

Insertion/Deletion Polymorphisms in Indian Tribal Populations

P. Veerraju¹, D. A. Demarchi², N. Lakshmi¹ and T. Venkateswara Rao¹

1. Department of Human Genetics, Andhra University, Visakhapatnam, Andhra Pradesh, India

2. Museo de Antropología, Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Córdoba, Argentina

KEYWORDS Alu markers; Indian tribes; gene diversity; genetic distance

ABSTRACT Five Alu markers (Alu APO, PV 92, TPA 25, D1 and ACE) were studied in five tribal populations namely, Konda Reddi, Koya Dora and Konda Kammara of East Godavari district, Lambada and Chenchu of Mahaboobnagar district of Andhra Pradesh. All the five loci were found to be highly polymorphic. While the lowest heterozygosity was observed in the Chenchu the Lambadi shows the highest. Both Neighbour Joining tree and Principal Component analysis based on genetic distances suggest two broad clusters, one formed by the Lambada and Chenchu and the other by the Konda Reddi and Koya Dora with Konda Kammara as an outer element to this three-point cluster. Another cluster analysis carried out along with 19 other Indian populations brings out no distinct cluster of the 5 AP tribes; instead these AP tribal populations are integrated into different subclusters of the UP and Bengal suggesting lack of distinct genetic identity of these AP tribes as far as the few Alu markers are concerned.

INTRODUCTION

The fascinating and complex nature of the Indian population structure presents many intricate and important issues for a population geneticist to resolve. The clear division of Indian population into strictly defined endogamous and hierarchical castes, tribes and religious groups facilitates a clear definition of the study/evolutionary units. This coupled with enormous variety of the Indian peoples has attracted wide spread interest on Indian populations and hundreds of populations have been investigated for traditional genetic markers and biological variables. With the advent of DNA technology, this interest in Indian populations has been redoubled and in the recent years a number of Indian populations have been analysed for a variety of DNA markers, and a number of anthropological hypothesis in vogue have been tested (Bamshad et al. 1996, 1998; Majumder et al. 1999a; Reddy et al. 2005; Kumar et al. 2006, 2007). One of the important observations that emerge from the earlier genetic studies is the clear differentiation of tribal population from the castes (Bhasin and Walter 2001). The tribal populations have also been observed to be relatively more heterogeneous among them, with a relatively long history of settlement in the Indian subcontinent when compared to the castes.

In the present study, we have screened five biallelic Alu insertion markers among the five tribal populations of Andhra Pradesh and compared the results with other caste and tribal populations of India for which similar data were published.

MATERIALS AND METHODS

Sample Collection and Laboratory Analyses:

A total of 277 individuals comprising 60 Konda Reddis, 63 Koya Doras and 29 Konda Kammaras belonging to nearby villages of Rampachodavaram of East Godavari district and 65 Lambadas and 60 Chenchus from Mahaboobnagar district, Andhra Pradesh, were selected for the present study. All individuals were unrelated and the blood samples were collected with the prior informed consent. DNA was isolated using a slightly modified, salting out procedure of Lahiri and Nurnberger (1991).

Using PCR, all the individuals were genotyped for five Alu indel polymorphisms: Alu APO, PV 92, TPA 25, Alu D1 and ACE. The protocols of these markers have been reported earlier (Batzer et al. 1996; Majumder et al. 1999b).

Statistical Analysis: Genotype and allele frequencies were determined by direct count. Hardy-Weinberg Equilibrium (HWE) was tested by calculating exact significance probabilities (analogous to Fisher's exact test for 2 x 2 contingency tables), for avoiding the difficulties encountered in using the chi-square distribution for small samples (Haldane 1954). Measures of

Corresponding address: Prof. Dr. P. Veerraju
Department of Human Genetics, Andhra University,
Visakhapatnam 530 003, Andhra Pradesh, India.
Telephone: 91-0891-284-4726/4888
E-mail: pvraju@rediffmail.com

gene diversity (Nei 1987) were employed in order to quantify the degree of total (H_t), and intrapopulation (H_s) genetic variability, and to estimate the coefficient of gene differentiation (G_{ST}) among populations. To further test significant heterogeneity in the populations, the variation among populations was subjected to a contingency chi-square analysis (Workman and Niswander 1970).

To measure intergroup genetic differences, standard genetic distances (Nei 1972) were calculated. From the distance matrices, neighbor-joining trees (Saitou and Nei 1987) were generated. Genetic relationships among populations were further examined by extracting principal components from the allele frequencies and plotting the average scores of populations onto the first two eigenvectors.

