Gene: X 5 (2020) 100026
https://www.academia.edu/97223753/The_Y_chromosome_of_the_Soliga_an_ancient_forest_dwelling_tribe_of_South_India

Contents lists available at ScienceDirect

Gene: X

journal homepage: www.journals.elsevier.com/gene-x

Research paper

The Y-chromosome of the Soliga, an ancient forest-dwelling tribe of South India

Diane J. Rowold[b,g], Shilpa Chennakrishnaiah[c], Tenzin Gayden[d], Javier Rodriguez Luis[e], Miguel A. Alfonso-Sanchez[f], Areej Bukhari[f], Ralph Garcia-Bertrand[a,*,1], Rene J. Herrera[a]

[a] Department of Molecular Biology, Colorado College, Colorado Springs, CO 80903, USA
[b] Foundation for Applied Molecular Science (FfAME), Gainesville, FL 32601, USA
[c] Department of Experimental Medicine, McGill University, Montreal, Quebec, Canada
[d] Department of Human Genetics, McGill University, Montreal, Canada
[e] Area de Antropología, Universidad de Santiago de Compostela, Spain
[f] Departamento de Genética y Antropología Física, Facultad de Ciencia y Tecnología, Universidad del País Vasco (UPV/EHU), Bilbao, Spain
[g] Department of Pediatrics, Nicklaus Children's Hospital, Miami, FL, USA

ARTICLE INFO

Keywords:
Soliga
STR
SNP
Sub-Sahara Africa

ABSTRACT

A previous autosomal STR study provided evidence of a connection between the ancient Soliga tribe at the southern tip of the Indian subcontinent and Australian aboriginal populations, possibly reflecting an eastbound coastal migration circa (15 Kya). The Soliga are considered to be among India's earliest inhabitants. In this investigation, we focus on the Y chromosomal characteristics shared between the Soliga population and other Indian tribes as well as western Eurasia and Sub-Saharan Africa groups. Some noteworthy findings of this present analysis include the following: The three most frequent haplogroups detected in the Soliga population are F*, H1 and J2. F*, the oldest (43 to 63 Kya), has a significant frequency bias in favor of Indian tribes versus castes. This observation coupled with the fact that Y-STR haplotypes shared with sub-Saharan African populations are found only in F* males of the Soliga, Irula and Kurumba may indicate a unique genetic connection between these Indian tribes and sub-Saharan Africans. In addition, our study suggests that haplogroup H is confined mostly to South Asia and immediate neighbors and the H1 network may indicate minimal sharing of Y-STR haplotypes among South Asian collections, tribal and otherwise. Also, J2, brought into India by Neolithic farmers, is present at a significantly higher frequency in caste versus tribal communities. This last observation may reflect the marginalization of Indian tribes to isolated regions not ideal for agriculture.

1. Introduction

The Soliga people, who are believed to be representative of India's original inhabitants (Ray, 1973; Thapar, 1966; Basu et al., 2003; Majumder, 2008; Morlote et al., 2011) reside within the Chamarajanagar district of Karnataka, a state at the southernmost tip of the Indian sub-continent (Morlote et al., 2011). The name “Soliga”, which is a Soliganudi (a dialect of the Dravidian language family) word for “people of the bamboo”, reflects the tribal community's forest-bound lifestyle and close relationship with nature (Zaraska, 1997; Morlote et al., 2011). In the last quarter of the 20th century, the Soliga population was estimated to be about 20,000, and approximately 2000 of these individuals inhabit the forests of the Biligiri Rangana Hills (Morab, 1977; Morlote et al., 2011). Soliga society is divided into several clans, or endogamous sects (Morab, 1977) but does not participate in the caste system (Sujatha, 2002; Morlote et al., 2011). Hence, Soliga is designated as one of India's several hundred Scheduled Tribes. Physical characteristics of the Soliga people such as dark complexion, curly hair, short-stature, dolicho-cephalic skulls and sunken nasal root fit the general description of the Australoids, an ethnic group described by T.H. Huxley in 1870 (Chopra, 1965; Morab, 1977; Majumder, 1998). Indeed, this close phenotypic similarity prompted Huxley to propose the existence of a strong genetic connection between the Soliga and Aboriginal Australian populations. Over a century later, J.B. Birdsell

Abbreviations: STR, short tandem repeat; SNP, Single Nucleotide Polymorphism; Kya, thousand years ago; IRB, Institutional Review Board; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; PAI, polymorphic *Alu* insertion

* 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).
[1] Dept. of Biology, Colorado College, 14 East Cache La Poudre Street, Colorado Springs, CO 80903-3294.

https://doi.org/10.1016/j.gene.2019.100026
Received 5 October 2019; Received in revised form 3 December 2019; Accepted 17 December 2019
Available online 13 January 2020

(Birdsell, 1993) not only endorsed Huxley's hypothesis regarding the physical similarities and possible link between tribal Indians and Australians, but extended the narrative by suggesting that multiple migratory waves out of India around 15,000 years ago resulted in the occupation of Northern and Central Australia (Birdsell, 1993). Birdsell's belief is supported by archeological and genetic data, which suggest that the sub-continent's coast was used extensively in the subsequent colonization of Southeast Asia and Australia (Quintana-Murci et al., 1999; Macaulay et al., 2005).

A mitochondrial DNA analysis by Redd (Redd and Stoneking, 1999) supported a recent link between the Australian Aborigines and southern populations of Southeast Asia. In 2002, Redd and co-workers followed this investigation with their study on the C-M216* Y chromosome marker, which is reported to be present in both India and Australia (Redd et al., 2002). The results of a subsequent chromosome analysis (Hudjashov et al., 2007) failed to detect Y chromosomal evidence of Indian migrations into Australia since these two groups displayed different C sub-haplogroups (C4 and C5, respectively). Nevertheless, other genetic studies (Kirk and Thorne, 1976; Nei and Roychoudhury, 1993; Cavalli-Sforza et al., 1994; Alfonso-Sanchez et al., 2008) corroborated the India-Australia connection. Moreover, this proposed association was further strengthened by a more recent autosomal STR analysis conducted by Morlote et al. (2011).

In addition to the Indian-Australian migrations discussed above, India's coastline may also constitute an integral part of the route used in the Out of Africa Diaspora, estimated to have occurred much earlier (approximately 70,000 Kya) (Cavalli-Sforza et al., 1994). Although it is scientifically accepted that initial journey from Africa into Asia and the eventual demographic movement southeast into Austronesia occurred by way of the Southern Coastal Route (Endicott et al., 2007), many questions remain regarding the genetic constitution of the ancient inhabitants of India.

In present study, we investigate the Y-chromosome constitution of the Soliga, one of India's most ancient and isolated tribal groups, in an attempt to assess its phylogenetic relationship to other subcontinent and worldwide populations. For the first time, the genetic diversity of the Soliga population is examined by analyzing both Y-SNP and Y-STR data.

2. Materials and methods

2.1. Sample collection and DNA isolation

Buccal DNA representing 77 unrelated males was collected from the Soliga tribe residing in the Biligiri Rangana Hills in Karnataka, India. Genealogical history of each donor was recorded for at least two generations in order to assess paternal ancestry. DNA was extracted and purified from the Soliga individuals with the Gentra Buccal Cell Kit as instructed by the manufacturer (Puregene, Gentra Systems, Minneapolis, MN) and samples stored at −80 °C until use. Ethical guidelines stipulated by the Institutional Review Board of Florida International University (IRB) were strictly followed during sample collection, processing and subsequent analysis.

2.2. Y chromosome genotyping

A total of 52 bi-allelic markers were tested using standard methods including PCR-RFLP (Luis et al., 2004), allele-specific PCR (Martinez et al., 2005; Regueiro et al., 2006), PCR amplification and electrophoretic detection of Y chromosome *Alu* insertional polymorphism (YAP or PAI) (Hammer and Horai, 1995), as well as direct sequencing (Luis et al., 2004) as deemed necessary. The Y-SNP markers were evaluated in a hierarchal order to determine Y-haplogroup status of each individual sample. Y-SNP haplogroup assignment and nomenclature is in accordance with the Y Chromosome Consortium (Karafet et al., 2008) and subsequent updates by Underhill and colleagues (Underhill et al., 2010).

2.3. Y-STR haplotyping

In addition to the Y-SNP analysis, 76 of the Soliga samples were also typed for 17 Y-STR loci (DYS19, DYS385a/b, DYS398I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA H4) using the AMP*l*FSTR Yfiler PCR amplification kit as per manufacture's (Applied Biosystems of Life Technologies/Fisher Scientific) recommended protocol. Resulting amplicons were separated on an ABI Prism 3130 XL Genetic Analyzer using the ABI GeneScan 500 LIZ as an internal size standard and fragment lengths were estimated by GeneMapper® v3.2 (Applied Biosystems of Life Technologies/Fisher Scientific). Y-STR alleles were assigned by comparison with an allelic ladder provided by the manufacturer. The size of the DYS389I locus was subtracted from that of the DYS399II locus for all analyses performed. Allelic nomenclature follows the recommendations of the International Society for Forensic Genetics (ISFG) (www.isfg.org).

2.4. Statistical phylogenetic analyses

The Y-SNP and Y-STR frequency distributions of the Soliga study population were evaluated within the context of 101 geographically targeted reference populations (Table 1) in order to assess trans-continental (Africa, Europe and Asia) Y chromosomal distribution patterns. Not all reference populations are included in every Y-SNP or Y-STR analysis.

Frequency contour maps of Central and South Asia based on the three most prominent Y-SNP haplogroups in the Soliga study group (H1, F* and J) were constructed with the Surfer® software version 12 (http://www.goldensoftware.com). In addition to the Soliga collection, the analysis included Y chromosome frequency data of 52 Eurasian reference groups (Fig. 1, Supplementary Table 1). A Correspondence Analysis (CA) was performed (NTSYSpc-2.02i) (Rohlf, 2002) to distill salient Y-SNP phylogeographical associations among the Soliga and 39 African-Eurasian reference collections (Supplementary Table 1). In addition, an analysis of molecular variance (AMOVA) using Arlequin software v 3.5 (Excoffier and Lischer, 2010) partitions Y chromosome haplogroup variation of these 40 collections with respect to both geography (9 groups: Africa, Central Asia, Eastern Europe, the Himalayas, Northeast Asia, South Asia, Southeast Asia, Southwest Asia and Western Europe) and linguistic family (10 groups: Afro-Asiatic, Altaic, Austro-Asiatic, Austronesian, Dravidian, Indo-European, Kartvelian, Niger-Congo, Sino-Tibetan and Tai-Kadai) at three hierarchical levels: within populations, among populations within groups, and among groups.

