European Journal of Human Genetics (2001) 9, 695 – 700

www.nature.com/ejhg

ARTICLE

Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India

Gutala Venkata Ramana[1], Bing Su[1], Li Jin[1], Lalji Singh[2], Ning Wang[1], Peter Underhill[3] and Ranajit Chakraborty*[,1]

[1]*Human Genetics Center, University of Texas, School of Public Health, Houston, Texas, TX 77030, USA;* [2]*Center for Cellular and Molecular Biology, Hyderabad, India;* [3]*Department of Genetics, Stanford University, Stanford, CA 94305, USA*

From observations of lack of haplotype sharing based on Y-chromosome specific short tandem repeat (STR) loci, previous reports suggested negligible gene flow among different geographic populations of India. Using Single Nucleotide Polymorphism (SNP) sites in combination with STRs, we observed evidence of haplotype sharing across caste-tribe boundaries in South India. We examined 27 SNPs in the non-recombining region of the Y chromosome to investigate gene flow in 204 individuals belonging to three caste groups (Vizag Brahmins, Peruru Brahmins, Kammas), three tribes (Bagata, Poroja, Valmiki) and an additional group (the Siddis) of African ancestry. Principal component and AMOVA analyses show that the between group component of variation is non-significant ($P > 0.05$), while that among populations within the caste and tribal groups is significant ($P < 0.001$). In particular, the Valmikis and Siddis are close to the caste groups. Of a total of 11 distinct SNP-haplotypes observed, the two tribal groups (Bagata and Poroja) lack the haplotypes H4, H4A, H5A and H16, which are seen in the caste groups. In contrast, all three tribal groups exhibit the Southeast Asian haplotype H11 that is absent in the caste populations. The presence of haplotypes H4, H5, H14, and H16 in the Siddis indicate that they have assimilated considerable non-African admixture. The evidence of haplotype sharing between castes and tribes is also found when the H14 lineage was further subdivided by five STR loci. We conclude that even though these SNP-based Y-haplotypes are able to distinguish the populations, gene flow in these South Indian populations is not as negligible as that inferred from other studies based on Y-specific short tandem repeat markers. *European Journal of Human Genetics* (2001) **9**, 695 – 700.

Keywords: Y-specific SNP haplotypes; gene flow; caste and tribes of South India; Siddis; admixture

Introduction

Uniparental transmission along the male lineage, small effective population size and absence of recombination (except pseudo-autosomal region) are the salient features of Y chromosome[1,2] that makes it suitable for tracing male-initiated migrations. Extensive studies using DNA sequencing and HPLC have enabled to identify numerous single nucleotide polymorphisms (SNPs) on the Y chromosome.[3,4] These SNPs are single base changes or insertion/deletions, which are slowly evolving in comparison with the short tandem repeat markers, which evolve more rapidly. India represents one of the most diverse regions in the world wherein the populations exhibit enormous diversity in terms of language, culture, and ethnicity. A vast majority of Indian populations belong to the Hindu religion and has over 2000 castes each of which belong to a socially stratified Hindu

*Correspondence: R Chakraborty, Human Genetics Center, University of Texas School of Public Health, P.O. Box 20186, Houston, Texas 77030, USA. Tel. (713) 500 9820; Fax (713) 500 0900;
E-mail: rchakraborty@sph.uth.tmc.edu
Received 23 March 2001; revised 3 July 2001; accepted 4 July 2001

caste cluster.[5] There are over 400 tribal populations in India in addition to other religious groups like Muslims, Sikhs, Christians, Jains and migrant groups such as the Parsees and Siddis.[6] Earlier studies from India, based on Y chromosome short tandem repeat (STR) polymorphisms have shown that there is either negligible or no male gene flow among populations of India.[7,8] In contrast, mtDNA d-loop sequence variation[7] showed higher levels of female gene flow between related caste groups. In this research article, we provide new data on 27 Y-chromosome SNP sites in three castes, three tribes, and Siddis (a migrant population of African ancestry) of Andhra Pradesh, South India, and demonstrate that while these SNP markers reveal a substantial genetic variation among these groups, they also detect an evidence of male gene flow among these population groups.