Systematic versus nonsystematic processes that might be in operation among the Indian populations were investigated by employing the Harpending and Ward (1982) model. This model proposes that a linear relationship exists in subdivided populations between mean per locus heterozygosity (H_o) and distance from the gene frequency centroid (r_{ii}). Thus, the relative roles of stochastic processes or nonsystematic pressures (genetic drift or founder effect) and systematic pressures (migration and selection) on the genetic structure of populations may be assessed by regressing r_{ii} against H_o . Subpopulations that have high r_{ii} and low heterozygosity most likely experienced stochastic processes. On the other hand, subpopulations with high H_o and low r_{ii} would have likely experienced high systematic pressure in the form of migration (gene flow) and/or selection. In small populations with reproductive isolation, migration rather than selection is the most likely explanation for the observed high H_o and low r_{ii} values (Crawford 2000).

Spatial Autocorrelation: We also subjected the allele frequencies to a spatial autocorrelation analysis in order to test for positive or negative spatial association of the allele distribution with distance. We used Moran's I product-moment correlation coefficient (Cliff and Ord 1973; Sokal and Oden 1978). One-dimensional correlograms were computed, deriving geographic distances between the localities as great circle distances. The plot of the autocorrelation coefficient I against distance, referred to as a correlogram the overall significance of which is assessed through a Bonferroni test. A spatially random distribution

results in a series of insignificant I values, at all distances. A decreasing set of I coefficients, from a positive significant to a negative significant value, describes a cline, whereas a decreasing correlogram from a significant positive value to insignificant values at large distances is expected for allele frequencies under isolation by distance, i.e., when genetic diversity reflects only genetic drift and short-range gene flow (Barbujani 1987).

Multiresponse Permutation Procedure (MRPP): MRPP (Mielke 1981) is a nonparametric procedure for testing the hypothesis of no differences among two or more groups of entities, equivalent to discriminant analysis or one way MANOVA. Because the probability value of a MRPP statistic is derived through a permutation argument, there are no distributional requirements on the data such as multivariate normality and homogeneity of variances. A permutation is a specific arrangement or assignment of all N objects (in this case population samples) to the specified groups (here grouped by social-ethnic affiliation). The null hypothesis for MRPP states that all the possible permutations are equally likely. The test statistic describes the separation between the groups. The observed δ (the average of the within-group distances) is compared with the expected $\bar{\delta}$, the latter calculated to represent the mean δ for all possible partitions of the data. Small values of δ would indicate a tendency for clustering while larger values of δ would indicate a lack of clustering. The variance and skewness of δ are descriptors of the distribution of all possible values of δ corresponding to the possible partitions of the items. The probability value expresses the likelihood of getting a δ as extreme or more extreme than the observed δ , given the distribution of possible δ .

Population Backgrounds: The present study includes four Proto-Australoid populations namely, Koya Dora, Konda Reddy, Konda Kammara and Chenchu and a Caucasian population Lambada. Konda Reddis live in small hamlets of the hills of East and West Godavari districts. They are in general short in stature, generally medium brownish people with wavy hair. Agriculture is the prime economy of the Konda Reddis. They also depend on nature's produce.

The Koya Doras form a major tribal community of Andhra Pradesh, inhabiting the forests as well as the plain areas of East and West Godavari, Adilabad, Warangal, Khammam and Karimnagar districts. They have immigrated to these areas

about two centuries ago from the neighboring Bastar plateau in Madhya Pradesh because of famines and disputes. They are characterized by dark skin, coarse black hair, flat and broad nose, high cheekbones and relatively thick lips. They depend on agriculture for their living.

The Konda Kammaras are mostly confined to East Godavari district of Andhra Pradesh. The word “Konda Kammaras” means “Hill Blacksmith”. Traditionally they subsisted mainly by making agriculture implements to the neighbouring agricultural tribal populations. Now majority of them have shifted to agriculture. They are generally dark brown in colour with black hair. The hair form is mostly wavy and hair texture is mostly medium. Beard and moustache are sparse in males and most of the men have scanty body hair. Konda Reddies who are economically better placed come first in the social hierarchy followed by Koya Dora and Konda Kammaras.

The Caucasoid Lambadas or Banjari are ethnically different from the other southern Indian tribes. They are mostly found in the Warangal, Khammam, Ananthapur, Mahaboobnagar and Hyderabad districts of Andhra Pradesh. They have hardly any interaction with the other tribes in the state. They are a class of traders, herdsman, cattle breeders and collect wood and jungle produce. The Lambadi language is similar to the Western Marwari and Gujarati languages and belongs to the Indo-Aryan language group. They have long stature, fair skin, well-chiseled face and prominent narrow nose.

