Molecular Anthropology: Population and Forensic Genetic Applications

Sarabjit Mastana

INTRODUCTION

Molecular anthropology - the study of human genetic polymorphisms - is fast and ever-growing branch of anthropology that holds a great promise for both past and future. Some anthropologists believe that genetic/molecular anthropology is a science of future but it must be emphasised that it is a science of the past and present too. DNA is an unbroken link to our ancestors, populations and relatives.

Molecular anthropology is useful in estimating the contribution of different gene pools to the make-up of present-day populations and test hypotheses about origin of linguistic and historical population movements. In addition, anthropology is playing a significant role in our understanding of gene-environment interactions and contribution of populations to the detection of genes in common and complex diseases. Anthropologists interested in reconstructing historical population movements and phylogenetic relationships initially used "classical" genetic data to achieve these aims. Classical genetic data uses proteins and blood groups. The classical genetic data comprehensively provided the basis for the "first phase" in the development of molecular anthropology, which has been remarkably documented in many international and Indian research compendiums (Bhasin et al., 1994; Bhasin and Walter, 2001; Cavalli-Sforza et al., 1994, Mourant et al., 1976, Nei and Roychoudhury, 1988). These large-scale studies have demonstrated that the gene pool is not a simple sum of genes, but is a dynamic system, which is hierarchally organised and which maintains the memory of past events in the history of populations. All genetic information has a historical, anthropological, geographical and statistical context, therefore requires co-operation and collaboration between researchers in different fields.

The "new" or "second phase," is utilizing DNA analysis for the reconstruction of human population structure, histories and evolution. The potential benefits from this research are vast and valuable including; a better understanding of the genetic and evolutionary factors that influence populations; an understanding of genetic architecture of common and complex diseases such as diabetes mellitus, dementias, heart and skeletal diseases; and a better understanding of the origin of modern humans. The pattern of genetic variation in modern human populations depends on our demographic history (population migrations, bottlenecks and expansions) as well as gene specific factors such as mutation rates, recombination rates and selection pressure. By examining patterns of genetic polymorphisms we can infer how past demographic events and selection have shaped variation in the genome. Thus, molecular anthropology has important implications for evolutionary biology, disease analyses, and forensics. In this paper, first the anthropological and genomic basis for genetic variation is overviewed followed by some specific empirical research examples highlighting the usage of the molecular anthropological investigations. The focus here is on Indian studies but it cannot be exhaustive due to space considerations.

ANTHROPOLOGICAL BASIS OF GENETIC VARIATION IN INDIA

Contemporary Indian population can broadly be classified into four major types: Negrito, Australoid, Mongoloid and Caucasoid (Malhotra, 1978; Bhasin et al., 1994; Bhasin and Walter, 2001). The Negrito is mostly confined to the Andaman Islands, Nilgiri Hills and a few regions along the western coast of India. The Australoids are primarily distributed in the central and southern regions. The Mongoloid element is mostly found in the north-east and the sub-Himalayan regions, while the Caucasoid element in India is the most widespread.

The Indian population is socially organised into distinct groups that are largely endogamous and reproductively isolated. The majority practice Hinduism and follow the Hindu caste system and are hierarchically organized in Upper, Middle and Lower strata. Cultural norms act as barriers to inter-caste marriages. Traditionally, each caste pursued a hereditarily prescribed occupation, and different castes were linked to each other through a pre-determined pattern of barter of services and produce (Karve, 1961). Many religious communities, such as, Muslim, Christian, Sikh, do not belong to the Hindu or the tribal groups. Indian populations exhibit a wide variety of cultural and linguistic diversities. The population of the subcontinent can be divided into five major linguistic families: Iranian, Indo-Aryan, Austro-Asiatic, Tibeto-Burmese and Dravidian. The majority of the people speak Indo-Aryan languages, followed by Dravidian. The Austro-Asiatic speakers were indigenous to the eastern region in ancient times. The Dravidian speaking tribes of South and Central India are thought to be descendants of the original inhabitants of the Indian subcontinent who adopted the Dravidian language in preference to their own mother tongue.

Within each linguistic and religious group, socio-cultural and biological characteristics delineate numerous endogamous ethnic groups. These ethnic groups fall into the broad categories of caste/scheduled castes, tribes/scheduled tribes, and some other communities. This social organisation may be further complicated due to territorial affiliation of various tribes and caste groups. It is evident from many studies that the high level of heterogeneity in Indian populations, governed by high level of endogamy produced by numerous social restrictions, along with genetic drift has kept the Indian gene pool distinct from other continental populations. Therefore, for biological and medical studies, the criteria for selection of population should be based on their regional, anthropological, linguistic, religious, tribal and ethnic attributes.

HUMAN GENOMIC VARIATION

The human genome consists of 3 billion base pairs of nuclear DNA (nDNA) and around 16.6 Kb of extra- nuclear mitochondrial DNA (mtDNA). The completion of the Human Genome Project and its descendant, the HapMap project, has provided researchers with enormous opportunities and genetic markers for disease, population and evolutionary studies. It is now well established that less than 5% of nuclear genome codes for proteins and the remaining nuclear genome consists of unique or low copy number sequences and moderate to highly repetitive sequences. This non-coding genome, whose biological function is still not clearly defined, is equally vulnerable to mutations and has become a goldmine for anthropological and population genetic studies. These newly defined DNA markers not only gave a new look to the investigation of human genetic diversity, but initiated a new and important era in the application of anthropological genetics to the field of forensic genetics and molecular medicine. Many types of DNA polymorphism have been observed in the coding and non-coding parts of the human genome, the prominent among them are outlined below.

a. Single Nucleotide Polymorphisms (SNPs)

