European Journal of Human Genetics (1999) 7, 435–446

http://www.stockton-press.co.uk/ejhg

ARTICLE

Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications

Partha P Majumder[1], Bidyut Roy[1], Sanat Banerjee[1], Madan Chakraborty[1], Badal Dey[1], Namita Mukherjee[1], Monami Roy[1], Piyali Guha Thakurta[1,2] and Samir K Sil[3]

[1]*Anthropology and Human Genetics Unit, Indian Statistical Institute*
[2]*Department of Biophysics, Molecular Biology and Genetics, University of Calcutta, Calcutta*
[3]*Department of Life Sciences, Tripura University, Agartala, India*

DNA samples from 396 unrelated individuals belonging to 14 ethnic populations of India, inhabiting various geographical locations and occupying various positions in the socio-cultural hierarchy, were analysed in respect of 8 human-specific polymorphic insertion/deletion loci. All loci, except *Alu* CD4, were found to be highly polymorphic in all populations. The levels of average heterozygosities were found to be very high in all populations and, in most populations, also higher than those predicted by the island model of population structure. The coefficient of gene differentiation among Indian populations was found to be higher than populations in most other global regions, except Africa. These results are discussed in the light of two possible scenarios of evolution of Indian populations in the broader context of human evolution.

Keywords: *Alu* insertion; mtDNA; genome diversity; human evolution

Introduction

Considerable insight into the peopling of India has been derived from past studies on genetic diversities and affinities among ethnic groups of India.[1] The vast majority of these studies was based on blood group, serum protein and red-cell enzyme polymorphisms. The levels of polymorphism at loci that code for expressed proteins and enzymes are generally low because mutations at these loci are commonly deleterious and, therefore, are often strongly selected against. On the other hand, DNA polymorphisms, especially in the non-coding regions of the human genome, are expected to be selectively neutral. Polymorphic DNA markers have, therefore, proved to be immensely useful in studies of human diversity and evolution. In recent years, several insertion/deletion polymorphisms have been discovered in the human genome which are particularly useful in human population genetic studies because

(i) the ancestral states of these polymorphisms are known since these elements are inserted/deleted at random into the nuclear genome but are never precisely deleted/inserted, thereby facilitating accurate rooting of population networks, and

(ii) all alleles of a particular type are identical by descent since the probability of two insertions/deletions of these elements in/from the same genomic location is vanishingly small.

Correspondence: Professor Partha P Majumder, Anthropology and Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India. Tel: +91 33 577 8085; Fax: +91 33 577 6680; E-mail: ppm@isical.ac.in
Received 9 June 1998; revised 27 November 1998; accepted 16 December 1998

We have studied eight such insertion/deletion polymorphisms in 14 ethnic populations of India. These populations belong to three of the four major linguistic groups present in India. The populations are at different levels of modernisation and socio-cultural hierarchy. The overall objective of this study was to shed light on the peopling of India and on human evolution.

A segment of the human mitochondrial DNA (mtDNA) was discovered[2] to have been inserted in the human nuclear genome. The inserted region comprises 540 bp of the control region of mtDNA, corresponding to nucleotide positions 59–16089 of the Cambridge reference sequence.[3] This insertion (denoted as mt → NUC) is human specific; the event most likely occurred after the separation of the chimpanzees and humans, but before the divergence of human populations. We have studied this locus, which is currently polymorphic in many populations.

The *Alu* family of short interspersed elements (SINEs) is found in about 500 000 copies in multiple chromosomal locations in the human genome.[4] *Alu* sequences are thought to be ancestrally derived from the 7SL RNA gene[5] and mobilise through an RNA polymerase III-derived transcript in a process termed 'retroposition'.[6] Although *Alu* sequences are also found in many mammalian genomes in addition to the human, a subset of these insertions are human specific.[7] Some of these human-specific *Alu* elements have retroposed so recently that they have not become fixed within the human genome.[7] Therefore, human populations are polymorphic with respect to these insertions. We have, in this study, examined six of these human-specific *Alu* polymorphisms.

The eighth polymorphic locus we have studied is the deletion of 256-bp of a 285-bp *Alu* element at the CD4 locus.[8] The lack of this deletion [*Alu*(+)] is the ancestral state of this polymorphism; chimpanzees, gorillas, orangutans and gibbons are monomorphic for the *Alu*(+) allele.[9]

Materials and Methods

Populations Studied

We have studied a total of 792 chromosomes from unrelated individuals belonging to 14 endogamous population groups of India. Of these groups four are tribal, speaking Austro-Asiatic (three groups) or Tibeto-Burman (one group) languages; nine are caste groups at different levels of social hierarchy (upper, middle and lower) speaking Indo-European languages, and one religious group of Indo-European (Hindi) speaking Muslims. The samples were drawn from single geographical locations from the States of West Bengal, Orissa, Tripura and Uttar Pradesh. Further details are provided in Table 1.

Laboratory Analysis

Blood samples (5–10 ml by venipuncture in EDTA) were drawn from individuals with prior informed consent. DNA was isolated using a standard protocol.[10] Oligonucleotide primers used in PCR amplifications of the eight loci along with corresponding annealing temperatures are given in Table 2. For the mt → NUC locus, the reaction mixture for amplification comprised 2.5 U Taq DNA polymerase, 50 ng each primer, 400 ng genomic DNA, in a total volume of 20 µl. PCR cycling temperature protocol was: 30 cycles × (94°C for 15 s, 63°C for 30 s, 72°C for 1 min). For the *Alu* insertion loci, the reaction mixture comprised 2.5 U Taq DNA polymerase, 50 ng each primer, 300 ng genomic DNA, in a total volume of 20 µl. PCR cycling temperature protocol was: 30 cycles × (94°C for 1 min, x°C (see Table 2 for the value of x) for 2 min, 72°C for 2 min). For the *Alu* CD4 locus, the reaction mixture comprised 2.5 U Taq DNA polymerase, 50 ng each primer, 200 ng genomic DNA, in a total reaction volume of 20 µl. PCR cycling temperature protocol was: 30 cycles × (94°C for 1 min, 58°C for 1.5 min, 72°C for 1.5 min).

Statistical Analysis

Maximum likelihood estimates of allele frequencies and their standard errors were calculated at each locus separately for each population. Heterozygosities at individual loci and the average heterozygosity were calculated using the estimated allele frequencies for each population. To assess the extent of gene differentiation among the population groups, gene diversity analysis[11] was performed separately for each locus as also for all loci considered jointly. To assess genomic relationships among the populations, dendrograms were constructed by three different methods – UPGMA, neighbour-joining (NJ) and maximum likelihood (ML) – using appropriate subroutines in PHYLIP 3.5c.[12] The D_A measure of genetic distance[13] was used. Genomic relationships among populations were also examined by extracting principal components of allele frequencies and plotting the positions of the populations using the values of the first three principal components.