The Chenchus or Chentsus are the most primitive Telugu speaking tribe inhabiting the hills of Kurnool, Mahaboobnagar and Guntur districts. They are a semiwild, innocent, inoffensive hill tribe, living on forest products. They are slender and medium in stature. The colour of the skin is dark brown to a rich copper colour. They have a low face with steep forehead, a deep depression at the root of the nose, a strong supra-orbital ridges and flat nose with wide nostrils. The chin is small and protruded. Linguistically they are Dravidian.

RESULTS AND DISCUSSION

The number of chromosomes examined and the frequencies for the insertion alleles are presented in Table 1 separately for the five populations. All the five screened biallelic loci are highly polymorphic in the five populations. Exact test probability values for deviation from Hardy-Weinberg equilibrium, together with observed and expected mean heterozygosities for each population are presented in Table 2. Significant departures from HWE were found in Chenchu and in Konda Reddi for *PV9*, *ACE* and *DI*, in Konda Kammaras for *DI*, and in Lambada for *PV9* and *DI*, in all the cases because of an excess of homozygotes. The observed excess of homozygotes may be attributable to inbreeding because of the practice of close consanguineous marriages among these tribal populations. Chenchu presents, by far, the lowest average

Table 1: Frequencies of Alu insertion markers in five Indian populations

Population	Alu APO		Alu PV92		Alu ACE		Alu TPA25		Alu DI	
	n	+	n	+	n	+	n	+	n	+
Chenchu	110	0.791	106	0.245	114	0.649	118	0.491	106	0.358
Koya Dora	116	0.673	110	0.654	80	0.775	118	0.644	110	0.409
Konda Reddi	108	0.611	106	0.604	106	0.717	114	0.597	100	0.320
Konda Kammaras	52	0.808	52	0.500	58	0.569	58	0.793	52	0.154
Lambada	94	0.672	108	0.454	114	0.465	110	0.500	104	0.385

n = number of chromosomes, + = insertions.

Table 2: Population wise exact test probability values for deviation from Hardy-Weinberg equilibrium, together with observed and expected mean heterozygosity.

Locus	Chenchu	Koya Dora	Konda Reddi	Konda Kammaras	Lambada
APO	0.682	1.000	0.266	0.545	0.701
PV92	0.001	0.383	0.009	0.056	0.002
ACE	0.022	1.000	0.015	0.453	0.598
TPA25	0.192	0.052	0.588	0.304	0.184
DI	0.000	0.151	0.000	0.000	0.003
Heterozygosity Observed	0.282	0.384	0.370	0.338	0.423
Expected	0.423	0.437	0.455	0.378	0.481

heterozygosity for the studied loci as well as the highest departure from the expected value.

Genetic Diversity Analysis: To determine the amount of genetic differentiation among populations, G_{ST} values (a measure of the interpopulation variability) for each polymorphic locus were determined. The results of gene diversity analyses are presented in Table 3, separately for each locus as well as the average. The total genetic diversity (H_T) among subpopulations is fairly high. However, most of the diversity corresponds to diversity between individuals within populations (H_S). The percentage of genetic diversity attributable to between populations relative to the total genomic diversity (G_{ST}) ranges from 0.028 for *APO* to 0.081 for *PV9*. On average, 5.1% of the total genetic diversity is attributable to between population differences. Contingency chi-square analysis (Workman and Niswander 1970) reveals significant differences ($p \leq 0.05$) among populations for all the five loci investigated, the overall value being highly significant (chi-square 112.076, d.f. 20, $P=0.000$).

Genetic Affinities among Populations: In order to assess the relationships between the populations, Nei's standard genetic distances

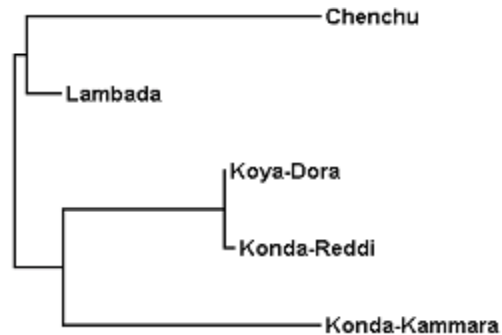


Fig. 1. Neighbor Joining tree of Nei's standard genetic distances based on allele frequencies for 5 Alu insertion markers

between pairs of populations were calculated and in the resulting neighbour joining tree Koya Dora and Konda Reddi clustered closely together to which the Konda Kammara joined as outer element. The Lambada and Chenchu constitute a second cluster. We have also examined affinities among populations using a different statistical approach, by extracting principal components of allele frequencies and plotting the scores of the populations on the first two principal components (Fig. 2), which portrays more distinct