DNA polymorphisms were first studied by Southern analysis of DNA digested with restriction enzymes. This type of polymorphism is called restriction fragments length polymorphism (RFLP), as alleles differ in the length of the restriction fragments obtained upon digestion with various restriction enzymes. The most common reason of RFLPs is a nucleotide replacement/ substitution in the recognition site. This type of polymorphism is now renamed as SNP (single nucleotide polymorphism). Single Nucleotide Polymorphisms (SNPs, commonly pronounced as “SNIPs”) are the most abundant types of polymorphisms in the human genome. The human genome contains more than 10 million SNPs. Due to the improvement and automation of sequencing methodologies, and the development of DNA micro-arrays, these markers are now extensively studied in the human genome for their association with different complex diseases, for understanding various aspects of population differentiation and human evolution (Rosenberg et al., 2002; Clark et al., 2003; Salisbury et al., 2003; Garrigan and Hammer, 2006). It is expected that a uniform analysis of vast numbers of randomly selected SNPs on various populations will provide a better understanding of genetic affinities, disease diagnostics, forensics and analysis of complex diseases. Comprehensive studies on anthropologically interesting populations from India are in progress (The Indian Genome Variation database, IGVdb, 2005).

b. Repeat Length Polymorphisms

Repetitive sequence elements are distributed over the entire human genome, and they are subdivided into tandemly arrayed (satellites, minisatellites, microsatellites) or interspersed (LNIES and SINES like *Alu* repeat) repetitive sequences. Jeffreys and colleagues (1985) were the first to demonstrate the potential of repeat length polymorphisms (minisatellites) for forensic and anthropogenetic applications. Minisatellites have a larger core repeat sequence (normally 10-70bp) and are highly polymorphic but their analyses are laborious and require high quality DNA and complex statistical methods. Minisatellites became a popular choice in forensic and population analyses in the late 1980s and early 1990s but increasingly have been replaced by other genetic markers (Bowock et al., 1994, Papiha et al., 1996a). Microsatellites, also called STRs (short tandem repeats), have a shorter core sequence (2-6bp), are PCR-able, can be multiplexed and automated, and have discrete alleles. Both Minisatellites and microsatellites are highly variable, polymorphic, co-dominant and heterozygous and thus are excellent tools for genetic individualization and population genetic studies (Bowock et al., 1994; Papiha et al., 1996a; Das et al., 2002; Mastana and Singh, 2002; Das and Mastana, 2003; Ranjan et al., 2003; Ghosh et al., 2003; Agrawal et al., 2003; Sachdeva et al., 2004; Kashyap et al., 2005; Mastana et al., 2006). Kashyap and colleagues have documented STR variation in a large number Indian populations for forensic and population genetic applications (see Kashyap et al., 2004, 2006 and references within these papers for further details).

Chakraborty (1990) suggested that even a single minisatellite locus could provide information concerning sub-structuring within a population with a statistical power greater than several classical genetic markers studied simultaneously. The minisatellite loci have high heterozygosity and a vast number of alleles but their high mutation rate makes them more useful to explore recent population history. However, microsatellite or STRs are used extensively to analyse intra and inter-population affinities and deeper evolutionary history. Overall, these studies have highlighted that these loci are polymorphic in Indian populations and there is no significant deficiency in heterozygosity levels as one would expect in endogamous populations. The F_{ST} values for the minisatellites and microsatellites are of low to moderate range (1.2-5.4%) in various studies. These values are slightly higher than the F_{ST} values reported for conventional system from different populations and regions of the Indian sub-continent (range 0.6% to 2.1%) (Papiha, 1996).

Interspersed repeated DNA sequences can be divided into two classes: short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs). The most extensively studied class of SINEs are *Alu* insertions due their abundance as well as their association with many biological functions and diseases (Batzer and Deininger, 2002). Further details of the *Alu* insertion polymorphisms are outlined below (see-examples of research applications).

A worldwide study of *Alu* and microsatellite polymorphisms revealed that it was possible to classify individuals as belonging either to African, European or East Asian continental clusters with high probability using around 100 genetic loci. However, Indian populations, when added to the STRUCTURE analyses, failed to form/belong to a discrete cluster indicating a significant genetic heterogeneity among the populations studied (Bamshad et al., 2003)

A recent on 1,200 genome-wide polymorphisms (microsatellites and insertion deletion polymorphisms) in 432 individuals from 15 Indian populations sampled in the United States demonstrated a low level of genetic differences and differentiation among these populations (Rosenberg et al., 2006). The selection of these populations is based on spoken language and as such may not be representative of true endogamous and heterogeneous nature of population of India, therefore results should be considered with caution. In addition, number of individuals studied from each group/region/language group is small.

c. Uni-parental DNA Polymorphisms

The mtDNA and Y chromosome analyses have proven to be the most useful for studying historical population movements because of their ease in analyses and non-recombining nature. In the absence of a recombination event, both mitochondrial and Y-chromosomes behave as single units, and various markers stretched across are inherited as single blocks. This synteny of markers generates haplotypes, and the frequency of these haplotypes show great diversity in human populations. The mtDNA is inherited through the maternal cytoplasm therefore; variation in mtDNA provides a record of the maternal lineages. Y chromosome DNA documents the paternal lineage (Jobling and Tyler-Smith, 2003). Essential data and information can be obtained using the Y and mtDNA, which is often helpful in understanding the difference between male and female migration and biological evolution to the present day. A large number of research papers have recently documented the mtDNA and Y-chromosome variation among Indian populations (Bamshad et al., 2001; Kivisild et al., 1999; Corduex et al., 2003; Kivisild et al., 2003; Basu et al., 2003; Sengupta et al., 2006; Sahoo et al., 2006; Gutala et al., 2006; Thanseem et al., 2006; Zerzal et al., 2006).

The mitochondrial DNA (mtDNA) studies showed that a number of haplogroups and haplotypes are specific to Indian populations. The unique mtDNA haplogroups in Indian populations include: U2a,b,c, R5-8, R30, R31, N1d and N5 in haplogroup N and M2–6, M30–47 in haplogroup M. More than 60% of Indians find their maternal roots in Indian-specific branches of haplogroup M. Because of its deep time depth and virtual absence in West Eurasians, it has been suggested that haplogroup M was brought to Asia from East Africa along the southern route about 60,000 years ago by the earliest migration wave of anatomically modern humans (Kivisild et al., 1999; Kivisild et al., 2003; Basu et al., 2003). Within M haplogroup, extensive variation has been documented which is geographically population specific. Mitochondrial DNA profiles from a larger set of populations all over the sub-continent have bolstered the view of fundamental genomic unity of Indians (Roychoudhury et al., 2001; Basu et al., 2003; Kivisild et al., 2003; Corduex et al., 2003). Thanseem et al. (2006) also did not find any significant difference between Indian tribal and caste populations in mtDNA, except for the presence of a higher frequency of west Eurasian-specific haplogroups in the higher castes from North India. Overall diversity in the mitochondrial genome in Indian populations was estimated to be nearly as high as in Africans, and higher than in Europeans and other Asians.