To assess the relative amount of gene flow experienced by each population, we have used a regression model originally proposed by Harpending and Ward.[14] For this, the heterozygosity of the i^{th} population was plotted against the distance of the population from the centroid (r_i), calculated as:

$$r_i = (p_i - P)^2/[P(1 - P)], \quad (1)$$

where p_i and P are, respectively, the frequency of the insertion (deletion) allele in population i and the total population. Under the island model of population structure, Harpending and Ward[14] have shown that there should exist a linear relationship between heterozygosity and distance from the centroid:

$$h_i = H(1 - r_i), \quad (2)$$

where h_i and H denote, respectively, the heterozygosities of population i and the total population. Of particular interest in this analysis are the outliers: populations that have experienced more gene flow than average will have higher

Table 1 Study populations, locations of sampling, approximate population sizes in the area of sampling (*N*) and anthropological information

Population name (code)	*Location of sampling*	*N*	*Anthropological information*
Agharia (AG)	Sundergarh Dist., Orissa	100 000	Hindu Middle caste; Indo-European language; primarily agriculturists
Bagdi (BA)	Hooghly Dist., West Bengal	80 000	Hindu Low caste; Indo-European language; primary occupations cultivation and fishing; usually admit members of any caste higher than themselves in social ranking
Brahmin-UP (BR-UP)	Garhwal, Uttar Pradesh	150 000	Hindu Upper caste; Indo-European language; traditionally priests, but now various occupations
Brahmin-WB (BR-WB)	Various locations, West Bengal	200 000	Hindu Upper caste; Indo-European language; traditionally priests, but now various occupations
Chamar (CH)	Garhwal, Uttar Pradesh	70 000	Hindu Low caste; Indo-European language; leather workers
Gaud (GA)	Sundergarh Dist., Orissa	150 000	Hindu Middle caste; Indo-European language; primarily agriculturists
Lodha (LO)	Medinipur, West Bengal	25 000	Tribe; Austro-Asiatic language; primarily agricultural labourers; numerically small; geographically isolated
Mahishya (MA)	Hooghly Dist., West Bengal	150 000	Hindu Low caste; Indo-European language; primarily agricultural labourers
Munda (MD)	Sundergarh Dist., Orissa; Medinipur, West Bengal	80 000	Tribe; Austro-Asiatic language; primarily agricultural labourers; numerically large; wide geographical distribution
Muslim (MU)	Garhwal, Uttar Pradesh	150 000	Religious group; Indo-European language; many are religious converts from lower social groups
Rajput (RA)	Garhwal, Uttar Pradesh	100 000	Hindu Middle caste; Indo-European language; wide georgraphical distribution
Santal (SA)	Medinipur, West Bengal	10 000	Tribe; Austro-Asiatic language; numerically large; wide geographical distribution; claims of relationship to Mundas in folklore
Tanti (TA)	Sundergarh Dist., Orissa	40 000	Hindu Low caste; Indo-European language; traditionally weavers
Tipperah — also known as Tripuri (TI)	Various locations around Agartala, Tripura	90 000	Tribe; Tibeto-Burman language; part of Bodo ethnic stock who migrated from Tibet several centuries ago; agriculturists

Table 2 Oligonucleotide primers and annealing temperatures of the loci studied

Locus	*Primer sequences*	*Annealing temperature (°C)*	*Reference*
mt→NUC	5′– ACA AAG TCC AGG TTT CTA ACA G – 3′ 5′– AGT CTT GCT TAT TAC AAT GAT GG – 3′	63	2
Alu FXIIIB	5′– TCA ACT CCA TGA GAT TTT CAG AAG T – 3′ 5′– CTG GAA AAA ATG TAT TCA GGT GAG T – 3′	56	15
Alu D1	5′– TGC TGA TGC CCA GGG TTA GTA AA – 3′ 5′– TTT CTG CTA TGC TCT TCC CTC TC – 3′	70	15
Alu APO	5′– AAG TGC TGT AGG CCA TTT AGA TTA G – 3′ 5′– AGT CTT CGA TGA CAG CGT ATA CAG A – 3′	50	15
Alu TPA25	5′– GTA AGA GTT CCG TAA CAG GAC AGC T – 3′ 5′– CCC CAC CCT AGG AGA ACT TCT CTT T – 3′	58	15
Alu ACE	5′– CTG GAG ACC ACT CCC ATC CTT TCT – 3′ 5′– GAT GTG GCC ATC ACA TTC GTC AGA T – 3′	58	15
Alu PV92	5′– AAC TGG GAA AAT TTG AAG AGA AAG T – 3′ 5′– TGA GTT CTC AAC TCC TGT GTG TTA G – 3′	54	15
Alu CD4	5′– AGG CCT TGT AGG GTT GGT CTG ATA – 3′ 5′– TGC AGC TGC TGA GTG AAA GAA CTG – 3′	58	8

heterozygosities than predicted, whilst those that have experienced less gene flow than average will have lower heterozygosities than predicted.

Results

Allele Frequencies and Genomic Diversity within Populations

The numbers of chromosomes examined and allele frequencies for the ancestral-state alleles [insertion (+) for mt→NUC and *Alu* FXIIIB, D1, APO, TPA25, ACE, PV92 loci; deletion (–) for *Alu* CD4 locus] are given in Table 3 separately for the 14 populations. It is seen that all loci, except *Alu* CD4, are highly polymorphic in most populations. *Alu* CD4 exhibits low levels of polymorphism in most populations; in fact, the deletion (–) allele is absent among Chamars and Mundas.

The heterozygosities at each locus and the average heterozygosity over all the eight loci are given in Table 4 separately for each population. It is seen that most populations show very high levels of diversity with respect to most of the loci; the heterozygosity at the *Alu* CD4 locus is low and consistently the minimum in all populations. It is noteworthy that in many cases the maximum attainable value (0.5) of heterozygosity for a biallelic marker is actually attained. The average heterozygosity ranges from 0.351 (Tipperah) to 0.449 (Mahishya).