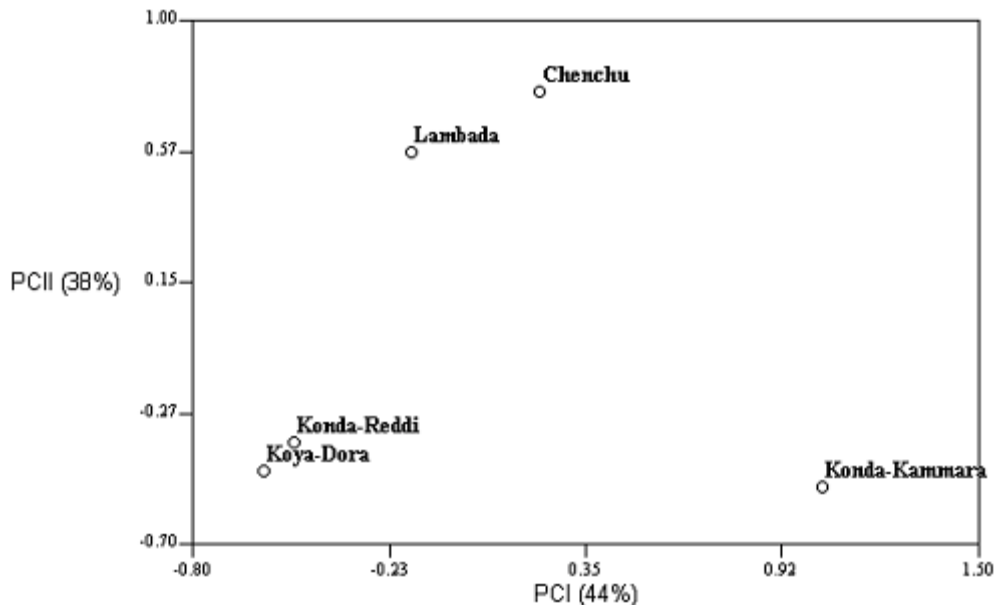


Fig. 2. Principal Component plot of the 5 Indian tribal populations based on the first two eigenvectors of allele frequencies of Alu insertion markers.

Table 3: Locus wise and average gene diversity values for the studied populations

Locus	H_r	H_s	G_{ST}
APO	0.411	0.399	0.028
PV92	0.499	0.459	0.081
ACE	0.464	0.440	0.052
TPA25	0.478	0.454	0.051
DI	0.439	0.422	0.037
All loci	0.458	0.435	0.051

Table 4: Results from the Multiresponse Permutation Procedure depicting the genetic conformity of the underlying socioeconomic stratification of the populations.

Group	N	Mean Distance
Upper Caste	2	0.282
Middle Caste	4	0.379
Lower Caste	4	0.347
Tribal	14	0.441
Observed δ		0.402
Expected δ		0.415
Significance		P=0.173

separation of Konda Kammara from the other tribes, but broadly conforming to the same pattern of population relationships in Figure 1.

The neighbor-joining tree presented in Figure 3 and the Principal Component Analysis plot of Figure 4 depict the genetic relationships among 24 Indian populations based on the same 5 Alu markers. The comparative data for the above analyses were gathered from Vishwanathan et al. (2003) and Majumder et al. (1999b). Broadly speaking the population clusters seem to have been based on geographical affiliation of the populations. However, the tribal populations of Andhra Pradesh do not form a single cluster but segregate into two broad clusters formed by populations of UP (Lambada, Chenchu and Konda Kammara) and Southern and Eastern Indian populations (Konda Reddy and Koya Dora). Given the ethnic /linguistic background of Lambadas, its clustering with the northern groups is expected. The unexpected position of Chenchu and Konda Kammara in the plot is somewhat intriguing.

Gene Flow among Populations: Figure 5 presents the plot of observed heterozygosities of the five populations against the distance from the gene frequency centroid along with the theoretical linear regression line. Three populations are in close agreement with the model: Konda Reddi presents the lowest distance (r_{ii}) to the hypothetical ancestral population, maintaining a fairly high heterozygosity. Koya

Dora presents intermediate values for both r_{ii} and H_o , falling exactly onto the theoretical regression line, whereas Chenchu shows high r_{ii} and a lower H_o . The two outliers to the model are Lambada with unexpectedly high heterozygosity (suggesting strong external gene flow), and Konda Kammara, with the highest r_{ii} and the lowest H_o value, which suggests the action of genetic drift.