The Y-chromosome analyses showed that the Indian caste populations were more closely related to Europeans than East Asians (Bamshad et al., 2001). The tendency of higher caste status to associate with increasing affinities to European populations hinted at a recent male-mediated introduction of West Eurasian genes into the Indian gene pool (Bamshad et al., 2001; Kivisild et al., 2003; Basu et al., 2003; Sengupta et al., 2006; Sahoo et al., 2006; Gutala et al., 2006; Zerzal et al., 2006). Recent studies on Indian Y lineages revealed distinct distribution patterns among caste and tribal populations. The paternal lineages of Indian lower castes showed significantly closer affinity to the tribal populations than to the upper castes. The frequencies of deep-rooted Y haplogroups such as M89, M52, and M95 were higher in the lower castes and tribes, compared to the upper castes (Thanseem et al., 2006).

Many additional types of DNA polymorphisms (Copy number variants, CNVs) have recently been discovered but the anthropological usage of these markers is yet to be examined in detail. The selection of the genetic markers for anthropological research is determined by the ability of the given marker to solve the tasks and by the technical aspects including type and nature of sample collection.

Overall studies on genetic diversity and affinities among contemporary human populations are useful for past population movements and identifying ancestral populations, and disease and association studies.

MOLECULAR ANTHROPOLOGY: SOME EMPIRICAL RESEARCH EXAMPLES

1. *Alu* Insertion Variation in India

In this section, using a part of research data from our on going studies, the applications of DNA analyses are shown. *Alu* insertion polymorphisms have been chosen as these are dimorphic, bi-parental, relatively easy to analyse and do not require specialist equipment.

Alu insertional elements represent the largest family of Short INterspersed Elements (SINEs) in humans. They are named due to the presence of an *AluI* recognition site in the sequence. The human genome contains more than 1 million *Alu* repeats, which account for ~10% of the total nuclear DNA. *Alu* repeats are generally located in non-coding regions. *Alu* insertions are of approximately 300 bp in length, dimeric in structure, and are composed of two nearly identical monomers joined by a middle A-rich region along with a 3’ oligo (dA)-rich tail and short flanking direct repeats (Fig. 1).

Alu elements increase in number by retro-

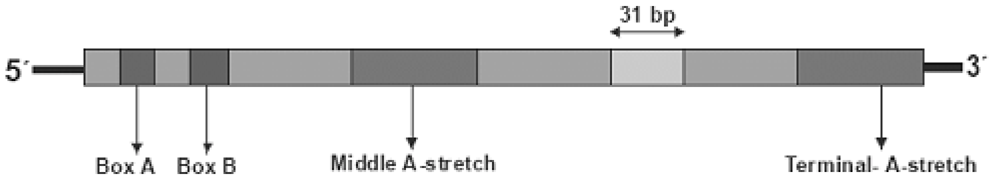

Fig. 1. The dimeric structure of the *Alu* element. The two halves are linked by an adenine-rich area. The right monomer includes a 31-base pair insertion, and the left half contains the RNA polymerase III promoter (boxes A and B). The total length of each *Alu* sequence is ~300 bp, depending on the length of the 3' oligo (dA)-rich tail.

transposition-a process that involves reverse transcription of an *Alu*-derived RNA polymerase III transcript (Batzer and Deininger, 2002). *Alu* repeats can affect the composition, organization and expression of the genome, therefore they play a significant role in the occurrence of human genetic diseases.

A variety of *Alu* subfamilies have been discovered based on the variation of the base DNA sequence. The youngest subfamilies have integrated into the human genome in the past 4-5 million years after the divergence of humans and African apes (Batzer et al., 1996; Roy-Engel et al., 2001). Approximately 5000 *Alu* elements in the human genome have integrated into the human genome relatively recently, following the human-chimpanzee split (Batzer et al., 1996; Stoneking et al., 1997). Furthermore, about 25% of this small subset are not fixed in human populations and therefore provide polymorphic markers in the form of presence/absence variants. These *Alu*-based variants provide useful DNA markers for human population studies, forensic studies and paternity analyses.

Alu insertions/repeats are convenient genetic markers. First, the insertion of an *Alu* element at a certain chromosomal site is an unique event, which means the individuals that share *Alu* insertion polymorphisms have inherited the *Alu* elements from a common ancestor, which makes the *Alu* insertion alleles identical by descent. Second, they are stable polymorphisms - once inserted, the elements are fixed in the genome, as there are no specific mechanism for removing them. Even when a rare deletion occurs, a significant remnant is left behind. Third, the ancestral state of the *Alu* insertion is known to be the absence of the insertion. Polymorphic *Alu* insertions are human specific and absent in nonhuman primates. It is possible to create a hypothetical ancestral population with frequencies of zero for all human specific *Alu* insertions therefore allowing phylogenetic analyses. The recent studies have shown that the root of population tree is located near the African Sub-Saharan populations presented evidence for an African origin of modern human populations (Stoneking et al., 1997). While there are many reports on *Alu* polymorphisms in different populations of the world (Batzer et al., 1996; Stoneking et al., 1997; Watkins et al., 2003), studies from the Indian Subcontinent are limited to relatively small number of loci and populations (Majumder et al., 1999; Viswanathan et al., 2003).