For finer genetic resolution several analyses at the level of Y-STR haplotypes were executed. The first is a Multidimensional Scaling (MDS) analysis (SPSS 14.0) generated from a matrix of pair-wise Rst distances (Arlequin v3.5) (Excoffier and Lischer, 2010) based on eight locus (DYS19, DYS389I-DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS439) Y-STR haplotypes of the Soliga collection as well as 38 African and Eurasian reference populations (Table 1). One thousand permutations of Rst estimations were performed on the 741 non-trivial pair wise comparisons and a Bonferroni corrected alpha value (0.05/741 = 0.00007) was used to evaluate significance of each distance estimation (Kayser et al., 2003).

The eight locus (DYS19, DYS389I-DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS439) Y-STR haplotypes of individuals belonging to the three most common Y-SNP haplogroups observed in the Soliga population (H1a, F* and J2b) were used to generate median-joining networks (NETWORK 4.5.1.6 at http://www.fluxus-engineering.com), in which the Y-STR markers are weighted inversely to their repeat variance and the Maximum Parsimony (MP) option to produce the least complex topology. In order to make informed comparisons, the above-mentioned analyses were performed at the highest

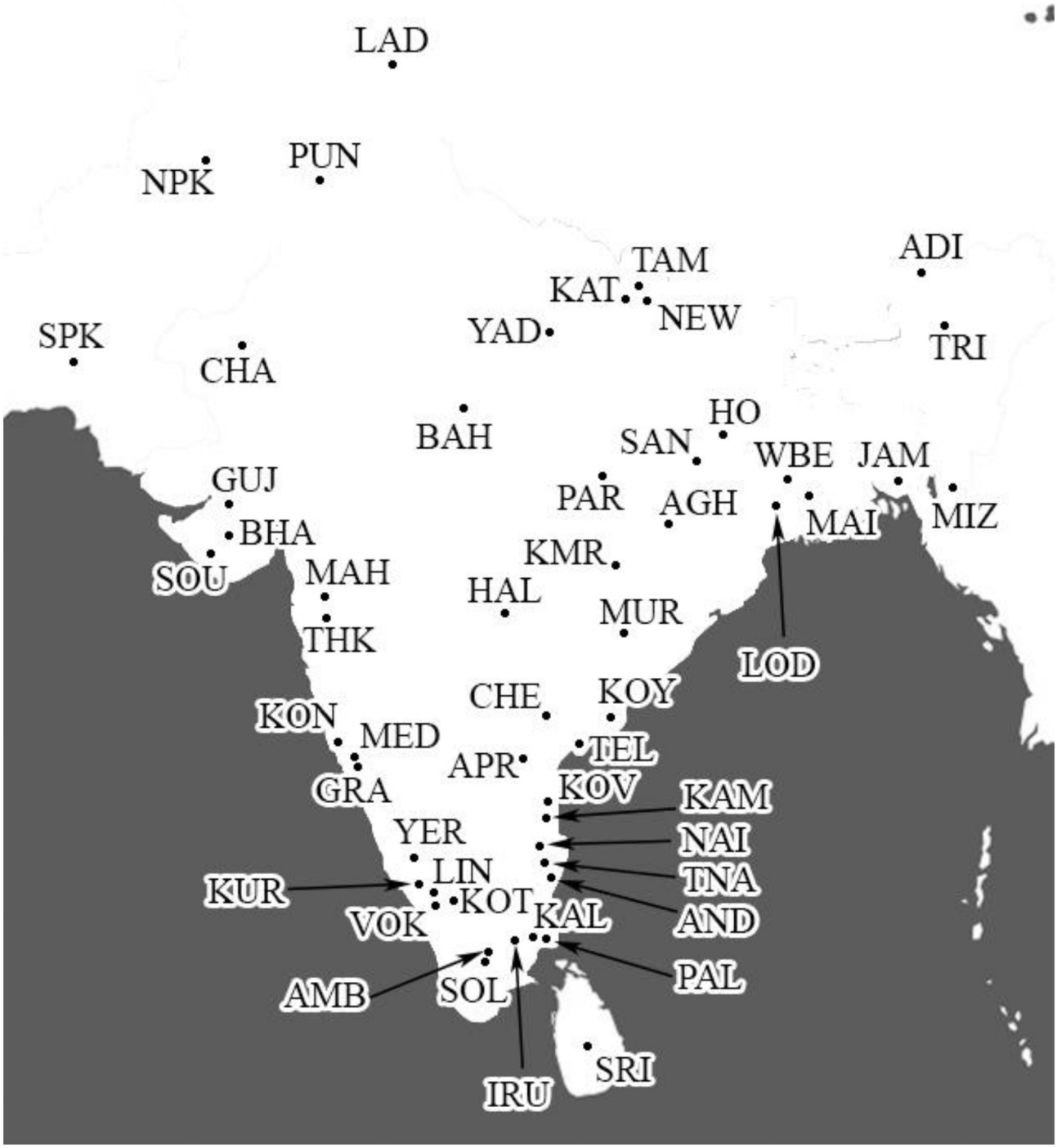


Fig. 1. Population map for frequency contour analysis.

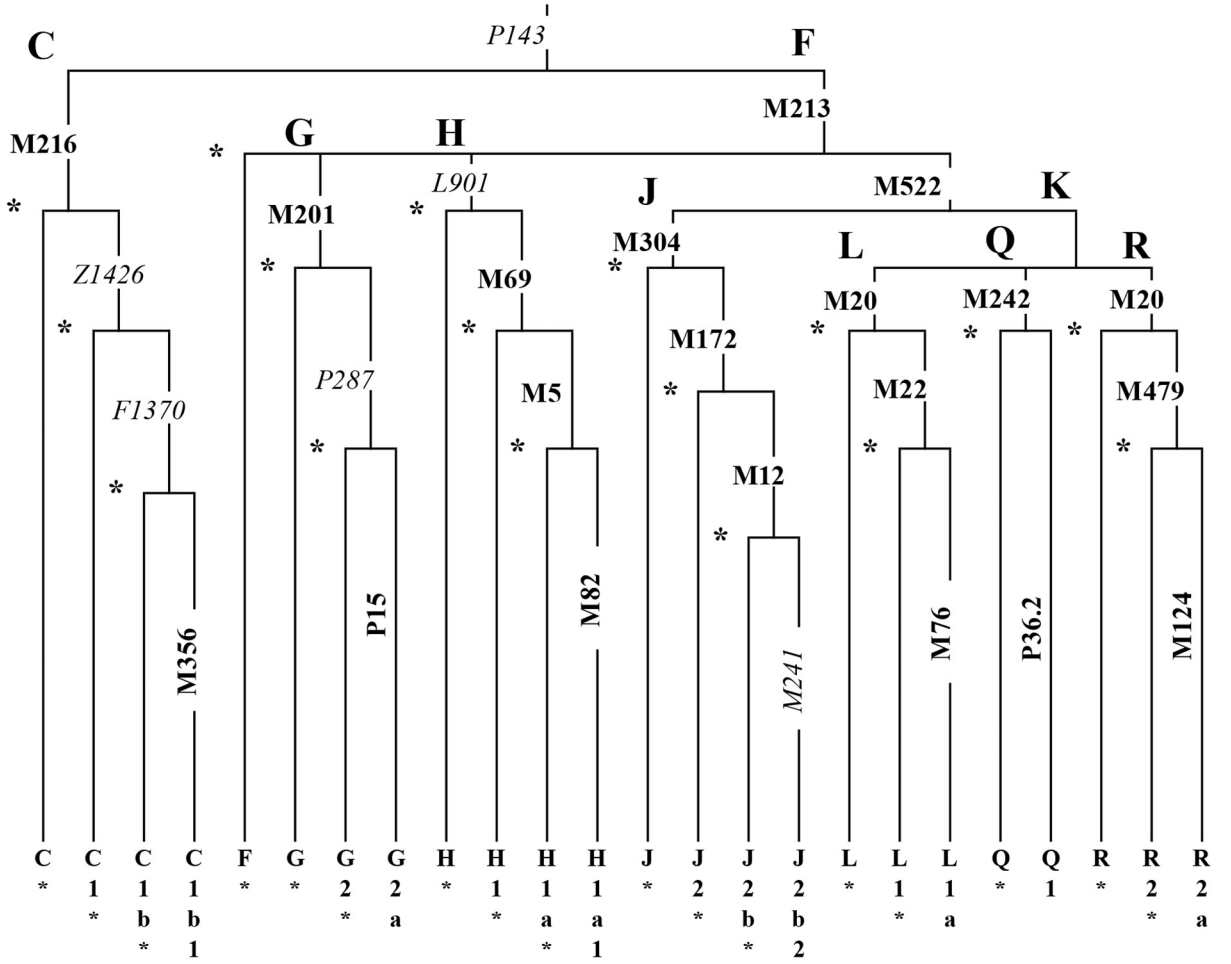
resolution available across all reference populations.

3. Results

3.1. Frequency contour maps

Frequency contour maps were generated from the Y-SNP haplogroup distributions of the Soliga (Fig. 2) and 51 reference collections representing India and the surrounding regions (West: Pakistan and Afghanistan; North: Kyrgyzstan, Kazakhstan, Karalkalpa, Shugun and Ladakh; Northeast: Kathmandu, Tamang, Newar and Tibet) (Fig. 1, Supplementary Table 1). Eight haplogroups comprise the Y-SNP frequency distribution of the Soliga: H1a1 (26%), F* (18%), J2b2 (18%), C1b1 (12%), R2a (10%), L1a (6%), Q1 (6%) and G2a (1%). Contour maps of the three most common Soliga haplogroups, F*, H and J are presented in Figs. 3–5, whereas additional contour maps featuring haplogroups C, G, L, Q and R can be found in the supplementary material (Supplementary Figs. 1, 2, 3, 4 and 5, respectively).