In Figure 6 we present the same analysis but including data from other Indian populations for the same loci (Vishwanathan et al. 2003 and Majumder et al. 1999b). Most of the populations are positioned close to the theoretical regression line presenting a H_o value between 0.42 and 0.46 and fairly low values of distances from the centroid. A second group (Irula, Tanti, Gaud, Kurumba, and Lodha), fall well above the regression line and far from the hypothetical ancestral population, suggesting more than average external gene flow into these populations. On the other hand, Muslims, Tipperah, Konda Kammara, and particularly Toda, fall below the regression line and also at high values of r_{ii} , suggesting the action of genetic drift. The position of Munda, with by far the highest r_{ii} (0.217 against a means value of $F_{ST}=0.082$, which is already high) and intermediate H_o , suggests that this population could belong to a different genetic origin.

Spatial Autocorrelation: Spatial correlogram is presented in Figure 7 with the horizontal axis containing geographic distance between populations (in kilometers) and the vertical axis standardized (z-score) value of Moran's I at each spatial lag. Moran's I values are plotted at the upper distance limit for the lags. The values of Moran's I across the different lags fluctuate around 0 and are not statistically significant, except for APO and PV92, which present significant negative values at intermediate distances. The overall results for the Bonferroni test also indicate that there is no significant departures from randomness in any of the 5 Alu insertion systems, i.e., there is a lack of spatial structure in the allele variation. This result somewhat contrasts with previous reports concerning the Indian subcontinent, which suggest clinal patterns at least for short range distances both for palmar dermatoglyphics (Reddy et al. 2004) and anthropometric and genetic markers (Reddy et al. 2001).

MRPP: The results of the MRPP are

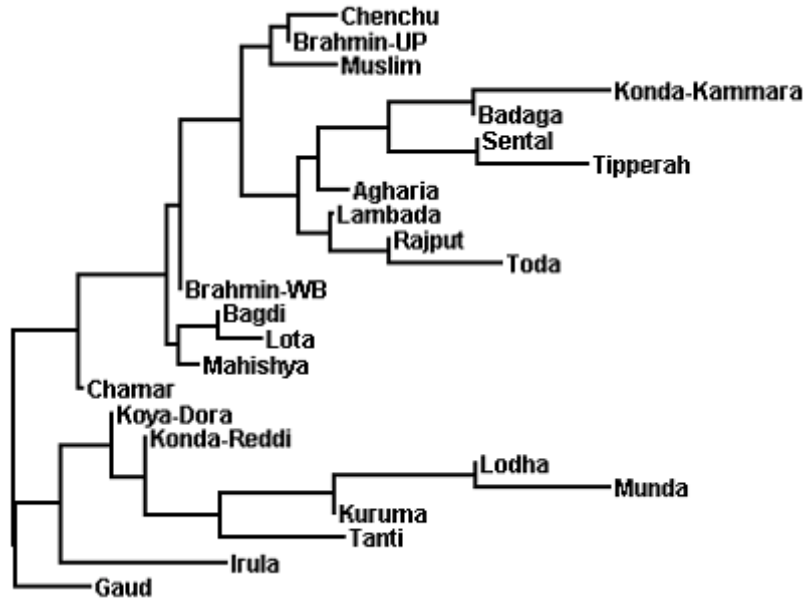


Fig. 3. Neighbor joining tree derived from Nei's genetic distances, based on five Alu insertions among the 24 Indian populations

