementary Figure 1). Haplogroup E, which includes Bantu speakers as well as other Niger Kordofanian members, covers all major geographical regions of the continent, while B2a appears to trace the overall path of the Bantu demic dispersion (Figure 5). Also notable, is that although Bantu speakers retained their ancestral Y chromosomal heritage, as indicated by the tight clustering of Bantu speaking populations seen in the CA plot of Supplementary Figure 2a and genome wide data (Brye et al., 2010), their genetic

composition may have been impacted at different locations by other major ethno-linguistic groups such as Khoisans and Pygmies as well as various Eurasian and Afro-Asiatic hoplgroups, as evidenced by the polymorphic levels haplogroups R1b1-P25 and E1b1b1-M35, respectively. Also, the proximity of Khoisan and Pygmies to the Bantu cluster in the CA plot could reflect introgression of Bantu haplogroups in these hunter-gatherers. Thus, Y chromosomal transfer appears to have been bidirectional to varying degrees. E1b1a-M180, traditionally regarded as a Bantu-specific signature, ranges from 54% to 12% in the Pygmy populations and 22% to 9% in three of four South African Khoisan groups (absent in the Tsummke-San). In contrast, the presence of conventional Bantu signals is scarce in the majority of East African Afro-Asiatic populations (E1b1a-M180: ≤ 1% in S. Cushitic and 0 in the remaining eight). Furthermore, the Chadic and Gur also display low frequencies of E1b1a-M180 (4% in Mandara and 0 in the Tupuri, Podokwo and Uldeme). Although B2a-M150 occurs at a considerable frequency (31%) in the Uldeme, it is low (either 4% or 0) in the other three Central African Afro-Asiatic groups.

4.2. Genetic affinity of the Southeast African Bantus

According to the MDS (Supplementary Figure 3) and the CA (Supplementary Figure 2) plots, the Southeast African Bantu-speaking populations do not exhibit higher affinity to the East compared to the Central African Bantu-speaking populations. The MDS, in particular, exhibits a pattern in which both East and West African Bantu populations are in close proximity to the three Southeast African Bantu-speaking groups. In the CA projection, the three Mozambiquean populations are located within the tight nucleus of mostly Bantu, which includes geographically dispersed populations. Furthermore, Maputo and Central Mozambique associate with east

(Tutsi), central (Ngumba 2), southwest (Ambo) and southeast (Shona) Bantu-speaking groups (Supplementary Figure 2b).

4.3. Origins of haplogroup B2a in Southeast African Bantu speakers

Although B2a-M150 has been dated at about 6.0 kya based on Y-STR diversity (Batini, 2011), more recent studies based on Khoisan groups from Southwest Africa (Botswana, Namibia, and Zambia) have generated much older time estimates in the order of 40 kya (Scozzari et al., 2014; Barbieri et al., 2016). These dates argue for B2a predating the Bantu Expansion and raise the possibility that at least some of the B2a lineages are not Babtu in origin. In connection with this possibility, in the three Southeast African populations genotyped in the present study, the B2a haplogroup was detected at a frequency of 15% (40/267). Furthermore, network analysis reveals an interesting bipartite partitioning of the Central Mozambique, Maputo and North Mozambique B2a Y-STR haplotypes into two distinct clusters (Figure 6). One of these forms a branch shared by members of four Central African Bantu-speaking populations (Duma, Nzebi, Punu and Shake) and is separated by approximately six mutations from the core haplotype. A second, more distant assemblage (approximately 18 mutations away from central node) is composed solely of Maputo, North and Central Mozambique males. These two distinct lineages may represent distinct migration waves from unique ancestral populations with unique Y-STR diversities, for example, haplotypes of east *versus* central Bantu sources. Alternatively, since some of the B2a lineages may represent Khoisan ancestry (Scozzari et al., 2014: Barbieri et al., 2016), this partitioning into two clusters may represent one group of B2a being Khoisan in origin (possibly the distant upper one) while the other, consisting of a number of several Bantu-speaking populations, may correspond to more recent gene flow from the advancing Bantu dispersal. However, although B2a was found at a frequency of 9.2% (n=53) among West

