Gene 593 (2016) 48–57

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

On the Bantu expansion

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 80903, 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

ARTICLE INFO

Article history:
Received 10 June 2016
Accepted 18 July 2016
Available online 19 July 2016

Keywords:
South East Africans
Forensic analysis
Y-chromosome
Y-STRs
Population genetics
Bantu

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 sustained 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 5000 years ago (ya) (Cavalli-Sforza et al., 1994; Vansina, 1990; Vansina, 1995; Bleek, 1862; Greenberg, 1972). Sometime between 4000 and 3000 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

Abbreviation: 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.

* Corresponding author at: Department of Molecular Biology, Colorado College, 14 East Cache La Poudre Street, Colorado Springs, CO 80903-3294, USA.
E-mail address: renejustoherrera@gmail.com (R. Garcia-Bertrand).

http://dx.doi.org/10.1016/j.gene.2016.07.044
0378-1119/

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 a 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 (Batini et al., 2011), and the role of linguistic and geographic factors on Y chromosomal distribution (De Filippo et al., 2011). Also, studies based on maternally-derive inheritance have uncovered genetic homogeneity among East and West Africa Bantu-speaking groups (Barbieri et al., 2014a, 2014b) 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., 2014a, 2014b). Genome-wide data, on the other hand, suggests only limited gene flow between the Bantu-speaking migrants and Khoisans in Southeast Africa (González-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 and Bellwood, 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, mtDNA and genome-wide SNPs markers, 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 ethnically 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 (Fig. 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 provides 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 (Fig. 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, Fig. 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 the 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 and Lischer, 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 and Lischer, 2003). For this analysis, North and Central Mozambique are combined due to