The contour maps of haplogroups F* and H (Figs. 3 and 4 respectively) indicate that these haplogroups are mostly concentrated in the Indian subcontinent and neighboring populations (Punjab, Ladakh, North Pakistan, South Pakistan, Kathmandu, Newar, Tamang and Tibet). Moreover, for both haplogroups F* and H, several high-density zones can be found along the Indian coastline (west, east and south).



Soligas (N=78)				9	14			1				20				14			5		5			8
%				11.5	17.9			1.3				25.6				17.9			6.4		6.4			10.3

Fig. 2. Soliga haplogroups and frequencies phylogeny.

The F* contour frequency pattern (Fig. 3) reveals four of such coastal pockets: northwest (Bharwad and Maharastra at 0.28 and 0.24 frequencies, respectively), southwest (Yerava: 0.43), southeast (Irula: 0.30) as well as the central east coast (Koya: 0.36).

The area of the highest H frequency (Fig. 4) is located near the central east coast and encompasses the populations of Muria (0.8), Koya (0.61) and Kamar (0.48). Several other high density regions are located on the southern (Kurumba: 0.73) and northeast (0.5 in Mahishya) coasts. Soliga, exhibits an H frequency of 0.28.

The contour frequency map of Haplogroup J (Fig. 5), which is associated with the Neolithic Diaspora from west Asia (Semino et al., 2000), reveals a geographical range encompassing those of F* and H. In addition, J exerts a strong presence in Pakistan along the coast (South Pakistan: 0.25) as well as inland (Punjab: 0.21). As in F* and H, several areas of high concentration are located along the Indian coastline (northwest: Thakur at 0.27, southeast: Andh at 0.35 and northeast: Lodha at 0.35). The J frequency of the Soliga study group is 0.17.

In interpreting the frequency distribution of the F*, H and J haplogroups as well as the other contour maps, it is important to keep in mind that there is a geographical bias due to due greater number of coastal populations represented in the number of reference populations (see Fig. 1).

3.2. Correspondence analysis

The correspondence analysis (CA), presented in Fig. 6, is based on the phylogeographical distribution of Y-SNP haplogroups among 40 populations (Supplementary Table 1). Together, Axis 1 (25%) and Axis 2 (20%) explain 45% of the total variation of the analysis. Three major assemblies are featured. The first is a tight, East Asian group at the bottom left of Quadrant II, characterized by a relatively high frequency of Haplogroup O. The second consists of a sub-Saharan African assemblage sequestered in the top right corner of Quadrant I. Most of these populations have high levels of haplogroup E. In addition to these, a large diffuse constellation comprising European, Western Asia, Central Asia and South Asia collections spans the left half of Quadrant IV and the far right section of Quadrant III near the X origin. Its formation is driven by the relatively strong presence of Haplogroup R and J. Within this third grouping, the two Himalayan groups (Kathmandu and Newar) and the majority of the South Asian collections veer to the left of the more western populations, most likely due to a higher proportion of Haplogroup H. Egypt's position between Eurasian and sub-Saharan African collections is, perhaps, attributed to the high frequency of J and R in conjunction with a polymorphic presence of Haplogroup E.

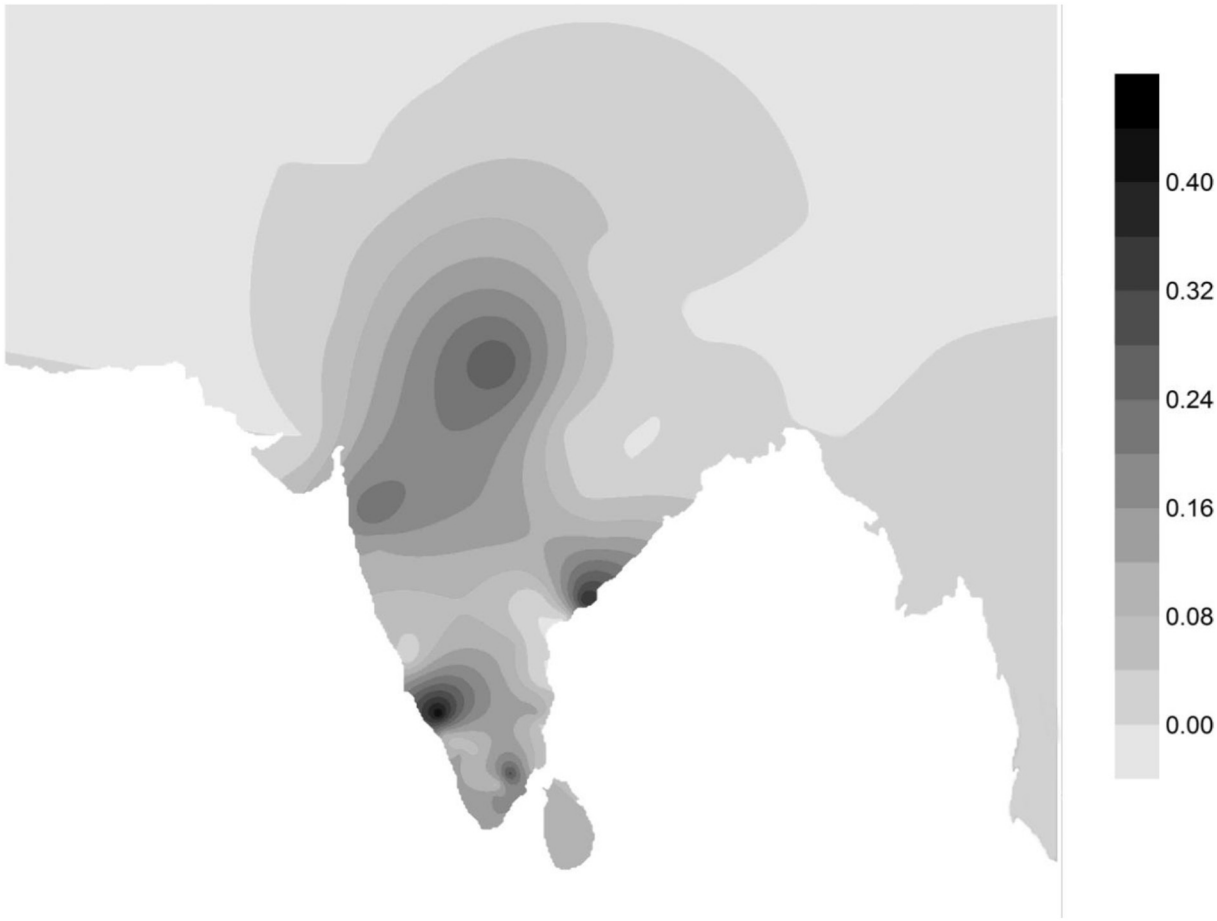


Fig. 3. Frequency contour map F*.

3.3. AMOVA

Supplementary Table 2 displays the results of the Analysis of Molecular Variance (AMOVA) based on the geographical and linguistic partitioning of Y-SNP haplogroup variation for 40 populations. According to these results, there is significant variation (p value < 0.00001) at all three levels (among groups, among populations within groups and within populations) for both partitioning criteria. Moreover, the distribution of variation among these three tiers is very similar for geography and linguistics.

3.4. Multidimensional Scaling

The Multidimensional Scaling (MDS) plot generated from Y-STR Rst pair-wise distances of 39 populations (Supplementary Table 1) is displayed in Fig. 7. This plot features two distinct clusters as well as a large dispersed scattering spread out between Quadrants II and III composed of eleven collections: five South Asian (Kurumba, Irula, Soliga, Thakur and Vokkaliga), two Central Asian (North Afghanistan, South Afghanistan), two Himalayan (Newar, Tamang) and two Southwest Asian (Yemen and Saudi Arabia1). Four of the South Asian populations (Irula, Soliga, Thakur and Vokkaliga) hover near the left of the plot origin, whereas the African cluster is located to its right in Quadrants I and IV and features all five sub-Saharan African populations (Cameroon1, Gabon, Mozambique, Rwanda1 and Uganda). Kurumba is sequestered in the upper left hand corner of the second quadrant. The third group is a tight cluster in Quadrant IV near that of the Africans and consisting of 23 geographically diverse collections. From an enlarged view of this assemblage (Supplementary Fig. 6), it is apparent that the following eleven collections are closest to the Africans: Southwest Asian (Iran and Oman1), South Asian (Lingayat, Madras and Mahadeo Koli), East European (Croatia, Greece and Serbia), West European (Austria and Italy) and two Himalayan (Kathmandu and Tibet), whereas, those populations most distant to the African groups are the Northeast (Japan, Korea, North China and South China) and Southeast (Indonesia, Malaysia, Philippines, Thailand and Vietnam) Asians, which lie in Quadrant IV.