This study employed 30 *Alu* insertion polymorphisms in 10 endogamous populations belonging to two geographical areas, consisting of a total of 984 unrelated individuals. An *Alu* locus consisting of presence and absence variants of an *Alu* element yields three possible genotypes. Assaying additional *Alu* variants exponentially increases the number of possible genotypes (3^n). *Alu* dimorphisms were selected in which both presence and absence alleles are common and hence more informative. Information on intermediate frequency *Alus* was collected from published literature (Carroll et al., 2001; Roy-Engel et al., 2001; Watkins et al., 2003,) including primer sets, expected amplicon sizes, and frequencies among various ethnic groups. The *Alu* loci analysed in this study included, TPA25, ACE, APO, D1, B65, Col3A1, HS4.32,HS4.65, PV92, Sb19.12,Sb19.3, Ya5NBC102, Ya5NBC123, Ya5NBC148, Ya5NBC171,Ya5NBC182, Ya5NBC 216, Ya5NBC241, Ya5NBC242, Ya5NBC27, Ya5 NBC333, Ya5NBC354, Ya5NBC61, Yb8NBC106, Yb8NBC120, Yb8NBC125, Yb8NBC13, Yb8NBC 201, Yb8NBC207 and Yb8NBC243 (Carroll et al., 2001; Roy-Engel et al., 2001; Watkins et al., 2003).

The cohort of the populations studied included North India (Punjab) [5] and Western India [5]. The Punjabi endogamous populations were sampled from towns and villages of Patiala district and included Brahmin (103, hereafter referred as Brahmin-N), Khatri (110), Jat Sikh (106), Lobana (104) and Scheduled caste (86). The Western Indian populations were sampled from Mumbai and included Brahmin (101 hereafter referred as Brahmin-W), Maratha 110), Muslim (96), Patel (104) and Parsee (64) (Mastana and Papiha, 1994; Papiha et al., 1996a,b). Insertion allele frequencies and associated standard errors were calculated for each locus separately in individual population sample. An estimate of unbiased heterozygosity was obtained from expected allele frequencies for each locus. The DA distance was calculated using the DISPAN programme and a unrooted UPGMA dendrogram was constructed from DA distance matrix. The correspondence analysis of allele frequencies is also used as an independent method to assess the level of genetic affinity among the different population groups. Relative amount of the gene flow was assessed using the Harpending and Ward's method of regression analysis (Harpending and Ward, 1982).

Significant departures from Hardy-Weinberg equilibrium expectations were observed in 20 of 300 comparisons. However, 8 of 10 populations deviated from the HWE at D1 locus. Ignoring the D1 locus, only 12 loci in different populations showed deviation, which is well below the rate one would expect by chance alone (5%). Similar deviation at the D1 locus has been observed in some other studies also and requires further investigations. Parsees differed at a number of pair-wise chi-square comparisons at different loci.

All loci were polymorphic in the populations studied and a range of insertion frequencies was observed at different loci. The overall pattern of allele frequency variation at different loci is extensive and comparable to other Indian and European studies. Single locus-by-locus descriptions are probably not very useful due to stochastic variation. A snap shot of genetic variation at six representative loci is given in figure 2. In this figure, TPA25 insertion frequency varies from 43% (Jat Sikh) to 61% (Maratha). Brahmins from western India show a relatively lower frequency of FXIIIB (0.203). For the APO locus, most populations had insertion frequencies above 70%. Parsee showed very low frequency of PV92 insertion (16%), which is towards the lower range found in the European populations (18-25%) (Stoneking et al., 1997; Roy-Engel et al., 2001; Watkins et al., 2003).

The average heterozygosity for each locus was substantial, with several values approaching the maximum attainable value of 0.5 for a bi-allelic locus. The average population heterozygosity ranges from 0.338 (Parsee) to 0.426 (Scheduled castes) for all loci. North Indian populations showed slightly lower of G_{ST} (0.048) compared to Western Indian (0.060), though the differences are statistically non-significant. Overall, the level of genetic diversity (G_{ST}) at all loci in the present set of populations is moderate, as only 5.4 % of the population diversity is attributable at genomic level. This level of differentiation is slightly lower than that observed in populations from East and North (6.8%), and South India (8.3%) (Majumder

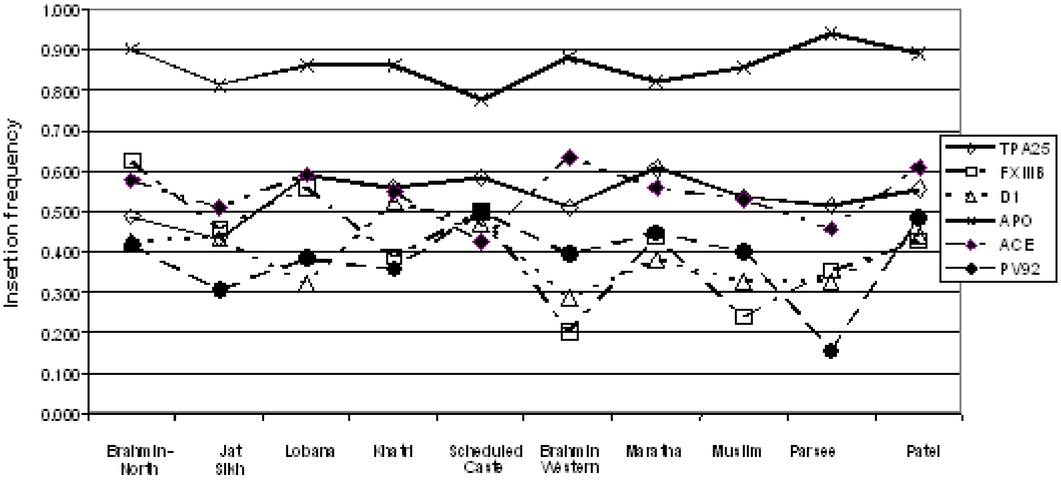

Fig. 2. Insertion frequencies of selected Alu loci

et al., 1999, Vishwanathan et al., 2003), but direct comparisons are not applicable as the number of loci are different in these analyses. The observed G_{ST} level is still nearly five times greater than using blood group, red cell enzyme and serum protein markers (Mastana and Papiha, 1994; Papiha, 1996). It provides evidence that an appreciable amount of inter-population differentiation is reflected in polymorphic *Alu* insertions. As there many more *Alu* polymorphisms in the human genome, the potential of these markers for use in studying population diversity and inter-relationships is remarkable.