Materials and methods

A total of 204 unrelated males were sampled from seven different populations of Andhra Pradesh, South India. DNA was extracted from fresh blood samples using standard procedure. The populations sampled are from castes (Vizag Brahmins, Peruru Brahmins, Kammas), tribes (Bagata, Poroja, Valmiki) and the Siddis who are migrants from Ethiopia with African ancestry.

Single nucleotide polymorphism genotyping

All markers used in the study were selected from 166 bi-allelic Y-chromosome markers[4] due to their polymorphism in individuals from the Indian subcontinent and Central Asia. A total of 27 Y-chromosome SNP markers were analysed using either AS-PCR or RFLP protocols.

The SNPs typed are M1 (Yap); M15 (9 bp insertion); M130 (C →T); M48 (A→C); M89 (C→T); M170 (A→C); M172 (T→G); M52 (A→C); M9 (C→G); M11 (A→G); M46 (T→C); M62 (T→C); M122 (T→C); M7 (C→G); M134 (1 bp deletion); M119 (A→C); M50 (T→C); M110 (T→C); M103 (C→T); M95 (C→T); M88 (A →G); M111 (4 bp deletion); M45 (G→A); M120 (T→C); M3 (C→ T); M17 (1 bp deletion); M5 (A→G). In addition, sequencing of M168 and M60 were carried out using the forward primer by means of standard cycle sequencing with fluorescent dideoxynucleotides. The fluorescently labelled extension products were run on ABI 373A DNA sequencer (Perkin Elmer). The Y chromosome specific STR loci[9] typed is: DYS 19, DYS 389I, DYS 390, DYS 391, DYS 393. The Y-STR haplotypes are defined by repeat size of alleles for each of the five loci. Analysis of molecular variance was analysed using the Arlequin software version 1.1.[10,11]

Results

Using the 27 biallelic markers we identified eleven haplotypes in 204 Y-chromosomes. The haplotype frequencies in various populations are shown in Table 1. Also, we present a phylogenetic tree for the present study (Figure 1) under the parsimony assumption,[4,12] which assign H1 as the ancestral haplotype (also observed in the Chimpanzees). Both H1 and H2 are ancient haplotypes present in African and non-African populations. Further, H5 defined by the M9 (C→G) mutation site appears to be the common ancestor for all haplotypes that are distributed in worldwide populations. H11, which is specific to Southeast Asia,[12] is also present exclusively in the tribal populations. Counting the mutation events shown in Figure 1 of Underhill *et al*,[4] in total five mutation events are needed to derive the haplotype H16 from H1.

The principal component analysis (Figure 2) of the haplotype distributions reveals that more than 87% of the haplotype variation is explained by the three principal components. The positions of the populations, by the three principal component scores, do not generally cluster them by their caste or tribal affiliation. For example, the Valmikis are closest to the Peruru Brahmins, both of which are also close to the Siddis (particularly based on the first two principal components). Also, the Vizag Brahmins and Peruru Brahmins are distant from each other (particularly based on PC1), although they belong to the same social hierarchy.

Table 1 Distribution of Y chromosome haplotype frequencies (per cent)

Population	n	*H1*	*H1A M130T*	*H2 Yap+*	*H4 M89T*	*H4A M172G*	*H4B M52C*	*H5 M9G*	*H5A M11G*	*H11 M95T*	*H14 M45A*	*H16 M17 (1 bp del)*
Castes:												
Vizag Brahmins	41				12.2	2.4	19.5	26.8			24.4	14.6
Peruru Brahmins	44		18.2		22.7	11.4	6.8	2.3	6.8		25.0	6.8
Kammas	40					7.5	2.5	32.5	17.5		32.5	7.5
Tribes:												
Bagata	23		8.70				8.7	4.3		47.8	30.4	
Poroja	20						10.0	10.0		30.0	50.0	
Valmiki	24		12.5		8.3	16.7	12.5	4.2	12.5	16.7	12.5	4.2
Migrant:												
Siddis	13	7.7		38.4	15.4			7.7			15.4	15.4