Genomic Diversity between Populations

The results of gene diversity analysis are presented in Table 5, separately for each locus as also for all loci taken together. It is seen that except for the *Alu* CD4 locus, the total genomic diversity (H_T) among the subpopulations is quite high. However, most of the genomic diversity is attributable to diversity between individuals within populations (H_S). The percentage of genomic diversity attributable to between populations relative to the total genomic diversity, G_{ST}, varies

Table 3 Allele frequencies at eight polymorphic loci in 14 ethnic populations of India

Population	*mt* →NUC		*Alu* FXIIIB		*Alu* D1		*Alu* APO		*Alu* TPA25		*Alu* ACE		*Alu* PV92		*Alu* CD4	
name	*n*	+	*n*	+	*n*	+	*n*	+	*n*	+	*n*	+	*n*	+	*n*	–
Agharia	46	0.500	46	0.348	48	0.417	44	0.795	46	0.587	48	0.417	46	0.435	48	0.063
Bagdi	62	0.452	62	0.532	62	0.645	62	0.855	62	0.484	62	0.694	62	0.468	62	0.065
Brahmin (UP)	54	0.444	54	0.593	54	0.370	54	0.889	50	0.500	54	0.593	54	0.333	52	0.115
Brahmin (WB)	46	0.478	46	0.609	46	0.521	46	0.869	44	0.545	46	0.609	46	0.565	46	0.152
Chamar	36	0.694	50	0.780	50	0.500	50	0.720	46	0.413	50	0.700	50	0.540	46	0.000
Gaud	28	0.500	30	0.733	30	0.200	30	0.500	30	0.433	30	0.600	30	0.333	26	0.038
Lodha	64	0.452	62	0.823	62	0.281	64	0.453	64	0.625	64	0.859	64	0.532	64	0.016
Mahishya	68	0.500	68	0.662	68	0.588	68	0.824	66	0.485	68	0.559	68	0.515	68	0.162
Munda	46	0.587	48	0.667	50	0.320	50	0.300	50	0.660	50	0.640	52	0.481	52	0.000
Muslim	50	0.560	54	0.500	56	0.464	56	0.946	52	0.346	56	0.643	54	0.315	54	0.037
Rajput	98	0.490	98	0.704	104	0.307	104	0.902	102	0.510	104	0.538	104	0.337	102	0.020
Santal	46	0.478	40	0.725	48	0.292	46	0.761	48	0.417	48	0.521	48	0.563	40	0.025
Tanti	28	0.535	30	0.767	32	0.406	30	0.533	32	0.718	30	0.433	32	0.656	32	0.031
Tipperah	74	0.459	78	0.846	80	0.313	80	0.863	82	0.549	78	0.590	74	0.811	82	0.012

n = number of chromosomes; + = insertion; – = deletion.

Table 4 Heterozygosities at individual loci and average heterozygosity based on 8 loci in each of 14 ethnic populations of India

Population name	*mt→NUC*	*Alu* FXIIIB	*Alu* D1	*Alu* APO	*Alu* TPA25	*Alu* ACE	*Alu* PV92	*Alu* CD4	*All loci* (S.E.)
Agharia	0.500	0.454	0.486	0.326	0.485	0.486	0.492	0.118	0.437 (0.049)
Bagdi	0.495	0.498	0.458	0.248	0.499	0.425	0.498	0.122	0.419 (0.052)
Brahmin (UP)	0.494	0.483	0.466	0.197	0.500	0.483	0.444	0.204	0.425 (0.048)
Brahmin (WB)	0.499	0.476	0.499	0.228	0.496	0.476	0.492	0.258	0.447 (0.042)
Chamar	0.425	0.343	0.500	0.403	0.485	0.420	0.497	0.000	0.401 (0.061)
Gaud	0.500	0.391	0.320	0.500	0.491	0.480	0.444	0.073	0.429 (0.055)
Lodha	0.495	0.291	0.404	0.496	0.469	0.242	0.498	0.031	0.378 (0.061)
Mahishya	0.500	0.448	0.485	0.290	0.500	0.493	0.500	0.272	0.449 (0.035)
Munda	0.485	0.444	0.435	0.420	0.449	0.461	0.499	0.000	0.416 (0.060)
Muslim	0.493	0.500	0.497	0.102	0.453	0.459	0.432	0.071	0.390 (0.066)
Rajput	0.500	0.417	0.426	0.177	0.500	0.497	0.447	0.039	0.383 (0.062)
Santal	0.499	0.399	0.413	0.364	0.486	0.499	0.492	0.049	0.418 (0.056)
Tanti	0.498	0.357	0.482	0.498	0.405	0.491	0.451	0.060	0.434 (0.056)
Tipperah	0.497	0.261	0.430	0.236	0.495	0.484	0.307	0.024	0.351 (0.061)

Table 5 Results of gene diversity analysis for individual loci and for all loci considered jointly

Locus	H_T	H_S	G_{ST}
mt→NUC	0.500	0.491	0.017
Alu FXIIIB	0.446	0.411	0.078
Alu D1	0.450	0.481	0.063
Alu APO	0.395	0.320	0.189
Alu TPA25	0.499	0.479	0.040
Alu ACE	0.457	0.480	0.048
Alu PV92	0.500	0.464	0.072
Alu CD4	0.100	0.094	0.053
All loci	0.425	0.396	0.068

between 1.7% (mt → NUC) and 18.9% (*Alu* APO). When all loci are jointly considered, 6.8% of the total genomic diversity is attributable to between populations.

Genomic Affinities among Populations

The affinities among the 14 populations, reconstructed using the neighbour-joining method is depicted in Figure 1 using allele frequency data of all the eight loci. The maximum-likelihood tree, based upon an examination of 379 trees, is not presented because its topology agreed with that of the neighbour-joining tree. It is seen that the affinities among the caste populations do not correlate well with their socio-cultural affiliation. Instead, populations that occupy closer geographical habitat show, by and large, closer genomic affinity. For example, the upper caste Brahmin groups sampled from distant geographical regions of Uttar Pradesh and West Bengal do not show close genomic similarity. Instead, the Brahmins of West Bengal are genetically close to low caste populations (Mahishya and Bagdi) who reside in close geographical proximity. Similarly, the Brahmins of Uttar Pradesh show close genomic affinities with two other populations – Rajputs (middle caste) and Muslims (religious group) – inhabiting contiguous geographical regions. It is, however, noteworthy that the other population – Chamar (low caste) – sampled from Uttar Pradesh is genetically quite distant from the Uttar Pradesh Brahmins, Rajputs and Muslims.

Of the four tribal populations included in this study, three (Santal, Lodha and Munda) are linguistically