Alu elements provide useful amounts of variation in evolutionary heritage, which is shown by the use of phylogenetic tree analysis. The figure 3 shows the un-rooted dendrogram produced from multidimensional DA distance matrix. Overall populations are differentiated according to the place of habitation and caste structure. It is interesting to note that two Brahmin groups do not show close genomic proximity. Instead, they are close to geographically proximal populations. Brahmins and Muslim from Western India also cluster together showing lower level of differences. Interestingly, the Parsee population, as expected, shows isolation and significant differences with other populations from Western India. Similar conclusions have been observed in other studies on conventional and molecular genetic markers (Mastana and Papiha, 1994; Papiha et al., 1996b; Rosenberg et al., 2006)

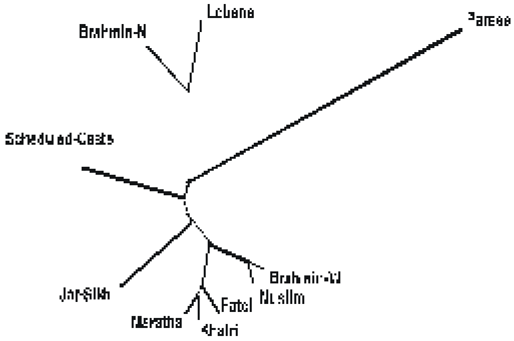

Fig. 3. Unrooted dendrogram of North and West Indian populations

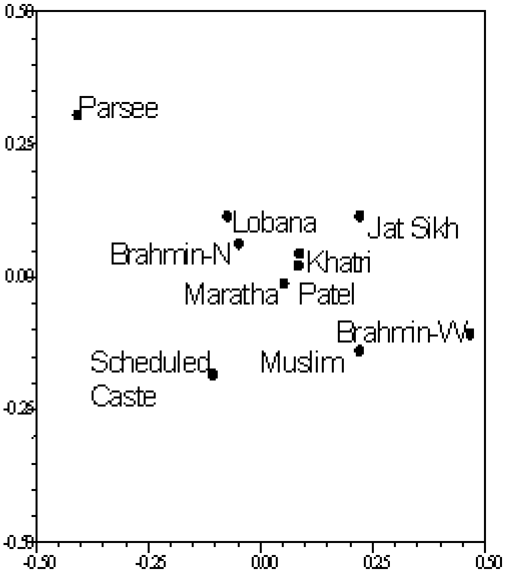

Fig. 4: Correspondence analysis plot of studied populations

Correspondence analysis was employed as an alternative method to assess genetic affinities of the population studied (Fig. 4) This figure illustrates that most populations form a loose central cluster and maintain geographical proximity, but Parsee are isolated, indicating their distinct origins and perhaps reproductive isolation.

The genetic drift and migration would determine whether a population group is incompletely isolated or distributed over a geographical space. This is supported by observed patterns of genetic diversities. These patterns are also generated by interactions with populations outside the set of populations under consideration. Harpending and Ward's method of regression of heterozygosity on genetic distance provides some evidence in support of the above hypothesis (Fig. 5). About half of populations (5) experienced lesser gene flow than predicted. These populations, like Brahmin-N, Lobana, Brahmin-West, Maratha and Parsee are below the regression line and therefore exhibit a lower level of gene flow or barriers to random mating. Interestingly, Scheduled Castes showed a higher level of gene flow along with Jat Sikhs, Patel and Muslims. Brahmin groups from North and West India along with other endogamous populations show lower levels of gene flow.

Overall *Alu i*nsertion results confirm that studied Indian populations are socially and geographically structured.

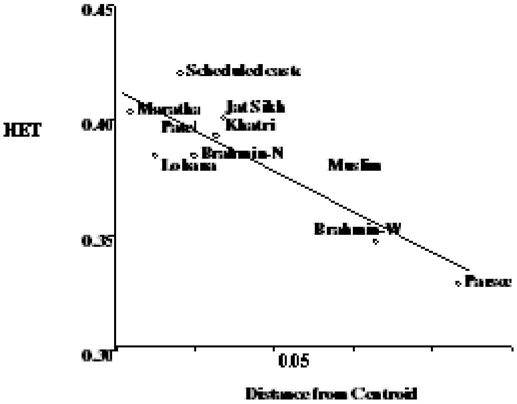

Fig. 5. Plot of Heterozygosity (Het) vs. Distance from the Centroid.

2. The Mystery of Sinhalese Origins: An *Alu* perspective

The origins of the Sinhalese population of Sri Lanka have attracted the attention of many anthropologists and geneticists. There are a number of theories about the origins but none of these provides conclusive proof. One theory is the legend of Prince Vijaya and the colonisation of Sri Lanka by a party of 701 individuals from north Western India, arriving at the Tambapani beach in the north western Sri Lankan province in 543 B.C (Saha, 1988; Kshatriya, 1995; Papiha et al., 1996; Papiha and Mastana, 1999).

Prince Vijaya is reported to be the grandson of Princess Kalinga of Bengal. The son of Princess Kalinga, Sihabahu is reported to have had 32 sons, all born in the West of India, Gujarat (Kshatriya, 1995). The eldest of the 32 sons was Prince Vijaya. It is hypothesised that Vijaya's party may have individuals from different parts of India who inhabited Sri Lanka and genetically contributed to the contemporary Sinhalese population. Prince Vijaya also married a daughter of the King of Madura (Tamilnadu). Many individuals travelled with her to Sri Lanka, therefore leading to theories that the Sinhalese have their ancestry in Tamil/South Indian populations. Many researchers have attempted to untangle the mystery of the Sinhalese origins as they seem to have genomic contributions from many areas of India (Saha, 1988; Kshatriya, 1995; Papiha et al., 1996; Papiha and Mastana, 1999) but results with conventional systems are contradictory and partial. New genetic markers may be able to provide a perspective on the origin of the Sinhalese.