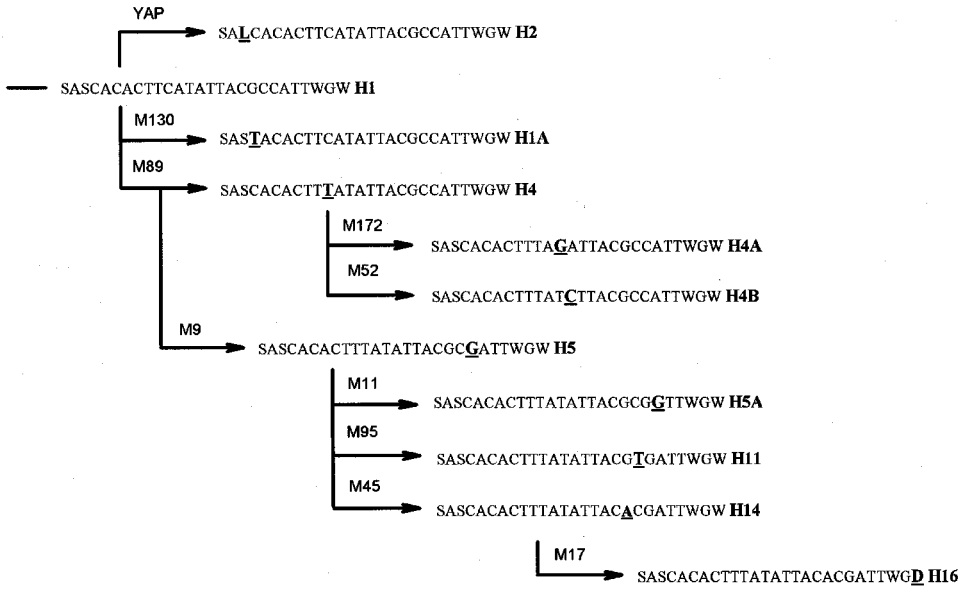

Figure 1 The parsimony tree of the 11 haplotypes based on 27 Y-chromosome SNP sites, observed in seven South Indian populations. The letters 'A', 'T', 'G', and 'C' refer to the nucleotides at the typed mutation sites. The letters 'S' and 'L' refer to the small and large alleles at the insertion/deletion mutation sites (including the YAP-insertion site). 'W' and 'D' refer to the wild type and deletion alleles at the 1 bp deletion sites. The tree is adapted from Underhill *et al.*[4] and Su *et al.*,[12] but abbreviated for the observed 11 haplotypes in the present study. The order of the loci (M15, M5, YAP, M130, M48, M3, M119, M7, M50, M110, M89, M170, M172, M52, M62, M122, M88, M103, M45, M95, M9, M11, M46, M120, M111, M134, and M17) does not necessarily represent their physical order on the Y-chromosome.