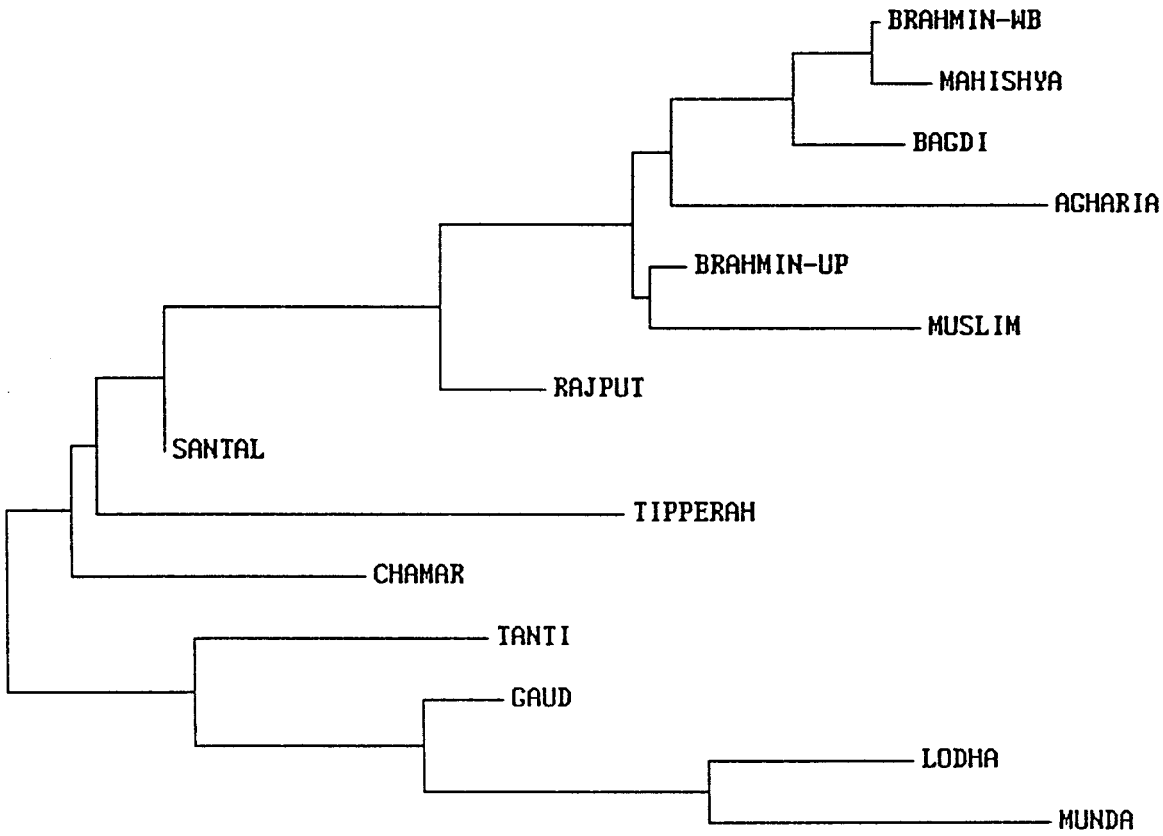

Figure 1 *Unrooted neighbour-joining tree depicting genomic affinities among 14 Indian ethnic populations based on eight human-specific insertion/deletion polymorphisms.*

Austro-Asiatic, whilst the fourth (Tipperah) is Tibeto-Burman (Sino-Indian). As is therefore expected, the Tipperahs stand out genetically, whilst the Lodhas and Mundas show close genomic affinities. In fact, the Lodhas and Mundas form a clear and distinct cluster as is evident from Figure 1. This cluster also includes another middle caste group (Gaud) who occupy overlapping geographical habitat with the Lodhas and Mundas. A notable exception is the Austro-Asiatic tribal group of Santals, who do not belong to this cluster. The Tibeto-Burman speaking Tipperahs seem to cluster with the Santals, although as is evident from Figure 1 they are genetically quite distant. The UPGMA and ML trees showed no significant topological dissimilarities with the NJ tree; therefore, these trees are not presented.

We have also examined affinities among the populations using a different statistical approach. This was done because each approach has its own limitation, and congruence of inferences using multiple approaches strengthens the overall conclusion. We have extracted principal components of allele frequencies, and have plotted the positions of the populations with respect to the first three principal components (Figure 2). As is seen from Figure 2, the first principal component, which explains about 40% of the variation in allele frequencies, broadly separates the tribal from the non-tribal populations. The relationships among the populations are also largely in conformity with those observed from the cluster analyses (UPGMA, NJ and ML).

To determine the genetic relationships of these ethnic populations of India with populations of other regions of the world, we have used the data on six *Alu* insertion loci presented in Stoneking *et al*[15] that are common with our study. There loci are ACE, TPA25, PV92, APO, FXIIIB and D1. The NJ tree of 45 global populations, including the 14 populations of the present study, is presented in Figure 3. (We have excluded data of the three ethnically and geographically ill-defined Indian population groups – Indian Christian, Hindu and Muslim – given in Stoneking *et al*.[15]) It is seen that, by and large, the Indian populations lie between the Mongoloid population groups (China, Filipino, Malayasian, etc) and Caucasoid population groups (Greeks,

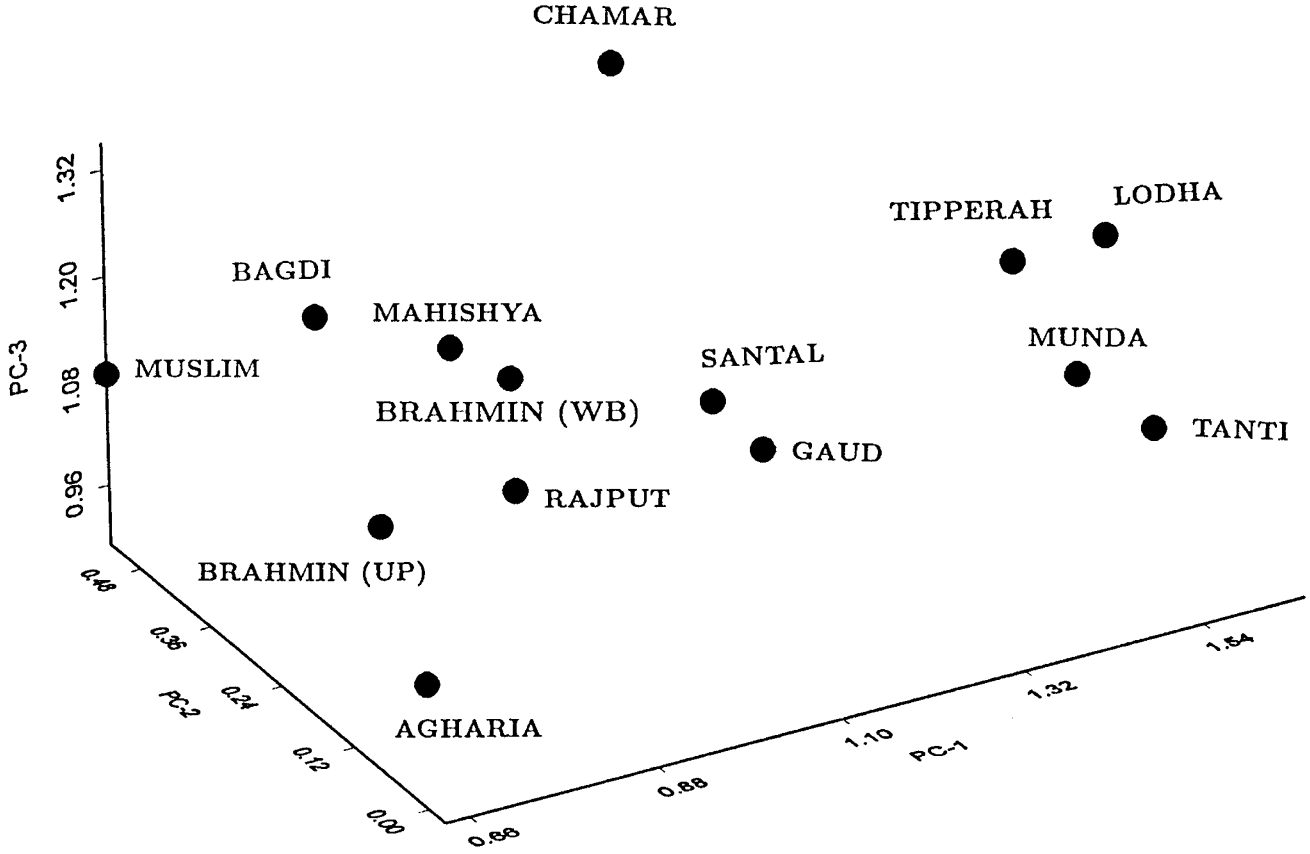

Figure 2 *Genomic affinities among 14 Indian ethnic populations based on first three principal components of allele frequencies at eight loci.*