In order to address these, we analysed the above mentioned 30 *Alu* polymorphisms in a sample of 121 Sinhalese collected from Colombo, Sri Lanka (Papiha et al., 1996b; Papiha and Mastana, 1999). In addition, *Alu* frequency data from Bengali (89) and Tamil (101), North and Western Indian populations (from the above study) were used for evaluation of genetic variation, affinities and genetic admixture. Overall pattern of allele frequencies is comparable to Indian populations but significant differences were observed at number of loci. Overall pattern of genetic relationships points towards substantial Bengali contribution as shown in DA distance derived dendrogram (Fig. 6) and admixture analyses.

A number of genetic admixture calculations were carried out using Tamil, Bengali, Gujarati (Patel), and Punjabi as parental populations. Admixture calculations were performed using two different methods- point estimates and maximum likelihood method. In all combinations, Bengali population seems to have higher contributions 57.49% (95%CI 36.89-78.59) compared to 42.51% (95%CI 0.7 - 9.15%) of Tamils by point estimates. Maximum likelihood method increased the Bengali contribution to 88.07% (95%CI 0.1-100%). When three parental populations were used Bengali contribution remained strong (50-66%) followed by North Western (20-23%) and rest contributed by Tamil. It should be pointed out that these results are based on relatively small number of loci (30 *Alus*), but the contributions are comparable to classic/conventional, mini-satellite and microsatellite systems (Papiha and Mastana, 1999). Overall analyses demonstrate that *Alu* insertion polymorphisms though only bi-allelic can be used to evaluate the admixture proportions.

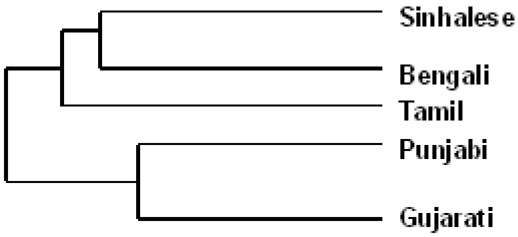

Fig. 6. UPGMA dendrogram of five populations of Indian Subcontinent.

3. *Alu* Polymorphisms: Forensic Applications

As *Alu* insertion polymorphisms are easy to analyse, identical by descent, their forensic usage have been proposed. There are limitations as these polymorphisms are only bi-allelic, difficult to multiplex with standard technologies and have high match probabilities at individual loci. We also examined the usage of *Alu* polymorphisms for forensic purposes in Indian populations. One of the aims of these analyses was to evaluate their suitability and to address substructure issues in forensic analyses. In the analysis of 10 populations from North and West India, it was observed that individual *Alu* match probabilities are relatively high (vary from 35 to 80%) but when all loci are combined it provides the low and conclusive match probabilities with high likelihood ratios which can be used in forensic analyses (Table 1). D1 locus was ignored in these calculations as it deviated significantly from HWE in a number of populations. There are some population specific differences but these are statistically non-significant. Overall, these results demonstrate that *Alus* are useful markers for forensic analyses and can be used in conjunction with other genetic markers like STRs and SNPs. Additional *Alu* insertion polymorphisms (20-25) will be required to provide match probabilities similar to STR loci.

Table 1: Forensic calculations for different populations

Population	*Combined Match Probability*	*Likelihood Ratio 1 in …*
Brahmin-N	1.13E-10	8859535282
Khatri	9.87E-11	10134990865
Scheduled Castes	3.53E-10	2830338360
Lobana	8.86E-11	11290083313
Jat Sikh	6.24E-11	16027186255
Brahmin-W	5.74E-11	17426158582
Maratha	9E-11	11106867825
Patel	1.53E-10	6541084968
Muslim	1.03E-10	9742888092
Parse	1.34E-10	7478191733

CONCLUSIONS

Since data of any particular kind are often too fragmentary to enable reconstruction of a composite picture of the people of any large geographical area, a multi-disciplinary approach are warranted. The above investigations show that geographical proximity, ethno-history, biosocial and cultural affiliation all seem to be important determinants of the genetic affinities among the populations of the Indian subcontinent. However, one should be careful about drawing conclusions about the observed diversity because differences might be attributed to the number and type of genetic systems analysed and the application of multivariate methods. Since the genetic differentiation in the regional and caste/tribal populations of the Indian subcontinent is low to moderate, although large enough, selection of patients and controls from appropriate populations for genetic epidemiology studies should be done with the help of Anthropologists to avoid any stochastic error that might generate erroneous results.

ACKNOWLEDGEMENTS

This paper is dedicated to the memory of my Guru, Prof. Surinder Papiha, who was a great mentor and guide. Sincere thanks to all research participants and Loughborough University for research funds to carry out this work.

REFERENCES

Agrawal, S., Muller, B., Bharadwaj, U., Bhatnagar, S., Sharma, A., Khan, F. and Agrawal, S.S.: Microsatellite variation at 24 STR loci in three endogamous groups of Uttar Pradesh, India. *Hum. Biol.*, **75**: 97-104 (2003).

Bamshad, M., Kivisild, T., Watkins, W.S., Dixon, M.E., Ricker, C.E., Rao, B.B., Naidu, J.M., Prasad, B.V., Reddy, P.G., Rasanayagam, A., Papiha, S.S., Villems, R., Redd, A.J., Hammer, M.F., Nguyen, S.V., Carroll, M.L., Batzer, M.A. and Jorde L.B.: Genetic evidence on the origins of Indian caste populations. *Genome Res.*,**11:** 994-1004 (2001).

Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A. and Jorde, L.B.: Human population genetic structure and inference of group membership. *Am J Hum. Genet.*, **72**: 578-589 (2003)

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N.P., Roychoudhury, S. and Majumder, P.P.: Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res***.,** **13**: 2277-2290 (2003)

Batzer, M.A., Arcot, S.S., Phinney, J.W., Alegria-Hartman, M., Kass, D.H., Milligan, S.M., Kimpton, C., Gill, P., Hochmeister, M., Ioannou, P.A., Herrera, R.J., Boudreau, D.A., Scheer, W.D., Keats, B.J., Deininger, P.L. and Stoneking, M.: Genetic variation of recent *Alu* insertions in human populations. *J. Mol. Evol*., **42**: 22-29. (1996)

Batzer, M.A., Deininger, P.L.: *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.*, **3**: 370-379(2002)

Bhasin, M.K., Walter, H. and Danker-Hopfe, H.: *People of India: An Investigation of Biological Variability in Ecological, Ethno-Economic and Linguistic Groups.* Kamla Raj Enterprises, Delhi (1994).