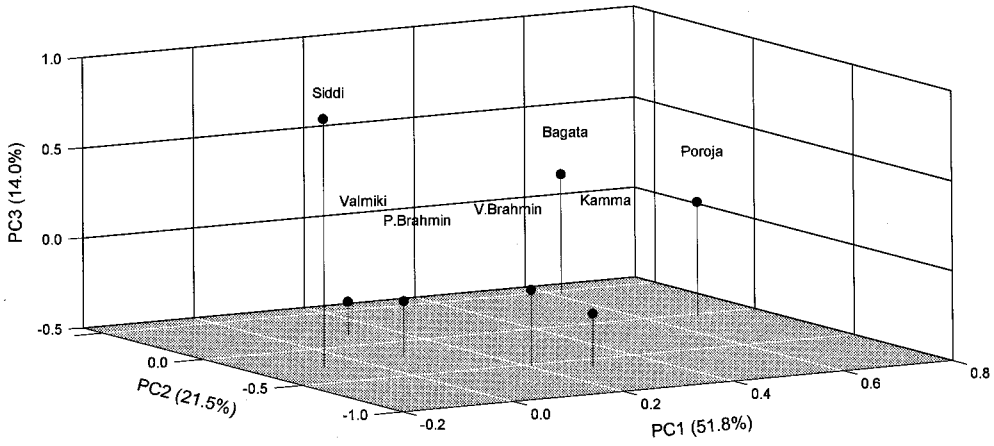

Figure 2 Relative positions of the seven populations, as reflected by the first three principal components based on the SNP-based haplotypes observed in the present study. The numbers in parentheses of the three dimensions represent the per cent of haplotypic variance explained by the respective principal components (PC1, PC2, and PC3).

The variance decomposition (AMOVA) analysis of the SNP haplotype frequencies provides a quantitative support of the same trend of genetic affiliation of these populations. With the seven populations divided into three groups (caste, tribe, and the migrant), and using the phylogeny of the 11 haplotypes as described in Su *et al.*,[12] the estimates of the variance components are shown in Table 2 (second column), along with their empirical levels of significance (column 3). While the populations are clearly distinguishable (V_p=8.8%, $P<0.001$), the variance component ascribed to among group difference (V_g=6.2%) is not significant ($P\approx0.063$). Genetic contact of the Valmikis and Siddis with the caste populations

Table 2 AMOVA showing per cent variation and p value

	SNPs		*H14 STRs*	
Type of comparison	*Seven populations*	*Five populations*	*Seven populations*	*Five populations*
Among groups	6.2 (0.063)	12.5 (0.103)	−9.7 (0.987)	−7.1 (1.000)
Among populations within groups	8.8 (<0.001)	7.1 (<0.001)	23.2 (<0.001)	24.8 (<0.001)
Within populations	85.8 (<0.001)	80.4 (<0.001)	86.4 (<0.001)	82.3 (<0.001)

Seven populations includes (Vizag Brahmins, Peruru Brahmins, Kamams), and (Siddis) and (Bagata, Poroja, Valmiki) in three groups. Five populations includes (Vizag Brahmins, Peruru Brahmins, Kammas), and (Bagata, Poroja) in two groups.

alone does not explain this. Excluding them from the analysis, while the numerical value between group differences becomes larger (V_g=12.5%), it still remains non-significant.

Haplotype sharing and frequency differences of haplotypes can be examined in the light of these observations. It is true that the caste populations (both Brahmin groups and the Kammas) can be distinguished from the two tribal groups (Bagata, Poroja), since the caste populations exhibit the haplotypes H4, H4A, H5A and H16, which are not present in the two tribal groups. In contrast, all the tribal groups show the presence of the Southeast Asian specific haplotype H11.[12,13] However, the Valmikis share haplotypes H1A, H4, H4A, H5A and H16 with caste populations. They also exhibit the Southeast Asian haplotype H11, which is present in the other two tribes, but neither in any of the caste populations nor in Siddis.

The Siddis exhibit H1 and H2 haplotypes, a signature of their African ancestry. Since the H1 individual showed ancestral alleles at M1 (Yap-), M89 (C), and M130 (C) loci, we sequenced this individual for the M168 locus and observed ancestral allele (C). We also sequenced for the M60 locus, and observed insertion of T (1 bp insertion) at this locus (belonging to the haplogroup II).[4] This haplogroup has been previously shown to occur widely in Africa. Thus, its presence in the Siddis corroborates the ancestry of the Siddis from Africa. In addition, they also have non-African haplotypes *viz*., H4, H5, H14, and H16 in their male gene pool, suggesting extensive admixture with the local Indian groups.