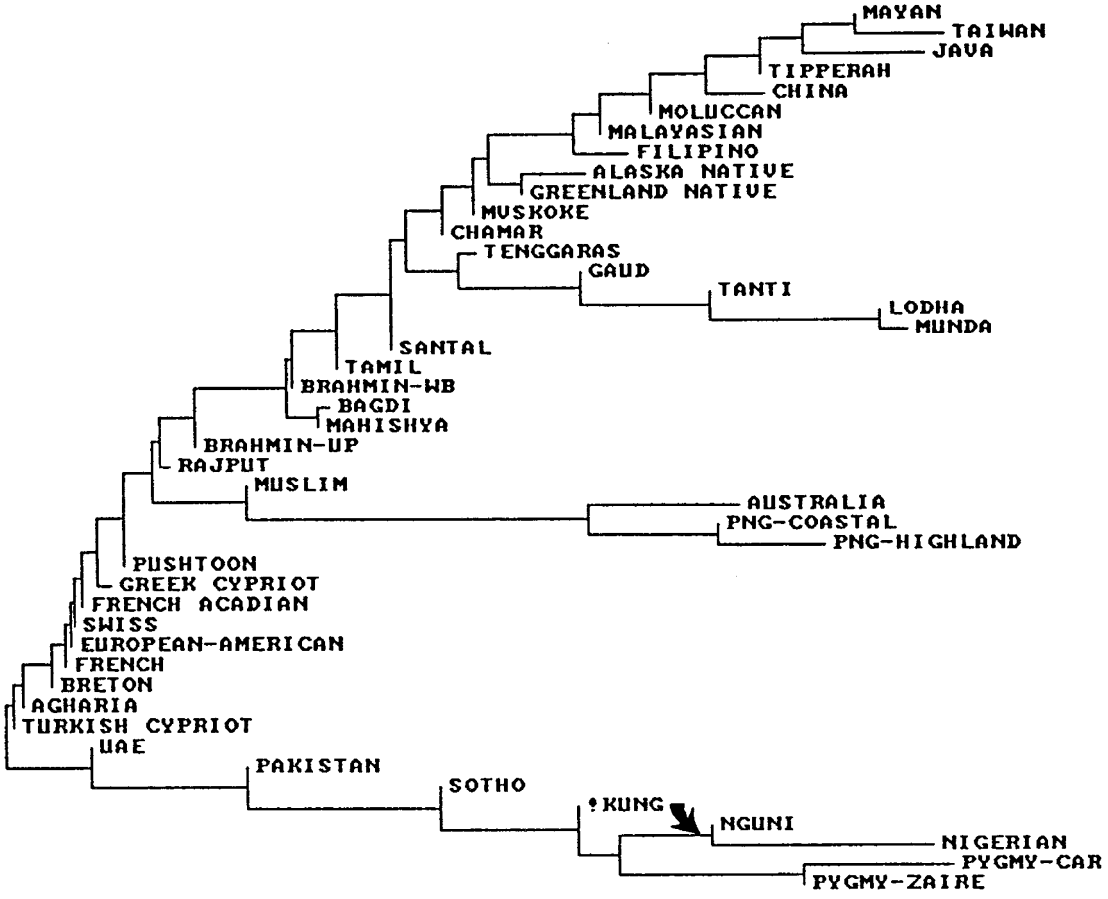

Figure 3 *Neighbour-joining tree depicting genomic affinities among 45 global populations based on allele frequencies at 6* Alu *insertion loci. The tree was rooted using a hypothetical ancestral population (see text for details); the root is marked with an arrow.*

French, Swiss, European-American, etc). Two of the three Indian Austro-Asiatic speaking groups (Lodha and Munda) again stand apart genetically. The Tipperah (a Sino-Indian speaking population) are genetically close to the Chinese, Javanese and Taiwanese. We also note that the African populations are also genetically quite distant from the other global populations. Since the ancestral states (non-insertion) of these *Alu* polymorphisms are known, it is possible to root the unrooted NJ tree by using the ancestral population (with zero insertion-allele frequencies at all loci) as an outgroup. The root lies very close to the cluster of African populations.

Gene Flow among Populations

With the objective of testing whether in a group of incompletely isolated populations distributed over a geographical space (Wright's island model), observed patterns of genomic diversities are the outcome of the processes of drift and migration among the populations or whether these patterns are generated by interactions with populations outside the set of populations under consideration, Harpending and Ward[14] derived a regression of heterozygosity on genetic distance. They have shown that the genetic distance of an island population from the gene frequency centroid (the overall mean gene frequencies of all populations) and the relative homozygosity of that island population should be linearly related if exchange with populations from outside is the same for each island. If gene flow from outside varies in amount from one population to another, this linear relationship no longer holds. Very isolated populations should be less heterozygous than the linear prediction, whilst populations which receive more genes should be more heterozygous than predicted.

We have plotted the observed heterozygosities of the 14 populations against the distance from the gene frequency centroid in Figure 4. The theoretical linear regression line is also drawn on this figure. It is seen

from this figure that of the 14 populations, three have experienced lesser gene flow than predicted, whilst the gene flow in 10 populations has been higher than predicted. The observed and expected heterozygosities are nearly equal for the Muslim group.

We have also performed this centroid analysis by pooling the data of the present study with those of Stoneking *et al*.[15] For reasons mentioned earlier, this analysis was based on data of six *Alu* polymorphic loci. The observed and predicted heterozygosities and the distance of the observed gene frequencies from the centroid are presented for each of the 45 populations in Table 6. These data are also graphically presented in Figure 5. It is seen from this figure that for 18 populations the observed heterozygosities are greater than predicted values. Of these 19 populations, 11 are Indian (10 of the present study and the Tamils studied by Stoneking *et al*[15]) and five are African. (The Indian populations for whom the observed heterozygosity is lower than the predicted are: Brahmin-UP, Muslim, Rajput and Tipperah; the African population exhibiting this pattern is !Kung.) Thus, overwhelmingly the African (5 out of 6) and Indian (11 out of 14) populations show higher heterozygosity than predicted.