African Khoisan groups (Barbieri et al., 2016), the presence of only one B2a individual in the eight Khoisan populations (from a total of 306 individuals or 0.3 %) from East and Southwest Africa (Tishkoff et al.2007) is not congruent with contemporary Khoisans as a source for the B2a detected among the Southeast African Bantu-speaking genotyped in the present study. This dichotomy in frequency of B2a among Khoisan groups may stem from regional differences. Thus, based on the larger Khoisan database that includes four East African populations (all 53 B2a individuals reported in Barbieri et al 2016 belong to West African groups), it is unlikely that extant Khoisans are responsible for the bipartite partitioning we observed in the B2a network. Yet, considering that B2a has been assessed to be very ancient, it is still possible that extinct Khoisan groups are responsible for at least some of the B2a seen in the Southeast African Bantu-speaking populations. Also, since the arrival of Bantu-speakers to Southeast Africa is relatively recent, it is possible that B2a was incorporated from hunter-gatherers groups into Bantu migrants en route to Southeast Africa. Alternatively, differential founder's effect and/or random genetic drift may be responsible for this dichotomy. Further, considering the substantial STR differences and the number of mutational steps differential between the two lineages from the core haplotype as well as the almost absence of B2a in contemporary Khoisans in Southeast Africa, the two independent Bantu migrations into Southeast Africa is a more likely scenario. In view of the rapid pace of the Bantu Expansion and the recent occupation of Southeast Africa by Bantu speakers, the proposed two-migrations model is the most parsimonious explanation. At this juncture, more Y-STR data from geographically/ethnically-targeted populations, especially from East African Bantu-speaking populations, is needed to resolve the genetic relationships of these two factions.

4.4. Eurasian admixture

The Eurasian ancestry of R1b1-P25 in the three Southeast African Bantu-speaking populations genotyped in the present study is supported by the detection of other Eurasiatic haplogroups such as R1a, K*, J2b, T and possibly E-M34. The presence of these foreign haplogroups in Southeast Africa supports gene flow from non-African origins, possibly of West European and Southwest Asian sources. For example, R-M269 exhibits high frequencies in Western Europe and represents about half of the Portuguese Y chromosomes. Thus, R-M269 may represent admixture involving Portuguese merchants or slave traders. Yet, the absence of haplogroups G, I and most of J, which are common in Portugal, points to a strong founder effect. Other haplogroups such as R1a, K*, J2b may derive from extraneous genetic inputs from Central Europe, the Balkans and Southwest Asia. Arab trade along the Africa east coast may have contributed to these signals.

4.5. Two sources for R1b1-P25

Especially interesting is the observation that although there appears to be an asymmetrical longitudinal partitioning of E1b1b1-M35 and R1b1-P25 in the eastern versus western regions of the continent, respectively, as reported in previous investigations (Luis et al., 2004; Rowold et al., 2014), both of these polymorphisms are present in the Maputo and Central Mozambique collections. The absence of E1b1b1-M35 and R1b1-P25 in North Mozambique may be due to detection failure considering the small sample size (N =10). The presence of the R1b1-P25 marker in Maputo and Central Mozambique may be a result of gene flow accompanying trade with the Arabs along this region of the southeast African coast since the 15th century (Newman, 1997; Garlake, 1973). This is supported by the higher level of R1b1-P25

in Central Mozambique (9%), which is geographically closer to the early trade centers of the Great Zimbabwe and Port of Sofala in the early to middle centuries of the second millennium, AD. It appears that the westward genetic diffusion of the Eurasian R1b1-P25 tapers off to sub-polymorphic levels in the more inland Southeast Bantu groups (Shona, Sotho-Tswana, Xhosa and Zulu). Perhaps, in relay fashion, the Shona of the Great Zimbabwe exchanged goods with the Bantu settlements closer to the coast, which in turn transacted with the Eurasian sailors. In addition to the Arab genetic contribution, Portuguese traders and settlers since the sixteen century AD may have introduced R1b1-P25 chromosomes into the two Mozambique populations[4] as well. These population sources are highly consistent with the presence of the M269 mutation (defining the R1b1a2 haplogroup) associated with the Neolithic spread of agriculture from the Near East westward and common throughout West and Southwest Europe (Underhill et al., 2000; Cruciani et al. 2002). Although the West and Central African populations exhibit extremely high frequencies (61% to 95%) of R1b1-P25, they harbor no M269. These individuals are just non-derived R1b1-P25*. In contrast, two of the three Southeast African Bantu populations genotyped in this study, displaying much lower frequencies of R1b1-P25 (≤ 9%) overall, contain polymorphic levels of M269 (Central Mozambique 9% and Maputo 1%). This asymmetrical distribution pattern indicates that the absence of the M269 mutation in the Central Africa is not due to sampling limitations but may suggest different sources for the R1b1-P25 and R1b1a2-M269 polymorphisms. In other words, R1b1-P25* in West and Central African populations may originate from West Asia gene flow while R1b1a2-M269 found in Southeast Africa may be of West European ancestry introduced via trade and/or during colonial times.