Discussion

Our study on the haplotypic diversity based on Y-chromosome SNPs demonstrates that the caste and tribal populations of Andhra Pradesh, South India can be distinguished by the presence of some haplotypes that are unique to these groups (H4, H4A, H5A, and H16 in the caste groups, and H11 in the tribals). However, the presence of haplotypes H4B, H5, and H14 in all of caste and tribal groups studied, and the presence of haplotypes H4A, H5A, H14 and H16 in the Valmikis raise the possibility of extensive gene flow across the caste-tribe distinction of populations in this region of the country. The AMOVA analysis of the frequency distributions of the 11 haplotypes supports this assertion.

Of the haplotypes (H4B, H5, H14, H4A, and H5A) providing the suggestion of caste-tribe gene flow, a more detailed study of the H14 haplotype (defined by M45, an Eurasian marker, and presumably the youngest among these group of haplotypes) provides a further confirmation of our assertion. We have typed Y STR markers *viz*., DYS 19, DYS 389I, DYS 390, DYS 391 and DYS 393 in all individuals (except in seven individuals, because the DNA was exhausted) carrying H14 (n=49 in the combined sample of seven populations). The haplotypes constructed for these five STRs based on repeat size for each locus showed 30 distinct haplotypes, of which five are shared between the seven population groups (Table 3). Even more important is the observation that three of the haplotypes are shared between caste and tribal groups, pointing at the possibility of a recent gene flow between castes and tribes.

Using analysis of molecular variance,[10,11] we partitioned the allele size variances of the five STR loci of the H14 lineage, according to their population affiliation (Table 2, columns 4 and 5). There is no significant difference between the three population groups (V_g=−9.7 $P\approx$0.987), while the distinction among populations is significant (V_p=23.2%, $P<0.001$). As in the analysis of SNP haplotype data, exclusion of the Valmikis and Siddis does not affect this result.

A longer antiquity of haplotypes, as compared to formation of caste and tribal groups, may be proposed to explain the observation of SNP-haplotype sharing of the Valmikis with the caste populations. Two lines of evidence suggest that this may not be the case. First, the non-significant group differences of SNP-haplotype diversity as well as STR-haplotype sharing between the castes and tribes (Tables 2 and 3) suggest evidence of gene flow across caste-tribe boundaries, rather than antiquity of haplotypes. Second, Underhill *et al*,[4] estimated that the average time of adding a new mutation in the non-recombining region of the Y chromosome is approximately 6900 years, which places H14 to have evolved (with three mutations) 20 700 years after H1, and H16 (with five mutations) 34 500 years after H1. With H1 estimated to be 44 000 years old,[4] these may indicate that H14 and H16 may have existed at a time predating the separation of caste and tribes in India.[14] However, we observed haplotype sharing between castes and tribes at the

Table 3 Y short tandem repeat (STR) haplotypes in H14 individuals

Repeat size and sites of polymorphism							*Frequencies in*				
DYS 19	*DYS 389 I*	*DYS 390*	*DYS 391*	*DYS 393*	*VB*	*PB*	*KAM*	*SID*	*BG*	*PO*	*VAL*
14	15	23	10	13	1						
15	14	25	11	13	1	4					
15	14	23	10	13	1						
14	15	23	10	15	1	1					
14	14	23	10	13	1						
15	14	23	10	14	1						
14	14	24	10	14	1						
14	14	23	9	14	1						
15	15	25	10	13	1				1		
15	15	24	10	13	1						
15	14	25	10	13		1	2				1
15	15	23	10	14			2				
16	15	22	10	14			2				
15	15	24	11	13			1				
16	14	25	10	13			1		1		
15	13	25	10	12			1				
15	15	24	11	14			1				
15	14	28	11	14				1			
16	15	24	10	13				1			
14	14	25	11	13		1					
14	14	24	11	13		1					
13	14	22	10	13		1					
15	14	23	11	13						4	
14	15	23	11	14						4	
14	15	22	10	14					1		
17	14	26	10	13					1		
15	15	23	10	13					1		
15	15	26	10	13					2		
15	15	25	11	13							1
16	14	24	10	13							1

VB: Vizag Brahmins, PB: Peruru Brahmins, KAM: Kammas, SID: Siddi, BG: Bagata, PO: Poroja, VAL: Valmiki.