3.5. Network analysis

The first of three network phylogenies is generated from Y-STR data (DYS19, DYS389I-DYS389II, DYS390, DYS391, DYS392, DYS393, DYS439) of F* males (Fig. 8). This network encompasses eight South Asian groups (Irula, Kurumba, Lingayat, Madras, Mahadeo Koli, Soliga, Thakar and Vokkaliga), three Southeast Asian (Indonesia, Philippines and Thailand) collections as well as one from Northeast Asia (South China). Except for a large central node representing ten South Asian males belonging to three ancient tribal populations (1 Kurumba, 4 Irula and 5 Soliga), there is no inter-population haplotype sharing. Radiating out from this multi-ethnic node by one mutation step are three Soliga, one Madras and one Vokkaliga male leading to a Vokkaliga cluster comprising an eight member node with relative short branches ending in singletons. In addition to the Vokkaliga grouping, there are two mono-ethnic branches, each five mutational steps from the network's center. One of these contains four Thakar and the other, three Irula individuals. However, for the most part, the remaining South Asian haplotypes (including the two Soliga singletons not part of the central cluster) are scattered throughout the network. The Indonesian singletons are not too far from the network's center, whereas the remaining Southeast Asian F* Y-STR profiles occupy peripheral and terminal

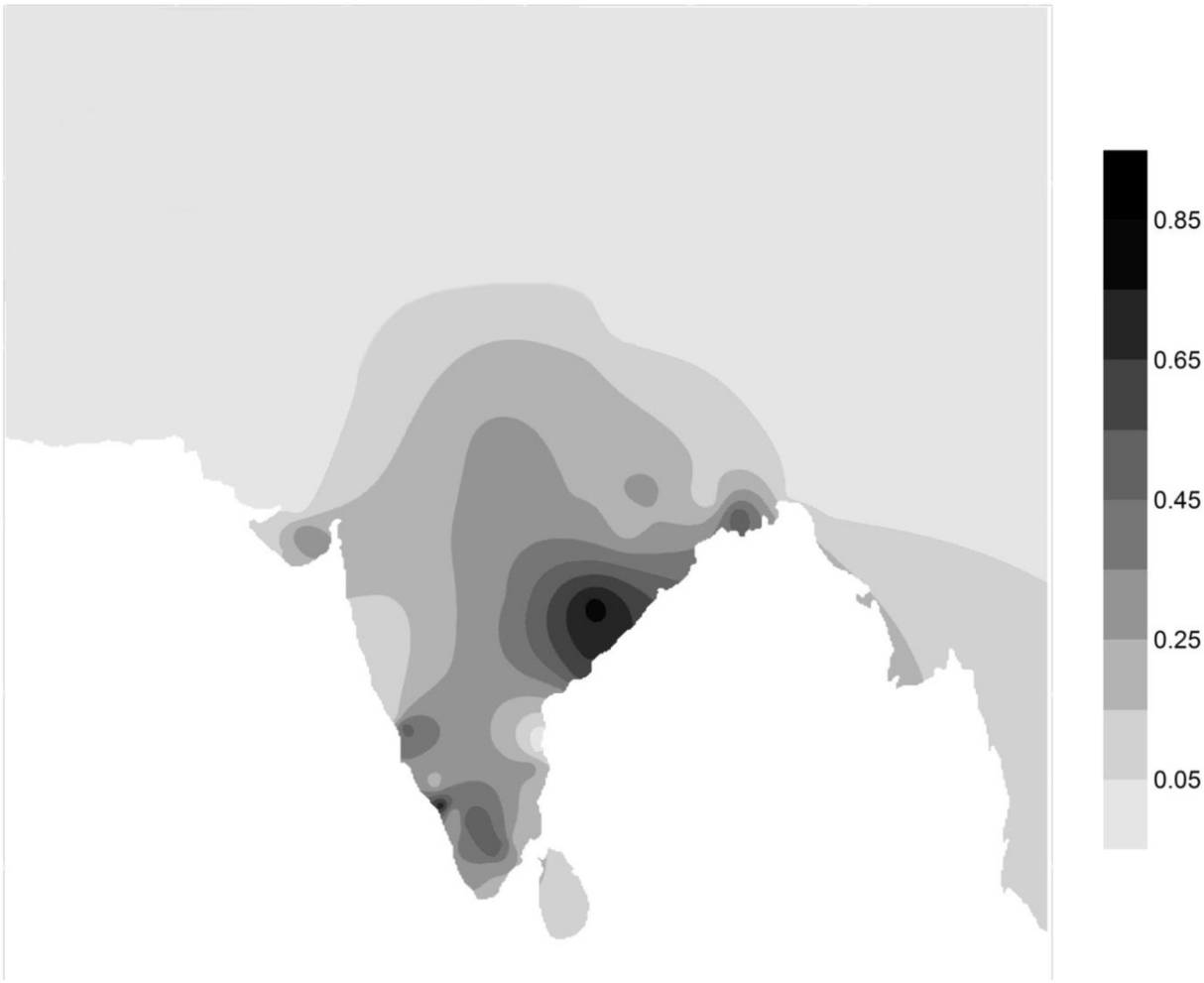


Fig. 4. Frequency contour map H.

positions. The four South China F* Y-STR haplotypes are assigned to the same marginal set of branches (Fig. 8, lower right) containing the majority of the Philippine profiles.

The H1 network of the current study (Fig. 9) represents individuals from eleven collections, one Central Asian (South Afghanistan), two Himalayan (Kathmandu and Newar) and eight South Asian (Irula, Kurumba, Lingayat, Madras, Mahadeo Koli, Soliga, Thakar and Vokkaliga). There are eight multi-ethnic haplotype sharing events. Seven of these involve only two populations either a South Asian pair or a combination of Himalayan and South Asian. A triplet node is composed of Kathmandu, Madras and Vokkaliga. Soliga participates in one of the dual-ethnic Y-STR haplotype events with its geographical neighbor, Kurumba. Apart from these heterogeneous intersections, there are numerous mono-ethnic nodes representing one or more individuals. Most notable, is the large node of seven Soliga profiles. Overall, however, intra-population substructure appears to be minimal since most collection-specific profiles are dispersed throughout the plot.

The J2b network of Fig. 10 includes collections from South Asia (Lingayat, Madras, Soliga and Vokkaliga), The Himalayas (Kathmandu, Newar), Central Asia (South Afghanistan), East (Turkey) and West (Spain) Europe. The J2b Y-SNP haplogroup, characterized by the presence of M12, is common to South Asia, East Mediterranean, Iran as well as parts of Europe. It is believed to have diverged from J2 (M172) in either the Caucasus or the Zagros Mountains at least 10 Kya since this is the estimated date of earliest documented J2b individual discovered in Western Iran (International Society of Genetic Genealogy at https://isogg.org) (Zalloua and Wells, 2004; Al-Zahery et al., 2003). This network is almost star-shaped with a large multi-ethnic center encompassing six profiles (two Spain, and one each of Kathmandu, Madras, South Afghanistan and Turkey). Four of five branches radiating from this central node lead to South Indian profiles differing by three or less mutation steps from the central haplotype. Other multi-ethnic sharing events include a profile common to two males, one South Afghanistan and one Madras as well as large node representing Turkey and Soliga (shared by 1 and 11 individuals, respectively). There appears to be little inter-population and intra-population structure except for the Soliga collection, which exhibits a severe reduction of Jb2 Y-STR haplotype diversity.

4. Discussion

Previous investigations highlighted possible genetic links between India, Indonesia and Australia using Mitochondrial (Redd and Stoneking, 1999) and Y chromosome (Redd et al., 2002) markers. In addition, an autosomal STR analysis (Morlote et al., 2011) provided genetic support for T.H. Huxley's proposed South Indian-Australian connection based on phenotypic similarities between the Soliga and Australian Aborigines (Huxley, 1870; Morlote et al., 2011). Whereas, Morlote et al. (2011) involves the use of autosomal SNPs to explore a South Asia to Australian colonization around15 Kya (Birdsell, 1993), this current study focuses on Y chromosomal linkages between Indian tribes and populations from West Eurasia and Africa. These genetic associations may illuminate earlier trans- and intra-continental migrations to and from the Indian sub-continent. More specifically, we

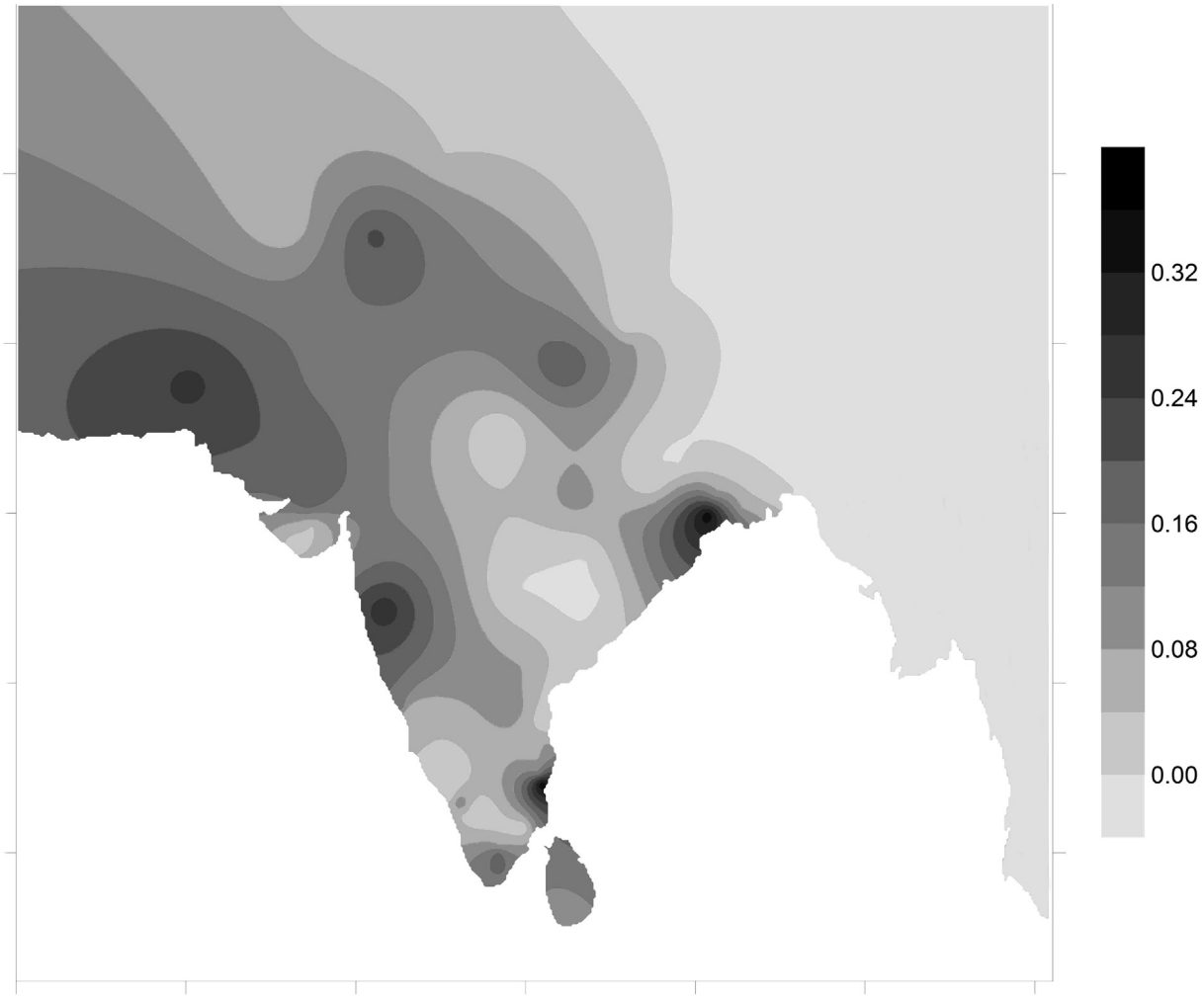


Fig. 5. Frequency contour map J.