Discussion

Human-specific insertion/deletion polymorphisms have already proved to be very useful in studies on genetic structure of human populations.[15–18] Since new alleles at these loci are not generated and also because these loci are unlikely to be under any selection pressure, allele frequency variation among populations in respect of these loci must necessarily have been generated by the effects of genetic drift and migration. As has been reported in ethnic populations from various parts of the world,[15–18] these loci show high levels of polymorphism in the population groups of India also. It is also seen that the levels of average heterozygosity are consistently high in all the populations investigated in this

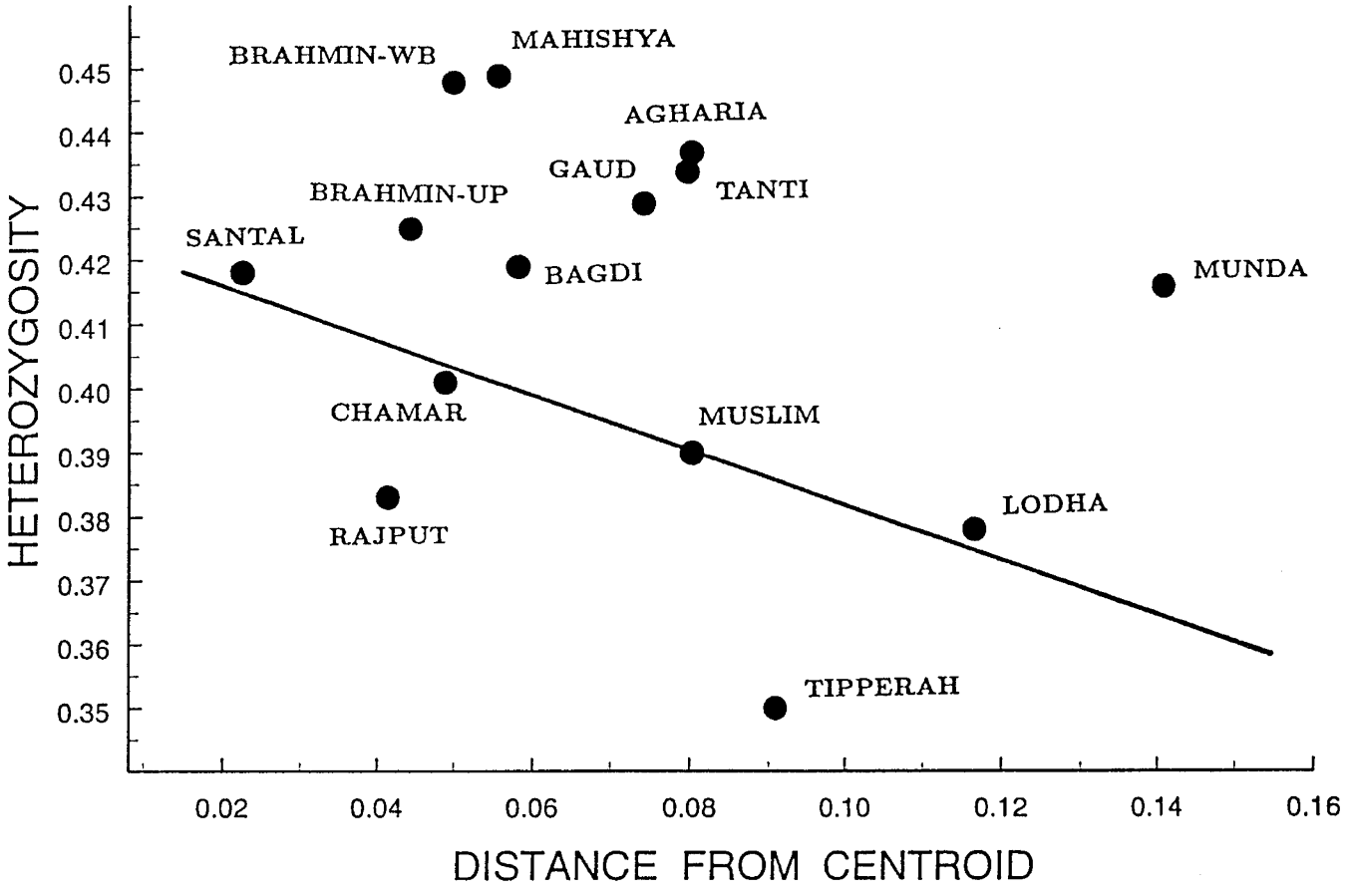

Figure 4 *Plot of average heterozygosity* vs *distance from allele frequency centroid of 14 Indian ethnic populations based on allele frequency data of eight loci.*

Table 6 Distances from gene frequency centroid, expected and observed average heterozygosities in 45 global populations based on six *Alu* insertion polymorphisms

Sl. no.	*Population name*[a]	*Distance from centroid*	*Expected heterozygosity*	*Observed heterozygosity*
1	Agharia	0.060	0.428	0.475±0.027
2	Bagdi	0.093	0.414	0.452±0.041
3	Brahmin-UP	0.021	0.446	0.446±0.049
4	Brahmin-WB	0.051	0.432	0.465±0.046
5	Chamar	0.084	0.417	0.460±0.027
6	Gaud	0.143	0.391	0.469±0.031
7	Lodha	0.272	0.332	0.413±0.046
8	Mahishya	0.061	0.428	0.466±0.034
9	Munda	0.315	0.312	0.470±0.011
10	Muslim	0.062	0.428	0.423±0.064
11	Rajput	0.035	0.440	0.419±0.050
12	Santal	0.032	0.441	0.462±0.024
13	Tanti	0.198	0.366	0.478±0.024
14	Tipperah	0.156	0.384	0.378±0.048
15	Alaska Native	0.144	0.390	0.369±0.070
16	Australia	0.291	0.323	0.237±0.052
17	Breton	0.062	0.428	0.428±0.051
18	China	0.167	0.380	0.362±0.038
19	European-American	0.091	0.414	0.406±0.067
20	Filipino	0.157	0.384	0.372±0.071
21	French	0.105	0.408	0.399±0.079
22	French Acadian	0.073	0.422	0.412±0.062
23	Greek Cypriot	0.074	0.422	0.393±0.062
24	Greenland Native	0.093	0.413	0.402±0.063
25	Java	0.267	0.334	0.338±0.057
26	!Kung	0.269	0.333	0.304±0.028
27	Malaysian	0.082	0.418	0.429±0.021
28	Mayan	0.226	0.353	0.342±0.065
29	Moluccan	0.108	0.406	0.405±0.029
30	Mvskoke	0.101	0.410	0.394±0.069
31	Nguni	0.260	0.337	0.386±0.042
32	Nigerian	0.475	0.239	0.310±0.092
33	Pakistan	0.126	0.399	0.419±0.035
34	PNG-Coastal	0.158	0.384	0.396±0.037
35	PNG-Highland	0.251	0.341	0.320±0.065
36	Pushtoon	0.022	0.446	0.436±0.042
37	Pygmy-CAR	0.415	0.266	0.321±0.076
38	Pygmy-ZAIRE	0.313	0.313	0.363±0.070
39	Sotho	0.159	0.383	0.423±0.026
40	Swiss	0.086	0.417	0.400±0.063
41	Taiwan	0.296	0.320	0.307±0.083
42	Tamil	0.033	0.441	0.450±0.029
43	Tenggaras	0.069	0.424	0.403±0.036
44	Turkish Cypriot	0.104	0.408	0.402±0.073
45	UAE	0.142	0.391	0.348±0.078