5. Conclusion

Our Y chromosomal investigation of 267 males from three Southeast African populations (Central Mozambique, Maputo, and North Mozambique) and 74 reference populations indicate that: 1. The high degree of Y chromosomal homogeny apparent in the compact Bantu cluster of the CA plot and MDS is likely a consequence of several factors including a common origin and the rapid pace of the expansive diaspora known as the Bantu Expansion; 2. In spite of this overall genetic similarity, geographic micro-variation in the Bantu Y chromosomal gene pool suggests that Bantus have been both a recipient as well as a contributing source to other ethno-linguistic groups featured in this investigation. The location-specific patterns of genetic diversity are likely the result of multiple episodes and routes of the Bantu migrations as well as the specific and unique interactions with local autochthonous populations; 3. Several analytical results (Y-STR: MDS and B2a network as well as Y-SNP: CA plots) fail to support a predominant Y chromosomal affinity of the three Southeast African Bantu groups (Maputo, Central or North Mozambique) to East *versus* Central African Bantu populations; 4. A bipartite network distribution based on Y-STR haplotypes under haplogroup B2a cannot be attributed to some lineages being Khoisan in origin; 5. The East-West African asymmetrical partitioning of E1b1b1-M35 (east) and R1b1-P25 reported in previous publications is apparent in this expanded data set as well. However, in the present study, we also uncover a longitudinal division of two R1b1-P25 subgroups. These are the R1b1a2-M269 subhaplogroup in two of the Mozambique genotyped populations, which are the only East African Bantu (and sub-Saharan) populations reported to exhibit it, and R1b1-P25* (minus the M269 mutation) in Central African Chadic and Bantus. The proximity of the study groups to the location of the Great Zimbabwe Empire and maritime commerce center suggests that R1b1a2-M269 may have been introduced by admixture with Arab and/or Portuguese traders, most likely the later.

Conflict of interest

No conflict of interest exists.

References

Balamurugan, K., Duncan, G., 2012. Y chromosome STR allelic and haplotype diversity in a Rwanda population from East Central Africa. Leg Med, Tokyo 14, 105-109.

Barbieri, C., Güldemann, T., Naumann, C., et al., 2014. Unraveling the complex maternal history of Southern African Khoisan populations. American Journal of Physical Anthropology. 153, 435–448.

Barbieri, C., Hübner, A., Macholdt, E., et al., 2016. Refining the Y chromosome phylogeny with southern African sequences. Human Genetics 135, 541-553.

Barbieri, C., Vicente, M., Oliveira, S., et al., 2014. Migration and Interaction in a Contact Zone: mtDNA Variation among Bantu-Speakers in Southern Africa. PLoS ONE 9(6):e9911.

Barbieri, C., Vicente, M., Rocha, J., et al., 2013. Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet. 92, 285-92.

Batini, C., Gianmarco, F., Destro-Bisol, G., et al., 2011. Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Molecular Biology and Evolution 28, 2603-2613.

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

Berniell-Lee, G., Calafell, F., Bosch, E., *et al.*, 2009. Genetic and demographic implications of the Bantu Expansion: Insights from human paternal lineages. Mol. Biol. Evol. 26, 1581-1589.

Bleek, W., 1862. A Comparative Grammar of South African Languages. Trübner & Co., London, UK.

Bostoen, K., Clist, B., Doumenge C., et al., 2015. Middle to Late Holocene Paleoclimatic Change and the Early Bantu Expansion in the Rain Forests of Western Central Africa. Current Anthropology 56, 354 – 384.

Bryc, K., Auton, A., Nelson, M., *et al.*, 2010. Genetics Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. US A 107, 786–791.

Cann, R., Stoneking, M., Wilson, A., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, NJ.

Chennakrishnaiah, S., Perez, D., Gayden, T., Rivera, L., Regueiro, M., Herrera, R. J., 2013. Indigenous and foreign Y chromosomes characterize the Lingayat and Vokkaliga populations of southwest India. Gene 526, 96-106.

Cruciani, F., Santolamazza, P., Shen, P., *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.

De Filippo, C., Barbieri, C., Whitten, M., et al., 2011. Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol. 28,1255-1269.

Diamond, J., Bellwood, P., 2003. Farmers and their languages: The first expansions. Science 300, 597-603.

Excoffier, L., Lischer, H., 2003. Arlequin suite v 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10, 564-567.

Garlake, P., 1973. Great Zimbabwe. Stein and Day, New York NY.

Gayden, T., Cadenas, A., Regueiro, M., *et al.*, 2007. The Himalayas as a directional barrier to gene flow. Am. J. Hum. Genet. 80, 884–894.

González-Santos, M., Francesco, M., Oosthuizen , O., et al ., 2015. Genome-wide SNP analysis of Southern African populations provides new insights into the dispersal of Bantu speaking groups. Genome Biol Evol. 9, 2560–2568.