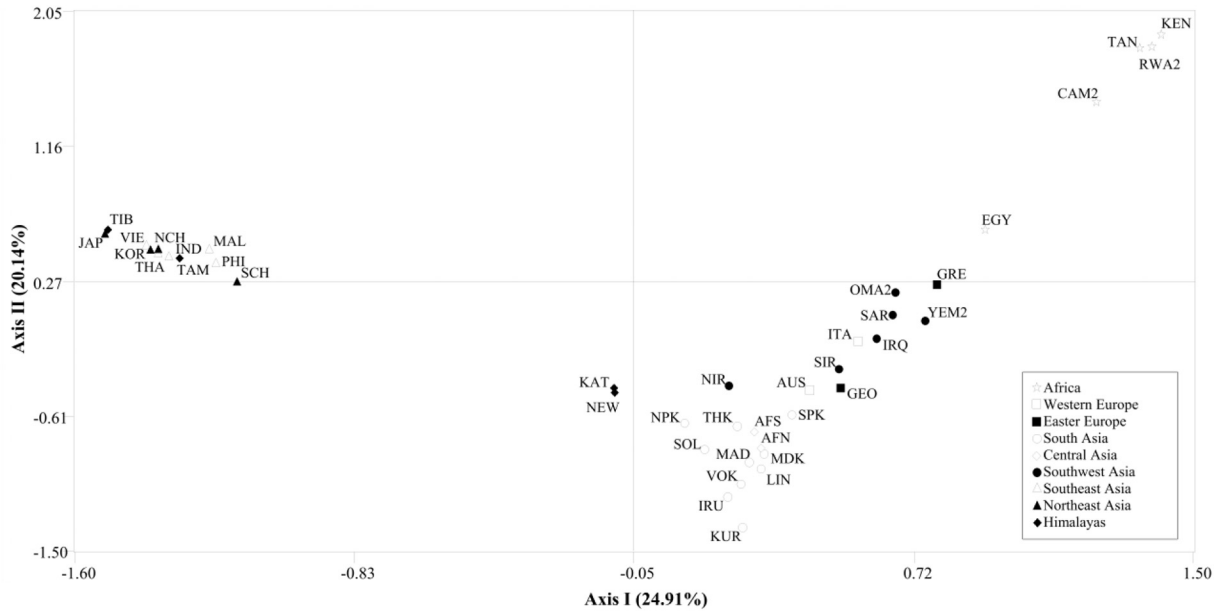


Fig. 6. Correspondence analysis.

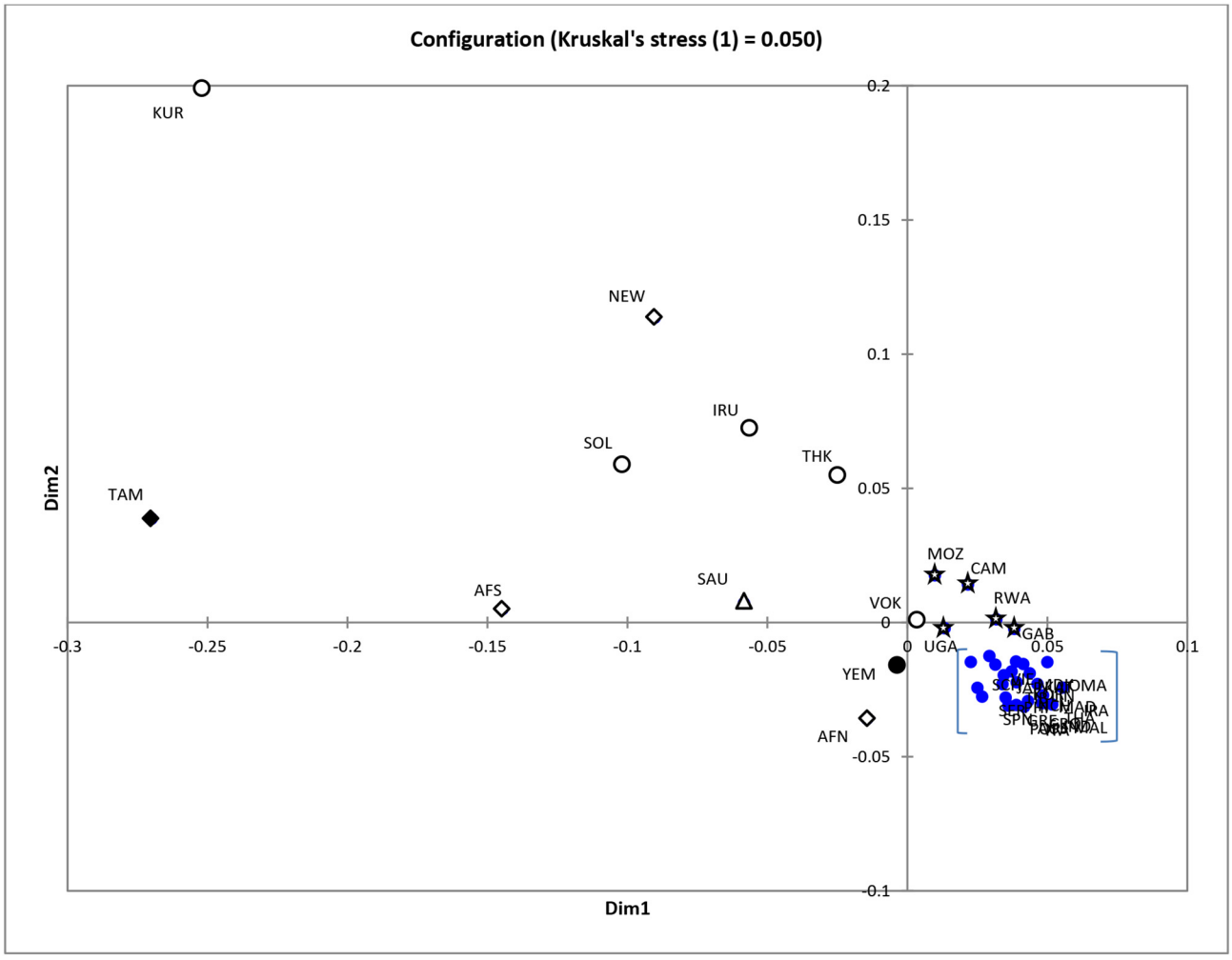


Fig. 7. Multidimensional Scaling.

investigate the Y chromosome composition of the Soliga people, a tribal community at the southern tip of India, against a back drop of globally and locally targeted reference collections (Supplementary Table 1) to elucidate paternal linage affiliations between the study group and populations at three geographic levels: sub-continental (Indian), continental (Eurasian) and trans-continental (African). We employ an array of analyses including contour mapping, correspondence analysis (CA), multidimensional Scaling (MDS) and network analysis as discussed above (Materials and Methods). Our aim is to assess Y-chromosomal phylo-geographical distributions and further delineate early and more recent migration patterns, which may underscore possible genetic connections involving the Soliga and other Indian, Eurasian and African communities. Below we discuss noteworthy results of the above analyses within the context of these objectives.

A comparison of the CA plot (Fig. 6), versus the MDS (Fig. 7) reveals several interesting findings. In the CA analysis (Y-SNP haplogroup frequency data), the west Eurasian (European and west Asian) and South Asian groups (including the Soliga and 7 other Indian populations) occupy Quadrant IV in relative proximity. Also, the west Eurasian groups are closer than the South Asians to the four sub-Saharan collections (Cameroon2, Kenya, Tanzania and Rwanda2) sequestered in the top right corner of Quadrant I. However, the MDS (generated from a Y-STR Rst distance matrix) inserts the sub-Saharan group (Cameroon1, Gabon, Mozambique and Rwanda1) between the South Asian and western Eurasian groups. Another interesting distinction between the two analyses is that the CA Indian congregation is much tighter (occupying the bottom left of Quadrant IV) than that of the MDS, which spans the entire second quadrant and spills into the compact west Eurasian assembly (Fig. 7) of Quadrant IV. These differences may highlight genetic connections and population dynamics at recent versus more ancient time depths due to the faster Y-STR effective mutation rates in comparison to those of the Y-SNP loci (a mean of 2.9×10^{-5} repeat units/locus/year $\pm$ 2.4×10^{-5} across loci within a Y-SNP haplogroup versus a mean of 7.5–8.9 $\times 10^{-10}$ substitutions/locus/year respectively) (Zhivotovsky et al., 2004, Balanovsky et al., 2017).

The CA (Fig. 6), based on major Y-SNP haplogroup frequency distributions, may reflect the initial Out of Africa Diaspora (70 Kya) (Cavalli-Sforza et al., 1994; Lahr and Foley, 1994; Quintana-Murci et al., 1999; Macaulay et al., 2005). This early Paleolithic migration is believed to have introduced the CT haplogroup to Asia. CT arose from haplogroup BT approximately 88 Kya (Karafet et al., 2008) in Northern Africa (Wang et al., 2014) by virtue of the M168 and M294 polymorphisms. Shortly after modern humans entered Asia through the Southwest corner of the Arabian Peninsula circa 70 Kya (Zhivotovsky et al., 2003; Stone et al., 2006; Underhill and Kivisild, 2007), CF emerged around 65 Kya (Underhill and Kivisild, 2007; Karafet et al., 2008) from the CT haplogroup with the gain of the p143 polymorphism. Subsequently, an eastward trajectory along the southwest Asian coastline brought CF into India (Underhill and Kivisild, 2007; Karafet et al., 2008).