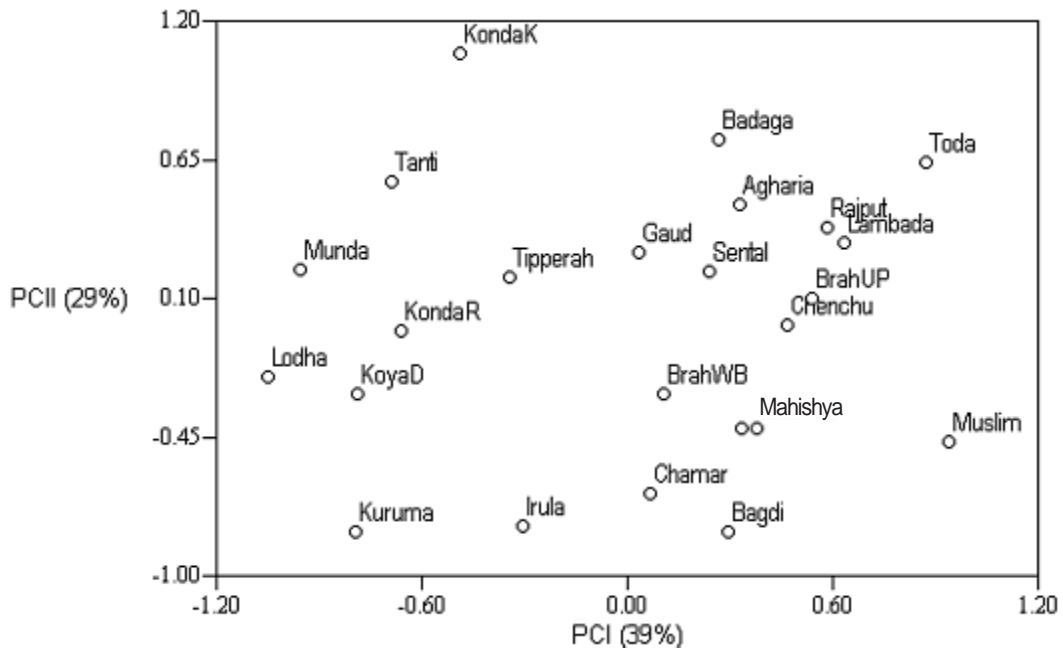


Fig.4. Principal Components plot of the 24 Indian populations based on five Alu insertions.

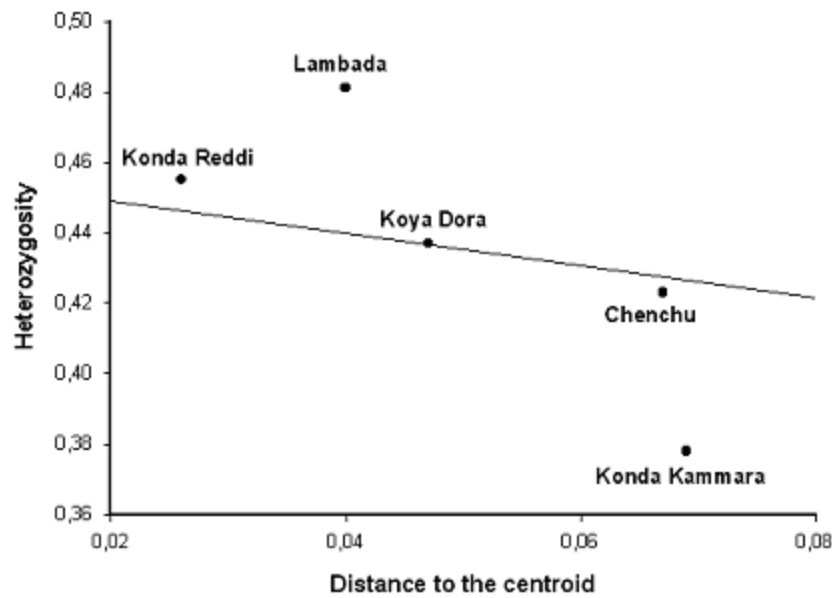


Fig. 5. Plot of average heterozygosity versus the distance from the gene frequency centroid of the 5 tribal populations