STR level as well within the H14 lineage (Table 3), some of which are at least two mutation steps different from each other. The non-significant caste-tribe group difference of the STR-haplotypes of the H14 lineage supports the gene flow hypothesis rather than the antiquity of the haplotypes.

Our data also suggests that Siddis have assimilated considerable non-African Y chromosomes (haplotypes H4, H5, H14, and H16) from the local Indian populations. The arrival of the Siddis in India dates back to AD 1100[15–17] and they have had social contacts with several local Indian populations. From the combined frequencies of the haplotypes of African (H1 and H2) and non-African haplotypes (H4, H5, H14, and H16), data shown in Table 1 indicates that at least 56% of the male genes of the Siddis could be of Indian origin, consistent with our estimate based on five STR loci we reported elsewhere.[18]

Acknowledgments

All DNA samples analysed here are anonymised, and were collected with informed consent of the subjects. Research supported by the US Public Health Services Research Grant GM 41399 from the National Institutes of Health.

References

1 Jobling MA, Tyler-Smith C: Father and sons: The Y chromosome and human evolution. *Trends in Genetics* 1995; **11**: 449–456.
2 Jobling MA, Tyler-Smith C: New uses for new haplotypes the human Y chromosome, disease and selection. *Trends in Genetics* 2000; **16**: 356–363.
3 Underhill PA, Jin L, Lin A *et al*: Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Research* 1997; **7**: 996–1005.
4 Underhill PA, Shen P, Lin AA *et al*: Y chromosome sequence variation and the history of human populations. *Nature Genetics* 2000; **26**: 358–361.
5 Malhotra KC: Population structure among the Dhangar caste-cluster of Maharashtra, India. In: Lukacs JR (ed) *The people of south Asia: the biological anthropology of India, Pakistan, and Nepal*. Plenum, New York, 1984, pp 295–324.
6 Majumder PP: People of India: Biological diversity and affinities. *Evolutionary Anthropology* 1998; **6**: 100–110.
7 Bamshad MJ, Watkins WS, Dixon ME *et al*: Female gene flow stratifies Hindu castes. *Nature* 1997; **395**: 851–852.
8 Bhattacharyya NP, Basu P, Das M *et al*: Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. *Genome Research* 1999; **9**: 711–719.
9 Kayser M, Cagila A, Corach D: Evaluation of Y-chromosomal STRs: A multicenter study. *Int. J Legal Med* 1997; **110**: 125–133.

10 Excoffier L, Smouse P, Quattro J: Analysis of molecular variance inferred from metric distances among DNA haplotypes. Application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479 – 491.
11 Schneider S, Kueffer JM, Roessli D *et al*: Arlequin ver 1.1 A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 1997.
12 Su B, Xiao J, Underhill P *et al*: Y-chromosome evidence for a northward migration of modern human into Eastern Asia during the last ice age. *Am J Hum Genet* 1999; **65**: 1718 – 1724.
13 Su B, Jin L, Underhill P *et al*: Polynesian origins: Insights from the Y chromosome. *Proc Natl Acad Sci USA* 2000; **97**: 8225 – 8228.
14 Basham AL: The Wonder that was India. New Delhi, India, Rupa and Co., 1981.
15 Enthoven RE: Tribes and castes of Bombay, **Vol III.** Bombay, 1922.
16 Palakshappa TC: The Siddis of North Karnataka. New Delhi, India, Sterling Publishers, 1976.
17 Sorley HT : The Siddis of Kanara – Census of India , **Vol 1**, Part III, 1931.
18 Ramana GV, Wang N, Singh L *et al*: Short tandem-repeat Y-chromosome haplotype data reveals a high level of admixture in the migrant populations, the Siddis, with local Indian populations. *Human Biology,* 2001, (manuscript submitted).