The Soliga Y-SNP frequency data reveals three dominant markers, F* (18%), H (as H1 at 26%) and J (as J2b at 18%). F* is the basal member of the F haplogroup, which, according to current scientific consensus, arose within India (Kivisild et al., 2003; Segupta et al., 2006;

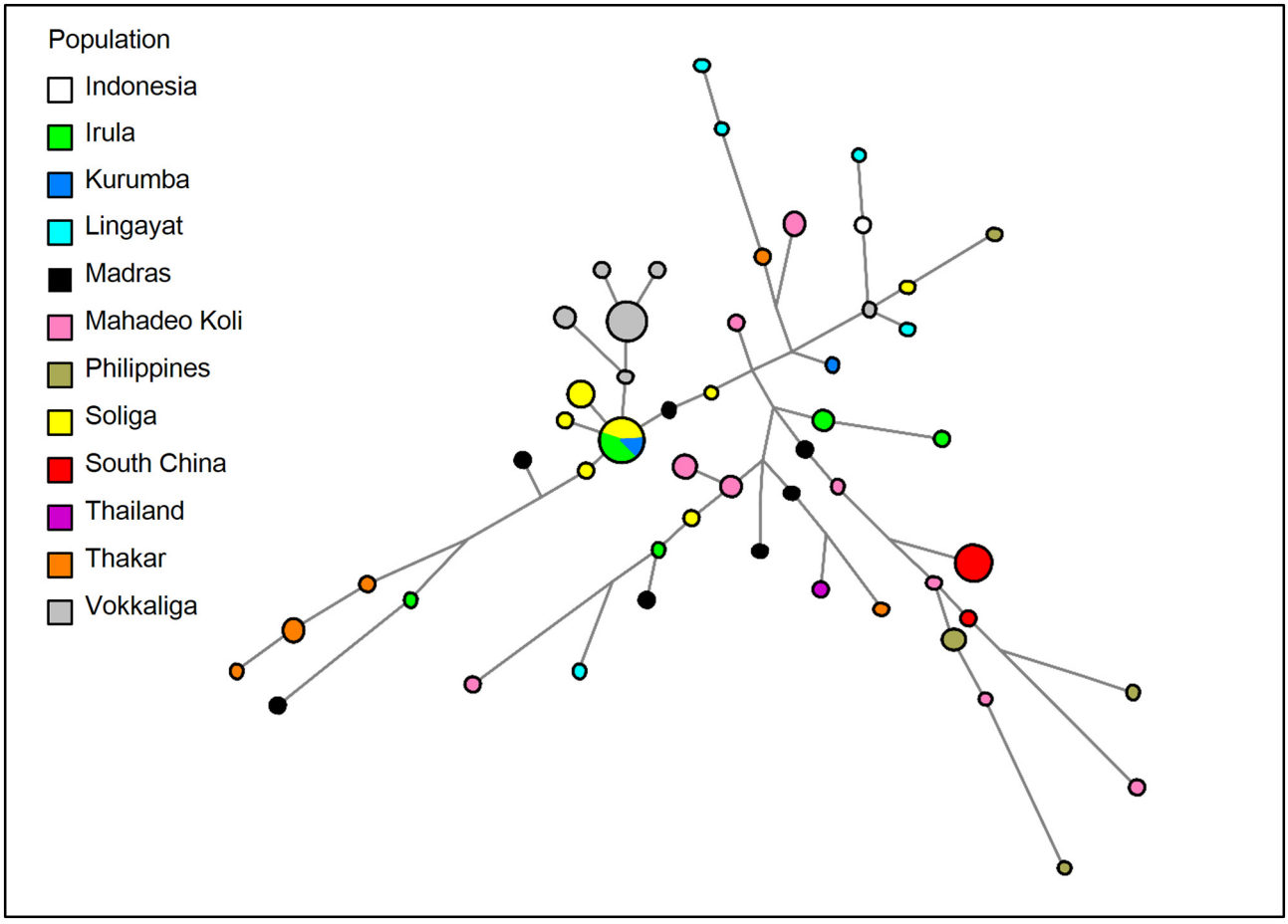


Fig. 8. F* network.

Sahoo et al., 2006; Arunkumar et al., 2012) from the CF haplogroup by an acquisition of the M89 polymorphism between 43 and 63 Kya (Hammer and Zegura, 2002; Karafet et al., 2008; Raghavan et al., 2014). In turn, the macrohaplogroup, GHIJK, descended from F in India around 45 to 50 Kya (https://isogg.org). H and J haplogroups radiated out from GHIJK at different times and locations within Asia.

The emergence of F (and F*) represents one of the earliest chapters in Asian colonization from Africa (Kivisild et al., 2003; Segupta et al., 2006; Sahoo et al., 2006). The contour map frequency distribution of the basal F (F*) shows that this haplogroup appears to be concentrated along the perimeter of the Indian sub-continent, a pattern consistent both with India as the proposed birthplace of F* as well as the scientifically endorsed trajectory of the Southern Coastal Route used in the initial Asian colonization by modern humans (Zhivotovsky et al., 2003; Stone et al., 2006; Underhill and Kivisild, 2007). Haplogroup F* appears to be more prevalent in the Indian tribal communities than in caste populations (Segupta et al., 2006; Thangaraj et al., 2010) (Table S2) ($p \ll 0.001$ at alpha = 0.05) and the majority of tribal F* is detected in collections from southern and western India (such as the Soliga, Irula, Kurumba, Mahadeo Koli and Thakur). The tribal/caste F* bias may result from severe genetic drift caused by restricted gene flow with the outside world as well as low rate of population growth, which, in turn, may stem from millenniums of relative isolation coupled with a subsistence economy (http://factsanddetails.com/india). Evidence of reduced admixture can be seen in the F* network (Fig. 8). There is only one multi-ethnic node (Soliga, Irula and Kurumba) in the midst of numerous singletons and intra-population sharing events involving Indian tribes. The diffuse arrangement of the tribal populations (Soliga, Irula, Kurumba Mahadeo Koli and Thakur) in the MDS (Y-STR Rst distance data) (Fig. 7) may also reflect protracted genetic drift.

Perhaps the most intriguing finding of this current study is the presence of nine, eight-locus Y-STR haplotypes (Supplementary Table 3) shared by members of three ancient Indian tribes (Irula, Kurumba and Soliga) with one or several males from five sub-Saharan African populations (Cameroon1, Gabon, Mozambique, Rwanda1 and Uganda). All of the Indian Y-STR haplotypes shared with Africans are on an F* background and no other collection included in this current study (Indian or otherwise) exhibits these profiles. A network of the nine Indian/African Y-STR (Supplementary Fig. 7) features a large node containing 29 members (Indian: 4 Irula, 1 Kurumba, 5 Soliga, sub-Saharan African: 2 Cameroon1, 14 Gabon, 2 Mozambique and 1 Rwanda1). Four branches radiate out from this center, which all lead to Indian-African nodes separated by a single mutational step (involving a single gain or loss of a repeat unit at one locus) from the central haplotype. Although the African members of this network (Supplementary Fig. 7) lack Y-SNP haplogroup assignments, we propose that these common Y-STR profiles may indicate a unique genetic connection between the three South Asian tribes and the sub-Saharan African groups. The simplest and most parsimonious explanation is that the African individuals involved are members of the F* haplogroup, the presence of which, may be a result of a Asian-African back migration, perhaps, to escape the harsh cold, dry climate of the Last Glacial Maximum (26.5 Kya). If, however, the African individuals belong to other haplogroups, especially those more commonly found in sub-Saharan populations (E, B and A), the following evolutionary sequence of events (or some variation thereof) may explain this finding.

1) African individuals of the CT haplogroup harboring the central profile of the Indian/African Y-STR network (Fig. S7) crossed over into Asia around 70 Kya.

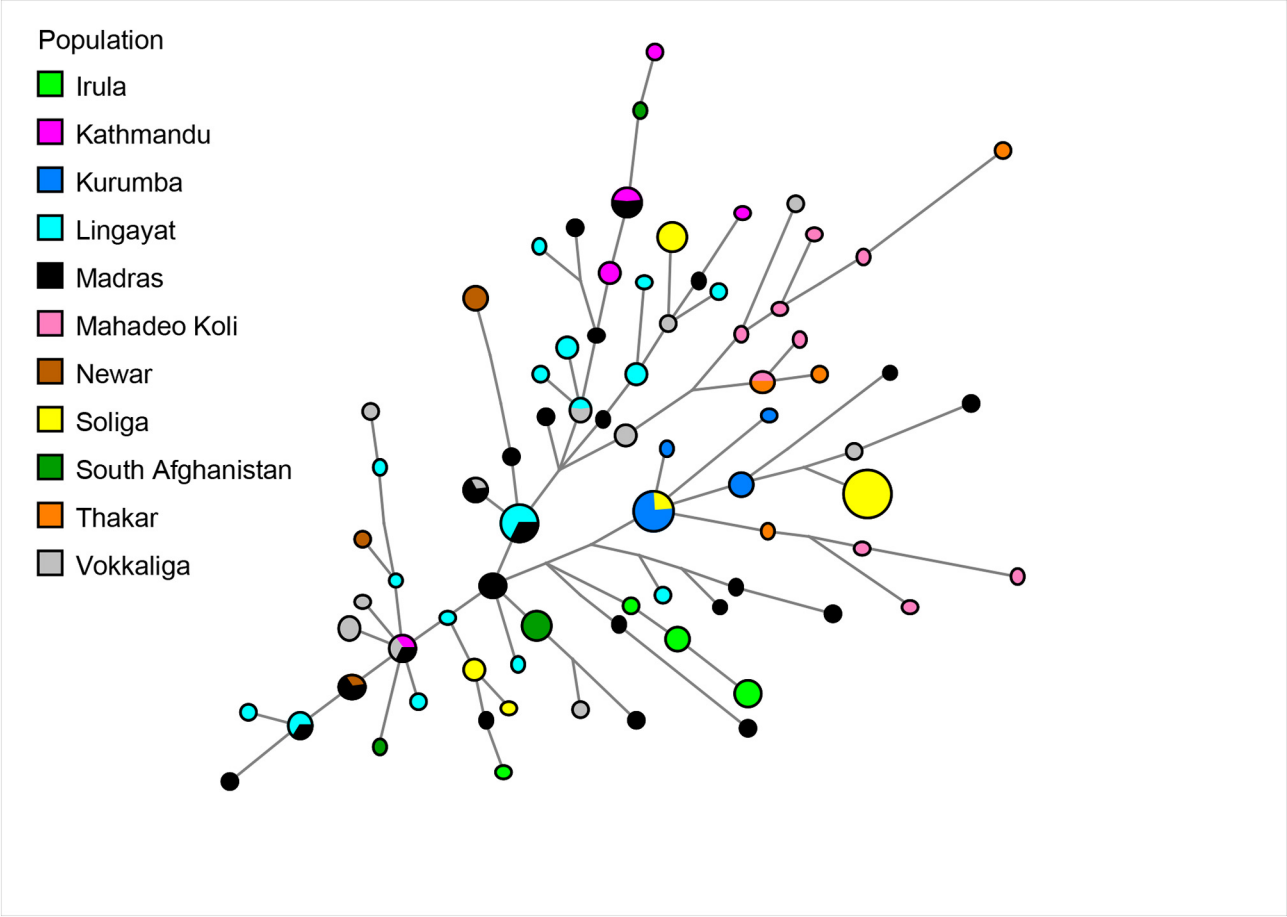


Fig. 9. H1 network.