Greenberg, J., 1972. Linguistic evidence regarding Bantu origins. Journal of African History 13, 189-216.

Hammer, M., Horai, S., 1995. Y chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. 56, 951–962.

Johanson, D., Edey, M., 1981. Lucy. Simon and Schuster, New York NY.

Karafet, T., Mendez, F., Meilerman, M., Underhill, P., Zegura, S., Hammer, M., 1987. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research 18, 830–838.

Kayser, M., Brauer, S., Stoneking, M., 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. 20, 893-900.

Knight, A., Underhill, P., Mortensen, H., et al., 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. J. Current Biology 13, 464–473.

Leakey, R., Lewin, R., 1982. Origins. Lodestar Books, New York, NY.

Leakey, R., Lewin, R., 1992. Origins Reconsidered. Doubleday, New York, NY.

Luis, J., Rowold, D., Regueiro, M., et al., 2004. The Levant versus the Horn of Africa: Evidence for bidirectional corridors of human migrations. Am. J. Hum. Genet .74, 532–544.

Marks SJ., Montinaro, F., Levy, H., et al., 2015. Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. Mol Biol Evol. 32, 29-43.

Martinez, L., Reategui, E., Fonseca, L., et al.: Superimposing polymorphism: The case of a point mutation within a polymorphic *Alu* insertion. Hum. Hered. 59, 109–117.

Martinez, L., Underhill, P., Zhivotovsky, L., et al., 2007.Paleolithic Y haplogroup heritage predominates in a Cretan highland plateau. Eur. J. Hum. Genet. 15, 485-493.

Mendez, F., Krahn, T., Schrack, B., 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.

Myres, N., Rootsi, S., Lin, A., 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.

Newman, J., 1997. The Peopling of Africa: A Geographic Interpretation. Yale University Press, New Haven CT.

Phillipson, D., 2005. African Archaeology. Cambridge University Press, Cambridge, UK.

Regueiro, M., Alvarez, J., Rowold, D., Herrera, R. J., 2013. On the origins, rapid expansion and genetic diversity of Native Americans from hunting-gatherers to agriculturalists. Am. J. Phy. Anthro. 150, 333-348.

Rowold, D., Garcia-Bertrand, R., Calderon, S., Perez-Benedico, D., Varela, M., Herrera, R. J., 2014. At the southeast fringe of the Bantu expansion: Genetic diversity and phylogenetic relationships to other sub-Saharan tribes. Meta Gene 2, 670–685.

Scozzari, R., Massaia, A., Beniamino, T., et al., 2014. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res. 24, 535-544.

Tishkoff, S., Gonder, M., Henn, B., et al., 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol. Biol. Evol. 24, 2180-2195.

Underhill, P., Myres, N., Rootsi, S., et al., 2010. Separating the post-glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur. J. Hum. Genet. 18, 479-484.

Underhill, P., Shen, P., Lin, A., et al., 2000. Y chromosome sequence variation and the history of human populations. Nat. Genet. 26, 358–361.

Vansina, J., 1995. New linguistic evidence and the Bantu Expansion. Journal of African History 36, 173–195.

Vansina, J., 1990. Paths in the Rainforests. University of Wisconsin Press, Madison WI.

Wood, E., Stover, D., Ehret, C., 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.

Y Chromosome Consortium, 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 12, 339–348.

C

Figure titles and legends

Figure 1: Population map.

Figure 1. Collection locations of study and reference populations.

Figure 2: YSNP haplogroup frequencies.

Figure 2. Hierarchy of YSNP haplogroups and frequency distribution of Central Mozambique, Maputo and North Mozambique populations.

Figure 3: Contour map, haplogroup A.

Figure 3. Frequency contour map of Y haplogroup A in Africa.

Figure 4: Contour map, haplogroup B2b.

Figure 4. Frequency contour map of Y haplogroup B2b in Africa.

Figure 5: Contour map, haplogroup B2a

Figure 5. Frequency contour map of Y haplogroup B2a in Africa.

Figure 6: Network, haplogroup B2a.

Figure 6. Network phylogeny of Y chromosomal haplogroup B2a.

Supplementary Files:

Supplementary Figure 1: Contour map, haplogroup E.

Supplementary Figure 1. Frequency contour map of Y haplogroup E in Africa.

Supplementary Figure 2: CA plots.

Supplementary Figure 2. CA based on YSNP haplogroup frequency data. **Panel a**: Forty nine populations are featured from Supplementary Table 1. A total of 26.4% of YSNP variation is represented in x (13.6%) and y (12.8%) axes. **Panel b**: CA of the "Bantu" cluster in the upper left quadrant of panel a containing 23 populations. A total of 33.9% of YSNP variation is represented in x (20.4%) and y (13.5%) axes.