[a]The first 14 populations are from the present study; the remaining populations have been studied by Stoneking *et al*[15].

study. Thus, consistent with the findings of classical markers,[1] these DNA markers confirm that Indian populations exhibit high levels of genomic diversity. The extent of genomic differentiation (G_{ST}) among the 14 Indian populations (0.068) is higher than those observed[15] in all other parts of the world except Africa. We note that some of the comparisons between the estimated G_{ST} values of the present study and those of Stoneking *et al*[15] may not be strictly valid because the populations included in the present study are anthropologically well defined ethnic groups, whilst some of those included in Stoneking *et al*'s[15] study are agglomerates of several ethnic groups. However, no disaggregated data sets on these polymorphisms are currently available; hence, no fully valid comparisons of estimated G_{ST} values are currently possible. The results of the present study also confirm earlier findings[1] based on classical markers that

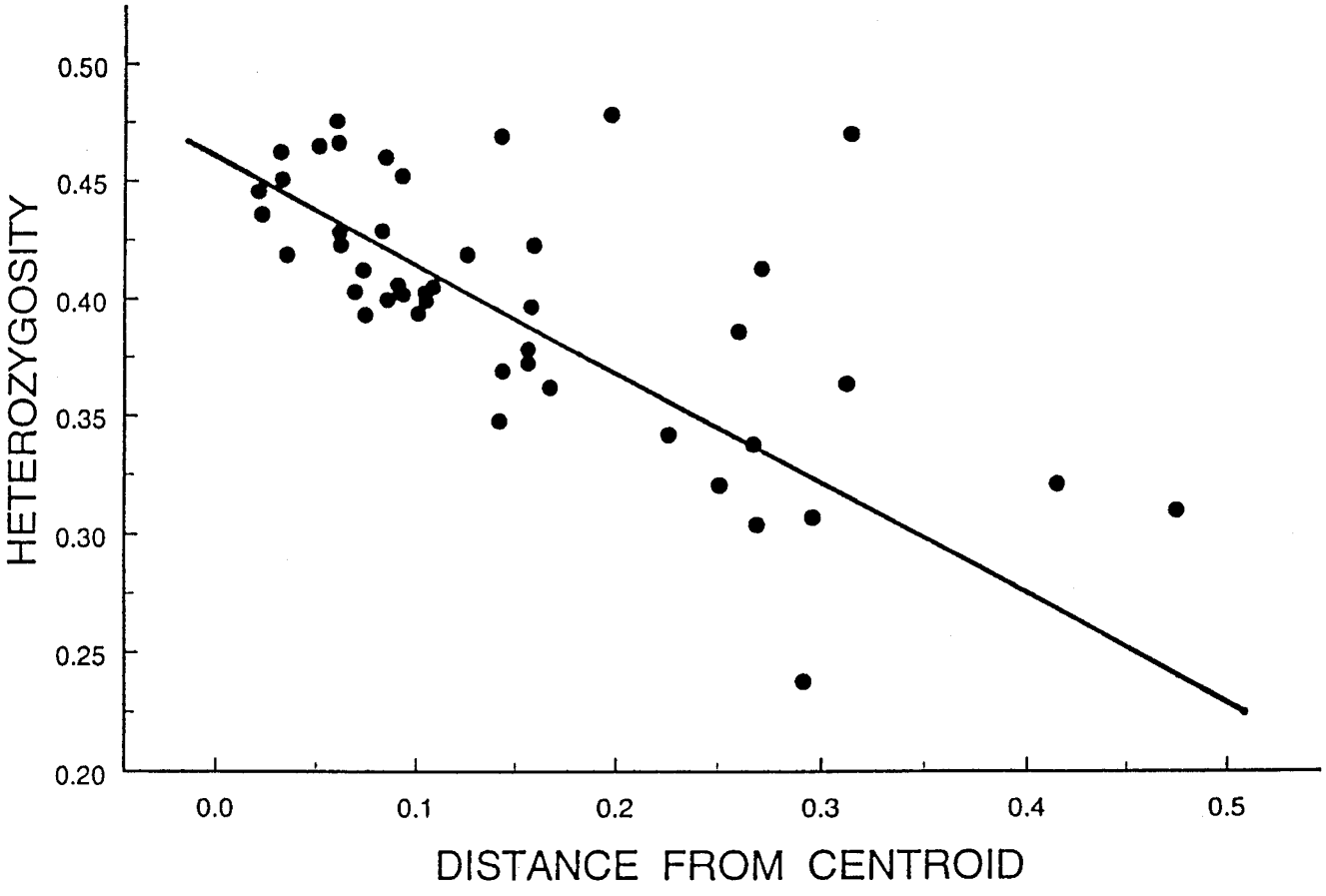

Figure 5 *Plot of average heterozygosity* vs *distance from allele frequency centroid of 45 global populations based on data of six* Alu *insertion loci.*

(i) genetic affinities among Indian populations do not correlate well with socio-cultural rankings; geographically closer populations are also genetically closer (Figures 1 and 2), and

(ii) Indian populations are genetically between Caucasoids and Mongoloids (Figure 3).

The centroid analysis (Figure 4) also shows that there has been a considerable amount of gene flow between the set of populations under consideration and other populations. This, by and large, is in agreement with anthropological findings.[1,19–26] In spite of considerable gene flow as inferred from the centroid analysis, the extent of gene differentiation among Indian populations continues to be high. Since this analysis does not permit timing of the period during which the gene flow may have occurred between these Indian and other populations, we are unable to offer a clear interpretation of this finding. A likely explanation is that gene flow occurred prior to the subdivision of these Indian populations into largely endogamous units.