2) Approximately 65 Kya, at some point along the southern coastal route, the CF haplogroup arose on this Y-STR background and its members migrated eastward into the southern tip of the Indian subcontinent.
3) F may have originated from CF on this Y-STR background somewhere between southern and western India.
4) In turn, the GHIJK macrohaplogroup emerged from F. Population growth and geographical dispersion may have caused a rapid expansion of Y-STR profiles. If the original Y-STR haplotype was carried over into the GHIJK pool, it was soon swamped by descendant haplotypes via rapid genetic diversification.
5) In contrast, the basal F*, associated with these South Asian tribal communities, experienced reduced rates of population growth and thus, a substantially reduced Y-STR diversity. In this setting, it is conceivable that the original Y-STR haplotype was retained at polymorphic frequencies.

According to the International Society of Genetic Genealogy (https://isogg.org/tree/ISOGG_HapgrpH.html), H diverged from the GHIJK macrohaplogroup, in South Asia 30 to 40 Kya. One of H's three sub-haplogroups, H1, which is defined by the presence of M69, is prominent in South Asia especially within the Indian subcontinent and can also be found in Nepal, Afghanistan, as well as the Romani populations of Spain (Kivisild et al., 2003; Di Cristofaro et al., 2013). The prevalence of H and H1 in India is seen both by the H contour frequency map (Fig. 4) and the H1 network (Fig. 9), which, except for Kathmandu, Newar and South Afghanistan, features only Indian groups. In the H1 network, seven multi-ethnic nodes exist within a constellation of singletons and intra-population sharing events including a 10-member Soliga node. Nevertheless, despite the relative lack of common Y-STR haplotypes among these collections, the network shows little structure. It is interesting, however, that the three southern tribal collections which share eight locus Y-STR profiles with sub-Saharan African groups do not participate in inter-population nodes with other South Asian, Himalayan or Central Asian collections. Perhaps, after the emergence and spread of H1, these tribal communities experienced an episode of substantial genetic drift induced by physical or cultural isolation coupled with low population growth.

Haplogroup J is believed to have originated in the Middle East about 48 Kya (https://isogg.org) from macrohaplogroup HIJK via the evolutionary sequence HIJK, IJK, IJ and J. In turn, J2 emerged somewhere within western Asia (Caucasus Mountains, Mesopotamia and the Levant) around 15 to 22 Kya (Zalloua and Wells, 2004; Al-Zahery et al., 2003, 2011), or 19 to 24 Kya (Batini et al., 2015). Associated with the spread of agriculture from the Fertile Crescent to points west (Europe), east (South Asia) and south (Africa) during the Neolithic Era (Zalloua and Wells, 2004; Al-Zahery et al., 2003, 2011), J2 is found in substantial concentrations in Europe, western Asia, South Asia and northern Africa. This distribution is consistent with the J2 contour map (Fig. 5) and J2 network (Fig. 10), which includes only populations from West Asia, South Asia and the neighboring Himalayan collections. The central core of the J2 Y-STR network is a multi-ethnic node representing Kathmandu, Turkey, Lingayat, Madras, and South Afghanistan. It is interesting to note that in South Asia, the J2 haplogroup is significantly more frequent in the Dravidian (19%) versus Indo-European (11%) speaking groups (Segupta et al., 2006), most likely due to the genetic impact the Neolithic expansion from Iran (Singh et al., 2016) west Asia (Abu-Amero et al., 2009; Segupta et al., 2006). In contrast, the Indo Europeans or Indo Aryans migrated from the north Central Asia around 3.8 to 4 Kya (Anthony, 2007; Kuz'mina and

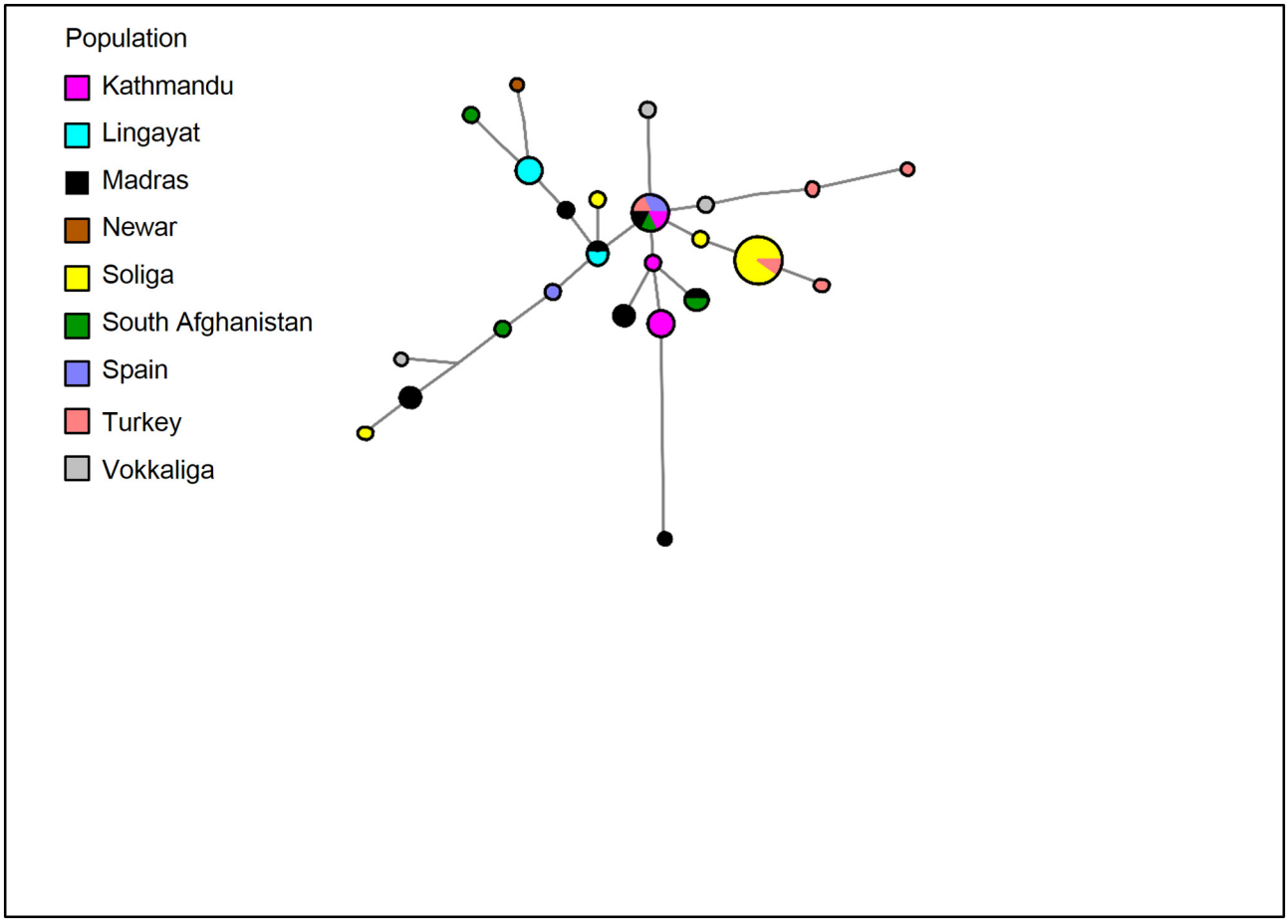


Fig. 10. J2b network.

Mallory, 2007). Initially, these people did not practice agriculture but were pastoralists (Anthony, 2007; Kuz'mina and Mallory, 2007). They also are believed to have introduced the caste system which had, since been was adopted by many non-tribal Dravidian speakers as well. A Chi square test reveals that the J2 frequency is significantly higher in castes versus tribes (Chi-square statistic = 11.8331, p-value < 0.0006) (Supplementary Table 4). The significantly higher Y-SNP contribution of West Asian farmers to castes versus tribal communities may be a result of the insular nature and much lower population growth rate of the later.

5. Conclusion

Together, the three modern day distributions of most common Y chromosomal haplogroups (F*, H1 and J2) present in the Soliga collection chronicle the demographic history of the South Asian tribal communities. However, each of these three haplogroups provides insights into different chapters shaping the genetic history of South Asian tribal communities.

F*, the most basal Soliga haplogroup may shed light on the initial African -Asian migration and subsequent South Asian colonization. F* emerged from CF shortly after its introduction into India via an eastward trajectory along the Southern coastal route (https://isogg.org). In turn, CF descended from the CT haplogroup in western Asia around 65 Kya (https://isogg.org). The frequency bias for F* in Indian tribes may be a result of genetic drift due isolation and low population growth. It may also indicate a common source group, a belief supported by the sharing of Y-STR haplotypes between the Soliga, and two other coastal tribes, Irula and Kurumba. Moreover, the presence of these Y-STR profiles in several sub-Saharan populations and conspicuous absence from the other Eurasian collections suggest a unique genetic connection between Indian tribal groups and sub-Saharan Africans.