Supplementary Figure 3: MDS.

Supplementary Figure 3. MDS based on YSTR haplotypes. ● Bantu, Central Africa,○ Bantu, East Africa, ▲Bantu, Southeast Africa, ■ Cushitic, □ Datog, Δ Khosian and♦Pygmy. Population codes are as indicated in Supplementary Table 1. YSTR data from Central and North Mozambique are combined (MOZ) to increase sample size.

Supplementary Figure 4: Network, haplogroup A.

Supplementary Figure 4. Network phylogeny of Y chromosomal haplogroup A.

Supplementary Figure 5: Network, haplogroup B2b.

Supplementary Figure 5. Network phylogeny of Y chromosomal haplogroup B2b.

Supplementary Figure 6: Network, haplogroup E2.

Supplementary Figure 6. Network phylogeny of Y chromosomal haplogroup E2.

Supplementary Table 1: List of populations.

Supplementary Table 1. List of populations examined.

Supplementary Table 2: Y-SNP frequencies in Southeast Africa.

Supplementary Table 2. Y-SNP frequencies of Southeast Africa populations.

Supplementary Table 3: Rst and corresponding p values.

Supplementary Table 3. Rst and corresponding p values.

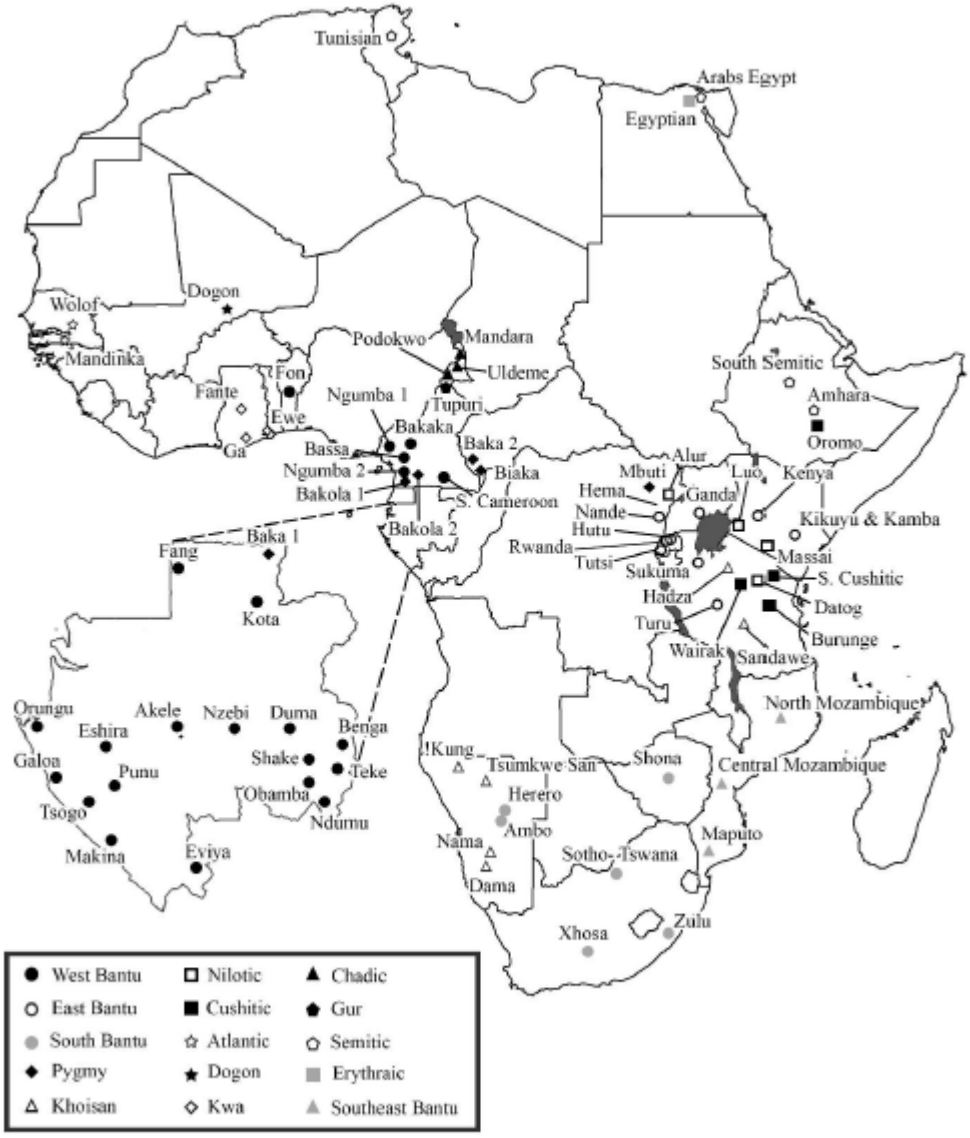
Supplementary Table 4: Y-STR haplotypes.