The dominant prevailing view of the origin and spread of modern humans is that *Homo sapiens* originated in Africa 100 000–200 000 years ago and that all present human populations outside sub-Saharan Africa are primarily descendants of a population that moved out of Africa about 100 000 years ago.[27] Harpending *et al*[28] have suggested that after the migration of modern humans from Africa, there were many rapid population expansions following an initial period of isolation. Ballinger *et al*,[29] on the basis of data on mtDNA polymorphisms, suggested that such an expansion may have taken place in southern China. Mountain *et al*,[30] using similar mtDNA data, have hypothesised that a centre of this expansion may have been in or close to India. Although the high levels of heterozygosity observed in African populations is compatible with a number of hypotheses that do not assume an African origin,[31] the findings that the 'root' (ancestral states) of these *Alu* polymorphisms lies close to the cluster of African populations[15,17] and that heterozygosities observed in African populations are higher

than those predicted have been interpreted[15,17] as evidence supporting the out-of-Africa theory and of a greater effective population size across Africa. Thus African populations most probably underwent a large expansion before they moved out of the continent and were to become the source of modern humans in other parts of the world. In the present study, we have presented evidence that, with respect to the *Alu* insertion polymorphisms, the Indian populations show levels of heterozygosity (0.448 ± 0.039) that are higher, although not always significantly, than most global populations, including African populations. In fact, among the 45 observed heterozygosity values presented in Table 6, the highest ten values are observed in Indian ethnic groups. Further, we have found that the vast majority of Indian populations show higher levels of heterozygosity than predicted by the Harpending–Ward[14] gene flow model. If this pattern of high heterozygosities were simply due to higher levels of gene flow, then one would have expected that Indian populations would be genetically less differentiated. However, we have found that the coefficient of gene differentiation among Indian populations is higher than among populations inhabiting all other regions of the world, except Africa.[15] Two explanations of the observation of higher than predicted heterozygosities coupled with a high level of genetic differentiation are:

(i) inflow of genes into the populations under study have been high (resulting in higher than predicted heterozygosities), but different study populations have had different sources of genes (resulting in high levels of genetic differentiation), and

(ii) an early inflow of genes into a population followed by a rapid expansion of this population (resulting in high heterozygosities) and subsequent splits of this population into largely isolated (endogamous) populations (resulting in high levels of genetic differentiation).

We are unable to provide any strong evidence favouring either of these two alternative possibilities. In the anthropological literature pertaining to the study populations, there are no observations to support the hypothesis that the different study populations have had inflow of genes from different external sources.

Our present data and analyses also do not permit evaluation of the process and estimation of rates of increase of heterozygosities. However, we do wish to emphasise that since our joint observation of higher than predicted heterozygosities and high level of genetic differentiation have earlier been accepted as hallmarks of population expansion,[15,17] the possibility of an early demographic expansion of modern humans within India cannot be ruled out. Support for such a possibility also comes from material culture remains found in India that show the evidence that upper palaeolithic (40 000 years before the present) cultures flourished in different parts of India.[32] We are now considering alternative ways of testing the two possible evolutionary scenarios that the present study has indicated.

Acknowledgements

This study was supported by a grant from the Department of Biotechnology, Government of India to PPM. We are grateful to Dr RS Balgir and Dr BP Dash for help in collecting samples from Orissa.

References

1 Majumder PP: People of India: Biological diversity and affinities. *Evol Anthrop* 1998; **6**: 100–110.
2 Zischler H, Geisert H, von Haeseler A, Paabo S: A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* 1995; **378**: 489–492.
3 Anderson S, Bankier AT, Barrell BG *et al*: Sequence and organisation of the human mitochondrial genome. *Nature* 1981; **280**: 457–465.
4 Deininger PL, Batzer MA: Evolution of retroposons. *Evol Biol* 1993; **27**: 157–196.
5 Ullu E, Murphy S, Melli M: Human 7S RNA consists of a 140 nucleotide middle repetitive sequence inserted in an Alu sequence. *Cell* 1982; **29**: 195–202.
6 Rogers J: Retroposons defined. *Nature* 1983; **301**: 460.
7 Batzer MA, Deininger PL: A human-specific subfamily of Alu sequences. *Genomics* 1991; **9**: 481–487.
8 Edwards MC, Gibbs RA: *Genomics* 1992; **14**: 590–593.
9 Tishkoff SA, Dietzsch E, Speed W *et al*: Global patterns of linkage equilibrium at the CD4 locus and modern human origins. *Science* 1996; **271**: 1380–1387.
10 Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; **16**: 1215.
11 Nei M: Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* 1973; **70**: 3321–3323.
12 Felsenstein J: PHYLIP, version 3.5c. University of Washington: Seattle, 1993.
13 Nei M, Tajima F, Tateno Y: Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol* 1983; **19**: 153–170.
14 Harpending H, Ward RH: Chemical systematics and human populations. In: Nitechi MH (ed.). *Biochemical Aspects of Evolutionary Biology*. University of Chicago Press: Chicago, 1982, pp 213–256.

15 Stoneking M, Fontius JJ, Clifford SL *et al*: *Alu* insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res* 1997; **7**: 1061–1071.
16 Batzer MA, Stoneking M, Alegria-Hartman M *et al*: African origin of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci USA* 1994; **91**: 12288–12292.
17 Batzer MA, Arcot SS, Phinney JW *et al*: Genetic variation of recent Alu insertions in human populations. *J Mol Evol* 1996; **42**: 22–29.
18 Novick GE, Novick CC, Yunis J *et al*: Polymorphic Alu insertions and the Asian origin of native American populations. *Hum Biol* 1998; **70**: 23–29.
19 Bodding PO: Traditions and Institutions of the Santals. Oslo Etnografiske Museum: Oslo, 1942.
20 Briggs GW: The Chamars. Association Press: Calcutta, 1920.
21 Risley H: The People of India. Thacker Spink & Co.: Calcutta, 1915.
22 Guha BS: The Racial Elements in Indian Population. Oxford University Press: Calcutta, 1938.
23 Sarkar SS: The Aboriginal Races of India. Bookland: Calcutta, 1954.
24 Mahalanobis PC, Majumdar DN, Rao CR: Anthropometric survey of the United Provinces, 1941: A statistical study. *Sankhya* 1949; **9**: 90–324.
25 Majumdar DN, Rao CR: Race Elements in Bengal: A Quantitative Study. Statistical Publishing Society: Calcutta, 1960.
26 Fuchs S: The Aboriginal Tribes of India. Macmillan: Calcutta, 1973.
27 Nei M: Genetic support for the out-of-Africa theory of human evolution. *Proc Natl Acad Sci USA* 1995; **92**: 6720–6722.
28 Harpending HC, Sherry ST, Rogers AR, Stoneking M: Genetic structure of ancient human populations. *Curr Anthropol* 1993; **34**: 483–496.
29 Ballinger SW, Schurr TG, Torroni A *et al*: Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics* 1992; **130**: 139–152.
30 Mountain JL, Herbert JM, Bhattacharyya S, Underhill PA, Ottolenghi C, Gadgil M, Cavalli-Sforza LL: Demographic history of India and mtDNA-sequence diversity. *Amer J Hum Genet* 1995; **56**: 979–992.
31 Templeton AR: 'Eve' hypothesis compatibility versus hypothesis testing. *Amer Anthropol* 1994; **96**: 141–147.
32 Misra VN: Stone age in India: An ecological perspective. *Man and Environment* 1989; **14**: 17–64.