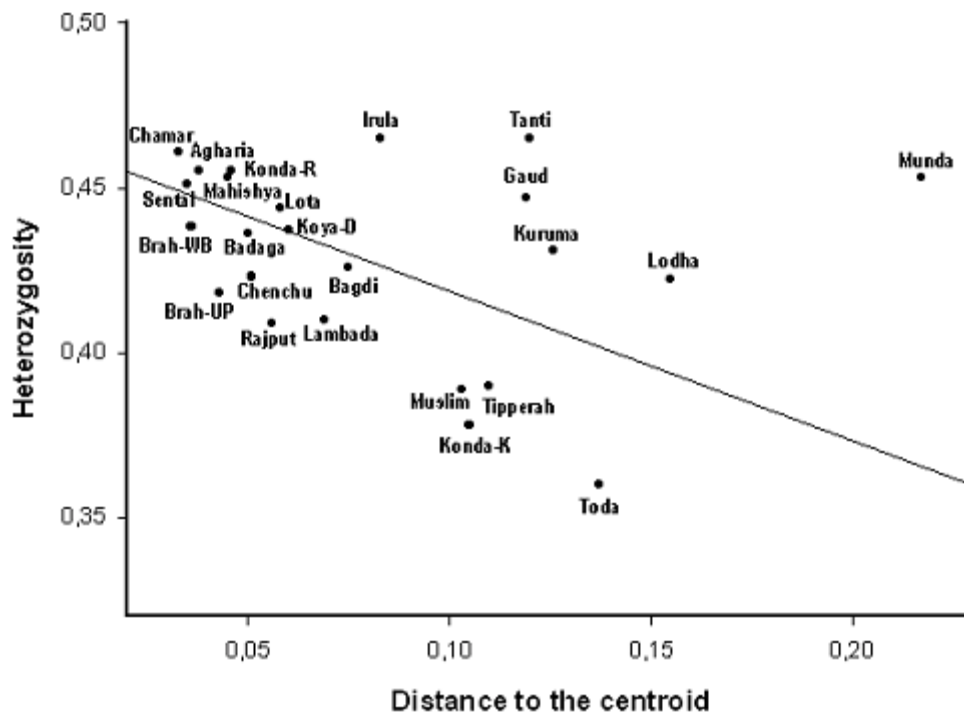


Fig. 6. Plot of average heterozygosity versus the distance from the gene frequency centroid of the 5 tribal populations

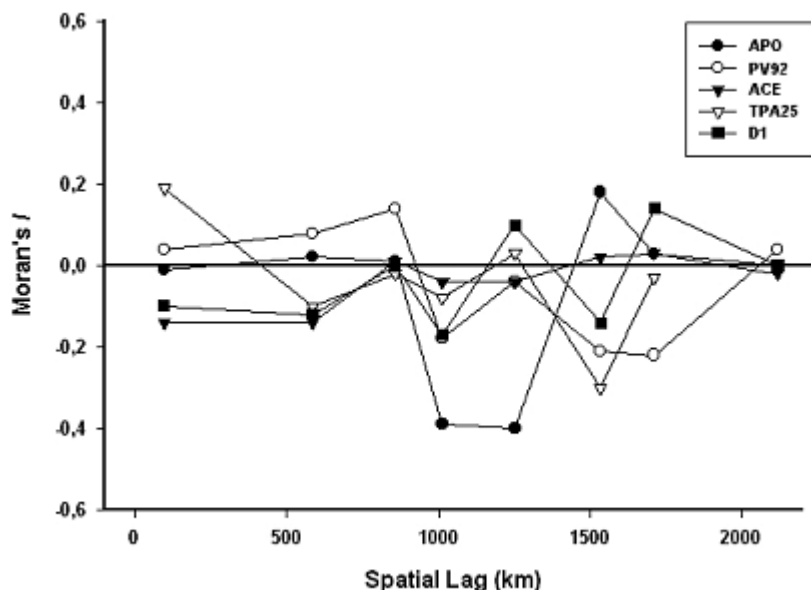


Fig. 7. Correlogram of Moran's I versus spatial lags at 250 kilometers

summarized in Table 4. The overall result suggests a lack of association between social-economic affiliation and Alu insertions allele frequency distribution pattern, since the observed δ , although slightly lower, does not differ statistically from the expected δ (the distance expected for groups of populations randomly generated). A more detailed observation, however, reveals a strong clustering pattern for the 3 castes groups (principally for the upper caste), since within-group distances are much lower than those expected for a random distribution. In contrast, the group that includes the tribal populations shows a complete lack of clustering, as the within-group distance well above the value expected for a random distribution. This suggests a much greater heterogeneity among the tribal populations when compared to the castes.

CONCLUSIONS

The five biallelic markers analysed in this study show high degree of variation among the five tribal populations studied, which is reflected in the clear separation/clustering of populations in Figure 1. Although there is tendency of populations to form clusters based on broad

geographic affiliations when the 24 Indian populations were subject to cluster analysis, the tribes of Andhra Pradesh fail to form a compact cluster. They are segregated into two broad clusters, one formed by North Indian populations, and the other formed by the Eastern (Austro-Asiatic) and Southern Indian tribal populations.

Given that the Indian population do not conform to the isolation by distance model of population structure and despite geographic proximity they behave like islands due to endogamy, it is not surprising that the Moran's I do not show monotonic decline with the increase in spatial lags which are as large as 250 kilometers. Average marriage distance in India and particularly in the rural/tribal areas of southern India is very small and a large proportion of marriages occur within the village and/or within the small radius (average < 10 kms). Further, most of the Indian caste and tribal populations still strictly adhere to the norms of endogamy, especially in the rural areas. Therefore, whatever trends expected in autocorrelation can be probably seen only at small geographic distances, not at the state or national level. The geographic representation of the sampled

populations in this study which is too disjoint and too few in number (24) which may also contribute significantly to the observed pattern in autocorrelation in this study.