Supplementary Table 4. Y-STR haplotypes.

Supplementary Table 5: AMOVA

Supplementary Table 5. Analysis of molecular variance (AMOVA) using YSNP haplogroups.

Figure 1. Geographical locations of populations examined



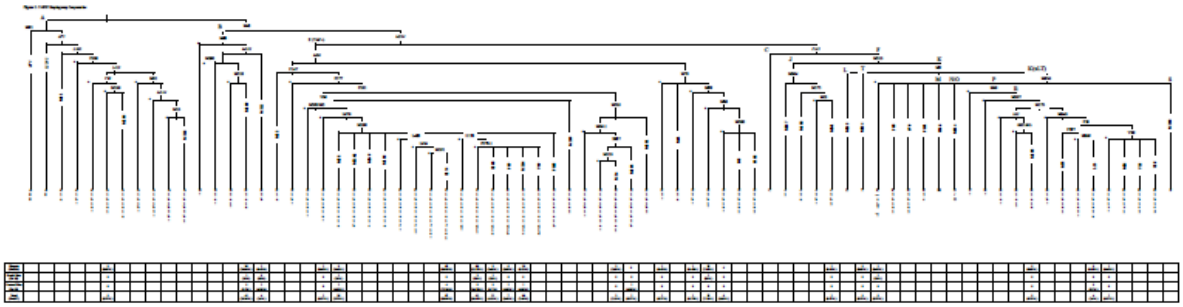


Figure 2

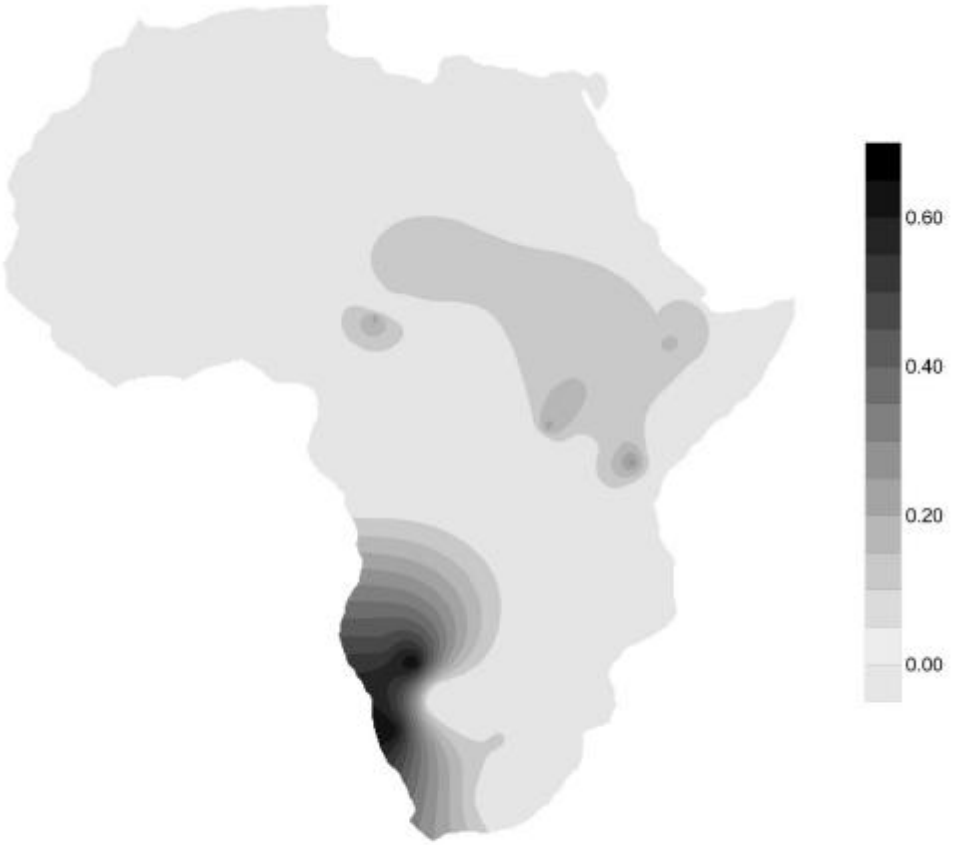