Bhasin, M.K. and Walter, H.: *Genetics of Castes and Tribes of India.* Kamla Raj Enterprises, Delhi (2001).

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L.: High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**: 455-457 (1994)

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W.S., Henke, J., Makalowski, W., Jorde, L.B., Deininger, P.L. and Batzer M.A.: Large-scale analysis of the *Alu* Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J. Mol. Biol.,* **311**: 17-40 (2001).

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

Chakraborty, R.: Genetic profile of cosmopolitan populations: effects of hidden subdivision. *Anthropol. Anz.,* **48**: 313-331 (1990).

Clark, A.G., Nielsen, R., Signorovitch, J., Matise, T.C., Glanowski, S., Heil, J., Winn-Deen, E.S., Holden, A.L. and Lai, E.: Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.,* **73**: 285-300 (2003).

Cordaux, R., Saha, N., Bentley, G.R., Aunger, R., Sirajuddin, S.M. and Stoneking, M.: Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.,* **11:** 253-64 (2003).

Das, B., Ghosh, A., Chauhan, P.S. and Seshadri, M.: Genetic polymorphism study at four minisatellite loci (D1S80, D17S5, D19S20, and APOB) among five Indian population groups. *Hum. Biol.,* **74**: 345-61 (2002).

Das, K. and Mastana, S.S.: Genetic variation at three VNTR loci in three tribal populations of Orissa, India. *Ann. Hum. Biol.,* **30:** 237-249(2003).

Garrigan, D. and Hammer, M.F.: Reconstructing human origins in the genomic era. *Nat. Rev. Genet.,* **7**: 669-680 (2006).

Gutala, R., Carvalho-Silva, D.R., Jin, L., Yngvadottir, B., Avadhanula, V., Nanne, K., Singh, L., Chakraborty, R. and Tyler-Smith, C.: A shared Y-chromosomal heritage between Muslims and Hindus in India. *Hum. Genet.,* **120**: 543-551 (2006)

Harpending, H. and Ward, R.H.: Chemical systematics and human populations. pp 213-256, In: *Biochemical Aspects of Evolutionary Biology*. M. H. Nitechi (Ed.).University of Chicago Press, Chicago (1982).

Indian Genome Variation Consortium. The Indian Genome Variation database (IGVdb): a project overview. *Hum. Genet*., **118**: 1-11 (2005).

Jobling, M.A. and Tyler-Smith, C.: The human Y chromosome: An evolutionary marker comes of age. *Nat. Rev. Genet.,* **4**: 598-612 (2003).

Karve, I.: *Hindu Society: An Interpretation.* Deccan College Post Graduate and Research Institute. Pune (1961).

Kashyap, V.K., Guha, S., Sitalaximi, T., Bindu, G.H., Hasnain, S.E. and Trivedi, R.: Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. *BMC Genet.,* **7**: 28 (2006).

Kashyap, V.K., Ashma, R., Gaikwad, S., Sarkar, B.N. and Trivedi, R.: Deciphering diversity in populations of various linguistic and ethnic affiliations of different geographical regions of India: analysis based on 15 microsatellite markers. *J. Genet.,* **83**: 49-63 (2004).

Kivisild, T., Bamshad, M.J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W.S., Dixon, M.E., Papiha, S.S., Mastana, S.S., Mir, M.R., Ferak, V. and Villems, R.: Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.,* **9**: 1331-1334 (1999)

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.V., Stepanov, V., Golge, M., Usanga, E., Papiha, S.S., Cinnioglu, C., King, R., Cavalli-Sforza, L., Underhill, P.A. and Villems, R.: The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.,* **72**: 313-332 (2003).

Kshatriya, G.K.: Genetic affinities of Sri Lankan populations. *Hum. Biol.,* **67**: 843-66 (1995).

Majumder, P.P., Roy, B., Banerjee, S., Chakraborty, M., Dey, B., Mukherjee, N., Roy, M., Thakurta, P.G. and Sil, S.K.: Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. *Eur. J. Hum. Genet.,* **7**: 435-446 (1999).

Malhotra, K.C.: Morphological composition of the people of India. *J. Hum. Evol.,* **7**: 45-63 (1978).

Mastana, S.S., Murry, B., Sachdeva, M.P., Das, K., Young, D., Das, M.K. and Kalla, A.K.: Genetic variation of 13 STR loci in the four endogamous tribal populations of Eastern India. *Forensic Sci. Int.,* May 6; [Epub ahead of print] (2006).

Mastana, S.S. and Papiha, S.S.: Genetic structure and microdifferentiation among four endogamous groups of Maharashtra, western India. *Ann. Hum. Biol.,* **21**: 241-262 (1994).

Mastana, S.S., Reddy, P.H., Das, M.K., Reddy, P. and Das, K.: Molecular genetic diversity in 5 populations of Madhya Pradesh, India. *Hum. Biol.,* **72**: 499-510 (2000).

Mastana, S. and Singh, P.P.: Population genetic study of the STR loci (HUMCSF1PO, HUMTPOX, HUMTHO1, HUMLPL, HUMF13A01, HUMF13B, HSFESFPS and HUMVWA) in North Indians. *Ann. Hum. Biol*., **29**: 677-684 (2002).

Mourant, A.E., Kopec, A.C. and Domaniewska-sobczak, K.: *The Distribution of the Human Blood Groups and Other Polymorphisms.* Oxford University Press, London (1976)

Nei, M. and Roychoudhury, A.K.: *Human Polymorphic Genes: World Distribution.* Oxford University Press, New York (1988).

Papiha, S.S.: Genetic variation in India. *Hum. Biol.,* **68**: 607-628 (1996).