ACKNOWLEDGEMENTS

Financial support for this work was provided by Department of Biotechnology, Government of India. We thank Prof. B.M. Reddy, Indian Statistical Institute, Hyderabad for his critical inputs, which helped in improved presentation of the manuscript. DA Demarchi is Investigator Career Member of the Consejo de Investigaciones Científicas y Técnicas de la República Argentina (CONICET).

REFERENCES

- Bamshad MJ, Farley A, Crawford MH, Cann RL, Busi BR, Naidu JM, Jorde LB 1996. Mt. DNA variation in caste populations of Andhra Pradesh, India. *Hum Biol*, **68**: 1-28.
- Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM et al. 1998. Female gene flow stratifies Hindu castes. *Nature*, **395**: 651-652.
- Barbujani G 1987. Autocorrelation of gene frequencies under isolation by distance. *Genetics*, **117**: 777-782.
- Batzer MA, Arcot SS, Phinney JW, Algeria-Hartman N, Kass DH, Milligan SM, Kimpton C et al. 1996. Genetic variation of recent Alu insertions in human populations. *J Mol Evol*, **42**: 22-29.
- Bhasin MK, Walter H 2001. *Genetics of Castes and Tribes of India*. Delhi: Kamla-Raj Enterprises.
- Cliff AD and Ord JK 1973. *Spatial Autocorrelation*. London: Pion.
- Crawford MH 2000. Genetic structure of Mennonite populations. In: MH Crawford (Ed.): *Different Seasons: Biological Aging of Mennonites of the Midwestern United States*. Publications in Anthropology Series 21, Lawrence: University of Kansas, pp.31-40.
- Haldane J 1954. An exact test for randomness of mating. *Journal of Genetics*, **52**: 631-635.
- Harpending H, Ward RH 1982. Chemical systematics and human populations. In: MH Niteeki (Ed.): *Biological Aspects of Evolutionary Biology*. Chicago, IL: University of Chicago Press, pp. 213-256.
- Lahiri DK, Nurnberger JI 1991. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nuc Acids Res*, **19**: 5444.
- Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, Thangaraj K, Singh L, Reddy BM (2006) Global Patterns in Human Mitochondrial DNA and Y-Chromosome Variation Caused by Spatial Instability of the Local Cultural Processes. *PLoS Genet*, **14**: e53.
- Kumar V, Reddy ANS, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L and Reddy BM 2007. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evolutionary Biology* **7**: 47.
- Majumder PP 1999a. Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y – chromosomal DNA polymorphisms. *Genome Res*, **9**: 711-719.
- Majumder PP, Roy B, Banerjee S, Chakraborty M, Dey B, Mukherjee N, Roy M, Thakurta PG and Sil SK 1999b. Human-specific insertion/deletion polymorphisms in Indian populations and their evolutionary implications. *Eur J Hum Genet*, **7**: 435-446.
- Mielke PK, Berry P, Brockwell and Williams J 1981. A class of non parametric tests based on multiresponse permutation procedures. *Biometrika*, **68**: 720-724.
- Nei M 1972. Genetic distance between populations. *Amer Naturalist*, **106**: 283-292.
- Nei M 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press, New York.
- Reddy BM, Demarchi DA, Malhotra KC 2001. Patterns of Biological Variation among the 20 endogamous groups of Dhargar caste-cluster of Maharashtra, India. *Collegium Anthropologicum*, **25**: 425-442.
- Reddy BM, Demarchi DA, Bharati S, Kumar V, Crawford MH 2004. Patterns of ethnic, linguistic and geographic heterogeneity of palmar interdental ridge counts in the Indian subcontinent. *Hum Biol*, **76**: 211-228.
- Reddy BM, Naidu VM, Madhavi VK, Thangaraj K, Kumar V, Langstieh BT, Venkatramana PV, Reddy AG, Singh L 2005. Microsatellite diversity in Andhra Pradesh, India: Genetic stratification versus social stratification. *Human Biology*, **77**: 803-823.
- Saitou N, Nei M 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**: 406-425.
- Sokal RR, Oden NL 1978. Spatial autocorrelation biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**: 229-249.
- Vishwanathan H, Edwin D, Usharani MV and Majumder PP 2003. Insertion/deletion polymorphisms in tribal populations of Southern India and their possible evolutionary implications. *Hum Biol*, **75**: 873-887.
- Workman PL and Niswander JD 1970. Population studies on southern Indian tribes. II. Local genetic differentiation in the Papago. *Am J Hum Genet*, **22**: 24-49.