Figure 3. Frequency contour map of Y haplogroup A in Africa

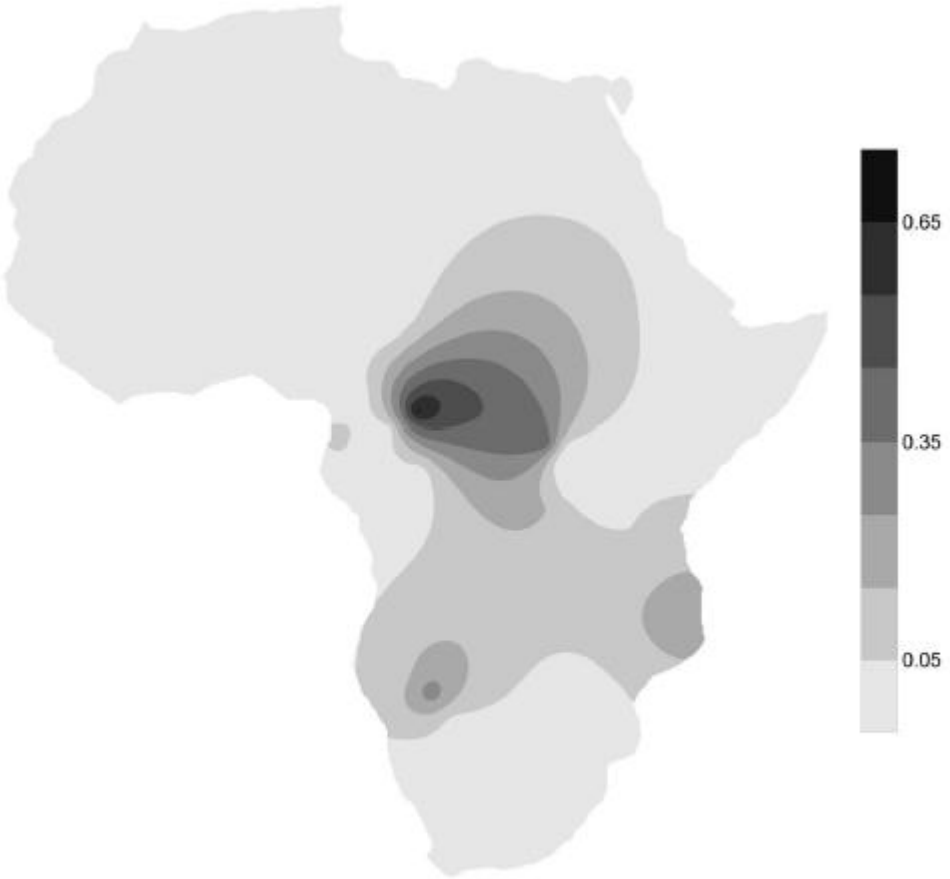


Figure 4. Frequency contour map of Y haplogroup B2b in Africa

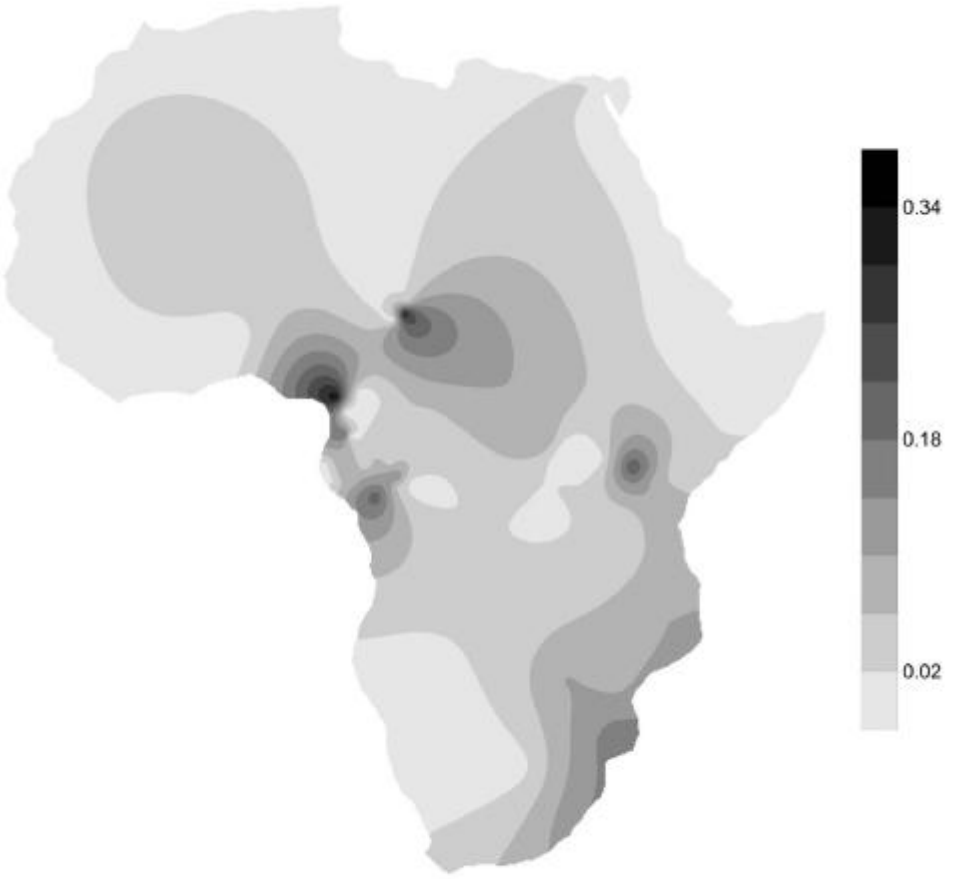
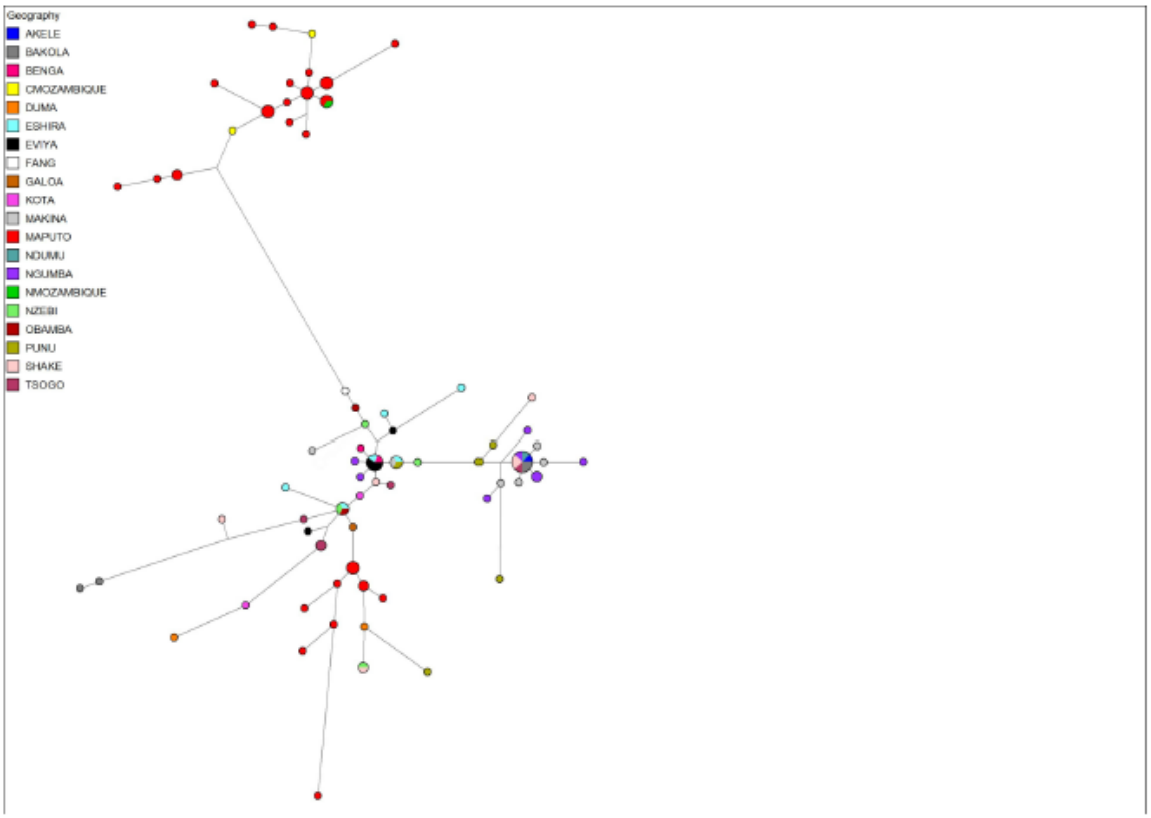


Figure 5. Frequency contour map of Y haplogroup B2a in Africa

Figure 6 Network phylogeny of haplogroup B2a



Abbreviation list:
SNP Single Nucleotide Polymorphism
STR: Short Tandem Repeats
mtDNA: Mitochondrial DNA
PCR: Polymerase Chain Reaction
AMOVA: Molecular Variance Analysis
MJ: Median Joining
MDS: Multi-Dimensional Scaling
CA: Component Analysis