Papiha, S.S. and Mastana, S.S.: Classical to molecular polymorphisms. pp1-21, In: *Genomic Diversity: Applications in Human Population Genetics*. S.S. Papiha, R. Deka and R. Chakraborty (Eds.). Kluwer Academic/Plenum Publishers, New York (1999)

Papiha, S.S., Mastana, S.S. and Jayasekara, R.: Genetic variation in Sri Lanka. *Hum. Biol.,* **68**: 707-737 (1996a).

Papiha, S.S., Mastana, S.S., Purandare, C.A., Jayasekara, R. and Chakraborty, R.: Population genetic study of three VNTR loci (D2S44, D7S22, and D12S11) in five ethnically defined populations of the Indian subcontinent. *Hum. Biol.,* **68**: 819-835 (1996).

Ranjan, D., Trivedi, R., Vasulu, T.S. and Kashyap VK.: Geographic contiguity and genetic affinity among five ethnic populations of Manipur, India: further molecular studies based on VNTR and STR loci. *Ann. Hum. Biol.,* 30: 117-31 (2002).

Rosenberg, N.A., Mahajan, S., Gonzalez-Quevedo, C., Blum, M.G., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L., Zapata, G., Weber, J.L., Belmont, J.W. and Patel P.I.: Low Levels of Genetic Divergence across Geographically and linguistically

Diverse Populations from India. *PLoS Genet.,* **2(12):** 2052-2061 (2006).

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W.: Genetic structure of human populations. *Science,* **298**: 2381-2385 (2002).

Roychoudhury, S., Roy, S., Basu, A., Banerjee, R., Vishwanathan, H., Usha Rani, M.V., Sil, S.K., Mitra, M. and Majumder, P.P.: Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum. Genet*., **109**: 339-350 (2001).

Roy-Engel, A.M., Carroll, M.L., Vogel, E., Garber, R.K., Nguyen, S.V., Salem, A.H., Batzer, M.A. and Deininger, P.L.: *Alu* insertion polymorphisms for the study of human genomic diversity. *Genetics,* **159**: 279-290 (2001).

Sachdeva, M.P., Mastana, S.S., Saraswathy, K.N., Elizabeth, A.M., Chaudhary, R. and Kalla, A.K.: Genetic variation at three VNTR loci (D1S80, APOB, and D17S5) in two tribal populations of Andhra Pradesh, India. *Ann. Hum. Biol.,* **31**: 95-102 (2004).

Saha, N.: Blood genetic markers in Sri Lankan populations—reappraisal of the legend of Prince Vijaya. *Am. J. Phys. Anthropol.,* **76**: 217-25 (1988).

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., Metspalu, M., Villems, R. and Kashyap, V.K.: A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc. Natl. Acad. Sci. U S A,* **103**: 843-848(2006).

Salisbury, B.A., Pungliya, M., Choi, J.Y., Jiang, R., Sun, X.J. and Stephens, J.C.: SNP and haplotype variation in the human genome. *Mutat. Res.,* **526:** 53-61 (2003).

Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.E., Lin, A.A., Mitra, M., Sil, S.K., Ramesh, A., Usha Rani, M.V., Thakur, C.M., Cavalli-Sforza, L.L., Majumder, P.P. and Underhill, P.A.: 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.,* **78**: 202-221(2006).

Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S., Saha, N., Jenkins, T., Tahir, M.A., Deininger, P.L. and Batzer, M.A.: *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.,* **7**: 1061-71(1997).

Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V.K., Bhaskar, L.V., Reddy, B.M., Reddy, A.G. and Singh L. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.,* **7**: 42 (2006)

Vishwanathan, H., Edwin, D., Usharani, M.V. and Majumder, P.P.: Insertion/deletion polymorphisms in tribal populations of southern India and their possible evolutionary implications. *Hum. Biol.,* **75**: 873-887 (2003).

Watkins, W.S., Rogers, A.R., Ostler, C.T., Wooding, S., Bamshad, M.J., Brassington, A.M., Carroll, M.L., Nguyen, S.V., Walker, J.A., Prasad, B.V., Reddy, P.G., Das, P.K., Batzer, M.A. and Jorde, L.B.: Genetic variation among world populations: inferences from 100 *Alu* insertion polymorphisms. *Genome Res.,* **13**: 1607-1618 (2003)

Zerjal, T., Pandya, A., Thangaraj, K., Ling, E.Y., Kearley, J., Bertoneri, S., Paracchini, S., Singh, L. and Tyler-Smith, C.: Y-chromosomal insights into the genetic impact of the caste system in India. *Hum. Genet.,* 2006; [Epub ahead of print]

KEYWORDS Molecular Anthropology. DNA. STR. *Alu*. mtDNA. Y Chromosome

ABSTRACT Molecular anthropology - the study of human genetic polymorphisms - is fast and ever-growing branch of anthropology that holds a great promise for both past and future. Indian subcontinent with its remarkable environmental, geographical, morphological, genetic, cultural and linguistic diversity provides immense promise for anthropological investigations to address origins of its people, population movements, analysis of genetic architecture in human diseases and gene-environment interactions. Several types of DNA polymorphisms have been discovered in the human genome and have been found to be very productive in addressing molecular anthropological questions. This paper traces the usage of some of these polymorphisms in the study of populations of India. Using a battery of *Alu* insertion polymorphisms, genetic structure and affinities of North and Western Indian populations are evaluated. *Alu* polymorphisms are also used to investigate the origin of the Sinhalese of Sri Lanka. In addition, forensic applications of *Alu* polymorphisms are demonstrated. These results suggest that there is fundamental genomic unity among Indian populations and geographic location, caste affiliation and migration/gene flow contribute significantly to the observed genetic variation in contemporary populations.

Author's Address: **Dr. Sarabjit Mastana,** Human Genetics Labortary, Department of Human Sciences, Loughborough University, Loughborough LE11 3TU, UK
Telephone: +44-1509-223041, *Fax:* +44-1509-223940, *E-mail:* S.S.Mastana@LBORO.AC.UK

Anthropologist Special Volume No. **3:** 373-383 (2007)
Anthropology Today: Trends, Scope and Applications
Veena Bhasin & M.K. Bhasin, *Guest Editors*