H evolved from F via the GHIJK intermediate 30 to 40 Kya (https://isogg.org). Its current distribution in Asia (found primarily in South Asia and immediate neighbors but relatively scarce elsewhere) represents genetic interactions within India and surrounding regions. Also, the H1 network topography reveals minimal sharing of Y-STR haplogroups among South Asian collections, tribal and otherwise. Of the three major haplogroups in Soliga, J2 is the youngest and may represent the later wave of Neolithic farmers into India. This haplogroup shows a frequency bias toward caste populations, which is understandable given that modern day tribes typically live in isolated regions, which may not be conducive to agriculture.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2019.100026.

Author contribution statement

The contribution of each author is indicated below:
Diane J. Rowold – analysis of data and writing.
Shilpa Chennakrishnaiah – experimentation.
Tenzin Gayden – sample collection.
Javier Rodriguez Luis – analysis of data.
Miguel A. Alfonso-Sanchez – analysis of data.
Areej Bukhari – analysis of data.
Ralph Garcia-Bertrand – analysis of data.
Rene J. Herrera – analysis of data and writing.

Declaration of competing interest

No conflict of interest exists.

References

Abu-Amero, K.K., Hellani, A., González, A.M., Larruga, J.M., Cabrera, V.M., Underhill, P.A., 2009. Saudi Arabian Y chromosome diversity and its relationship with nearby regions. BMC Genet. 10. https://doi.org/10.1186/1471-2156-10-59.

Alfonso-Sanchez, M.A., Perez-Miranda, A.M., Herrera, R.J., 2008. Autosomal microsatellite variability of the Arrernte people of Australia. Am. J. Hum. Biol. 20, 91–99.

Al-Zahery, N., Semino, O., Benuzzi, G., Magri, C., Passarino, G., Torroni, A., et al., 2003. Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. Mol. Phylogenet. Evol. 28, 458–472.

Al-Zahery, N., Pala, M., Battaglia, V., Grugni, V., Hamod, M.A., Kashani, B.H., et al., 2011. In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. BMC Evol. Biol. 11. https://doi.org/10.1186/1471-2148-11-28.

Anthony, D.W., 2007. The Horse the Wheel and Language. How Bronze-Age Riders From the Eurasian Steppes Shaped the Modern World. Princeton University Press.

Arunkumar, G., Soria-Hernanz, D.F., Kavitha, V.J., Arun, V.S., Syama, A., et al., 2012. Population differentiation of southern Indian male lineages correlates with agricultural expansions predating the caste system. PLoS ONE 7, e50269. https://doi.org/10.1371/journal.pone.0050269.

Balanovsky, O., Chukhryaeva, M., Zaporozhchenko, V., Urasin, V., Zhabagin, M., Hovhannisyan, A., et al., 2017. Genetic differentiation between upland and lowland populations shapes the Y-chromosomal landscape of West Asia. Hum. Genet. 136, 437–450.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., et al., 2003. Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res. 13, 2277–2290.

Batini, C., Hallast, P., Zadik, D., Maisano Delser, P., Benazzo, A., Ghirotto, S., et al., 2015. Large-scale recent expansion of European patrilineages shown by population resequencing. Nature Com. 6.

Birdsell, J.B., 1993. Microevolutionary Patterns in Aboriginal Australia: A Gradient Analysis of Clines. Oxford University Press, New York.

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

Chopra, P.N., 1965. The Gazetteer of India. Ministry of Education and Social Welfare, India.

Di Cristofaro, J., Pennarun, E., Mazie'res, S., Myres, N., Lin, A.A., Temori, S.A., et al., 2013. Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS ONE 8, e76748. https://doi.org/10.1371/journal.pone.0076748.

Endicott, P., Metspalu, M., Kivisild, T., 2007. Genetic evidence on modern human dispersals in South Asia: Y chromosome and mitochondrial DNA perspectives: the world through the eyes of two haploid genomes. In: Petraglia, M.D., Allchin, B. (Eds.), The Evolution and History of Human Populations in South Asia: Inter-disciplinary Studies in Archaeology, Biological Anthropology, Linguistics and Genetics. Springer Press, Dordrecht 229-224.

Excoffier, L., Lischer, H., 2010. Arlequin suite v 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Eco. Res. 10, 564–567.

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

Hammer, M.F., Zegura, S., 2002. The human Y chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. Ann. Rev. Anthro. 31, 303–321.

Hudjashov, G., Kivisild, T., Underhill, P.A., Endicott, P., Sanchez, J.J., Lin, A.A., et al., 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc. Natl Acad. Sci. 104, 8726–8730.

Huxley, T.H., 1870. On the geographical distribution of the chief modifications of mankind. J. Ethnol. Soc. London. 2, 404–412.

Karafet, T., Mendez, F., Meilerman, M., Underhill, P., Zegura, S., Hammer, M., 2008. New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree. Genome Res 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.

Kirk, R.L., Thorne, A.G., 1976. The Origin of the Australians. Humanities Press, New Jersey.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., et al., 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am. J. Hum. Genet. 72, 313–327.

Kuz'mina, E., Mallory, J., 2007. The Origin of the Indo-Iranians. Brill.

Lahr, M., Foley, R., 1994. Multiple dispersals and modern human origins. Evol. Anthropol. 3, 48–60.

Luis, J., Rowold, D., Regueiro, M., Caeiro, B., Cinniog˘lu, C., Roseman, C., et al., 2004. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. Am. J. Hum. Genet. 74, 532–544.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., et al., 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034–1036.

Majumder, P.P., 1998. People of India: biological diversity and affinities. Evol. Anthropol. 6, 100–110.

Majumder, P.P., 2008. Genomic inferences on peopling of South Asia. Curr. Opin. Genet. Dev. 18, 280–284.

Martinez, L., Reategui, E., Fonseca, L., Sierra-Montes, J., Terreros, M., Pereira-Simon, S., et al., 2005. Superimposing polymorphism: the case of a point mutation within a polymorphic Alu insertion. Hum. Hered. 59, 109–117.

Morab, S.G., 1977. The Soliga of Biligiri Rangana Hills. Anthropological Survey of India, India.

Morlote, D.M., Gayden, T., Arvind, P., Babu, A., Herrera, R.J., 2011. The Soliga, an isolated tribe from Southern India: genetic diversity and phylogenetic affinities. J Hum Genet 56 (4), 258–269.

Nei, M., Roychoudhury, A.K., 1993. Evolutionary relationships of human populations on a global scale. Mol. Biol. Evol. 10, 927–943.

Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K., Santachiara-Benerecetti, A.S., 1999. Genetic evidence of an early exit of *Homo sapiens* sapiens from Africa through eastern Africa. Nat. Genet. 23, 437–441.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., et al., 2014. Upper Paleolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505, 87–91.

Ray, N., 1973. Nationalism in India. Aligarh Muslim University, Aligarh, India.

Redd, A.J., Stoneking, M., 1999. Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations. Am. J. Hum. Genet. 65, 808–828.

Redd, A.J., Roberts-Thomson, J., Karafet, T., Bamshad, M., Jorde, L.B., Naidu, J.M., et al., 2002. Gene flow from the Indian subcontinent to Australia evidence from the Y chromosome. Curr. Biol. 12, 673–677.

Regueiro, M., Cadenas, A., Gayden, T., Underhill, P., Herrera, R., 2006. Iran: Tri-continental nexus for Y-chromosome driven migration. Hum. Hered. 61, 132–143.

Rohlf, F., 2002. NTSYSpc. Exter Publishing, Setauket, NY.

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., et al., 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. Proc. Natl Acad. Sci. 103, 843–848.

Segupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.E., et al., 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am. J. Hum. Genet. (2), 202–221.

Semino, O., Passarino, G., Oefner, P., Lin, A., Arbuzova, S., Beckman, L., et al., 2000. The genetic legacy of Paleolithic *Homo sapiens* sapiens in extant Europeans: a Y-chromosome perspective. Science 290, 1155–1159.

Singh, S., Singh, A., Rajkumar, R., Kumar, K.S., Samy, S.K., Nizamuddin, S., et al., 2016. Dissecting the influence of Neolithic demic diffusion on Indian Y chromosome pool through J2-M172 haplogroup. Scientific Reports 6. https://doi.org/10.1038/srep19157.

Stone, L., Lurquin, P.F., Cavalli-Sforza, L.L., 2006. Genes, Culture, and Human Evolution. Wiley-Blackwell.

Sujatha, K., 2002. Education among scheduled tribes. In: Govinda, R. (Ed.), India Education Report, 1st edn. Oxford University Press, New Delhi, pp. 87–94.

Thangaraj, K., Naidu, B.P., Crivellaro, F., Tamang, R., Upadhyay, S., Sharma, V.K., et al., 2010. The influence of natural barriers in shaping the genetic structure of Maharashtra populations. PLoS ONE 5, e15283. https://doi.org/10.1371/journal.pone.0015283.

Thapar, R., 1966. A History of India. Vol. 1 Penguin Books, London.

Underhill, P., Kivisild, T., 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu. Rev. Genet. 41, 539–564.

Underhill, P.A., Myres, N., Rootsi, S., Metspalu, M., Zhivotovsky, L., King, R., et al., 2010. Separating the post-glacial co-ancestry of European and Asian Y-chromosomes within haplogroup R1a. Eur. J. Hum. Genet. 18, 479–484.

Wang, C., Thomas, M., Gilbert, P., Jin, L., Li, H., 2014. Evaluating the Y chromosomal timescale in human demographic and lineage dating. Investig. Genet. 5. https://doi.org/10.1186/2041-2223-5-12.

Zalloua, P., Wells, S., 2004. Who Were the Phoenicians? *National Geographic Magazine*.

Zaraska, N.A., 1997. Health Behaviors of the Soliga Tribe. Women Master's Thesis. Queen's University, Canada.

Zhivotovsky, L.A., Rosenberg, N.A., Feldman, M.W., 2003. Features of evolution and expansion of modern humans, inferred from genome wide microsatellite markers. Am. J. Hum. Genet. 72, 1171–1186.

Zhivotovsky, L.A., Underhill, P.A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., et al., 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population divergence time. Am. J. Hum. Genet. 74, 50–61.