American Journal of Human Biology

Genetic analysis of novel *Alu* insertion polymorphisms in selected Indian populations.

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

<u>**Genetic analysis of novel *Alu* insertion polymorphisms in selected Indian populations.**</u>

<u>**Susie Laybourn[1], Elizabeth Claire Akam[1], Nick Cox[1], Puneetpal Singh[2], Sarabjit S Mastana[1]**</u>

1. Human Genomics Lab, School of Sport, Exercise and Heath Sciences, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom. General telephone: +44 (0)1509 263171
2. Department of Human Genetics, Punjabi University, Patiala, Punjab, India.

•Number of text pages, plus bibliography, number of tables, figures, graphs, and charts

Short title - New *Alu* insertion elements in Indian Populations

Address correspondence to this author at the Human Genomics Lab: Dr Sarabjit Mastana, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom : Phone +44 (0)1509 223041: Fax : +44 (0)1509 226301; E mail s.s.mastana@lboro.ac.uk

Or

Dr Puneetpal Singh, Department of Human Genetics, Punjabi University, Patiala, India: Email singh.puneet@rediffmail.com

Abstract

Objectives: Indian subpopulations (Chenchu, Koya and Lobana Sikh) were analysed at the genetic level for 12 *Alu* polymorphisms. These markers were then utilised to establish levels of genetic identity between the Indian populations and more widely between the Indian populations and a European population.

Methods: Previously collected blood samples were extracted using the phenol-chloroform method. The samples were utilised as templates for PCR using *Alu* specific primers and then analysed by agarose gel electrophoresis for the presence and absence of the approximately 300 bp insertion. Allele frequencies were calculated by the gene counting method and were tested for Hardy-Weinberg equilibrium, heterozygosities, inbreeding coefficient and G_{ST} to assess the level of genetic differentiation.

Results: All of the *Alu* loci were polymorphic in the three Indian populations studied and their average observed heterozygosity ranged from 0.294 (Punjabi) to 0.357 (Koya). Allele and genotype frequency variation at 2b, 9a and ACE loci led to statistically significant pairwise differences among the three study populations. Overall population heterogeneity was observed for 7 out of 12 *Alu* polymorphisms.

Conclusion: The overall results show that these Indian samples, though displaying significant genetic variation and differences among themselves, form an Indian cluster, which as expected, is distinct from European sample (Russian). As *Alus* are easily analysed and quantified by standard and cost effective methodologies this further reinforces their utility as an effective population genetic marker

Introduction

Alu insertion elements (*Alus*) represent the largest family of Short INterspersed Elements (SINEs) in humans and are named due to the presence of an *Alu*I recognition site within the characteristic 300 bp unit. The human genome contains more than 1 million *Alu* repeats, which account for ~10% of the total nuclear DNA (Batzer et al 1996; Batzer and Deininger 2002; Stewart et al 2011). *Alu* repeats are organised into families and subfamilies, one of which is human specific (HS). HS *Alus* were inserted into the genome after the human divergence from a common ancestor with chimps thus these only appear in humans (Batzer et al. 1996), with insertions sites being commonly found in non-coding regions.

*Alu*s have also been identified as highly stable due to their low insertion rate with only 100-200 new insertions occurring per millennia. These elements are also fixed due to lack of specific removal mechanisms in human genome (Batzer *et al*. 1996). This means individuals sharing *Alu* insertion polymorphisms have inherited from a common ancestor thus making the *Alu* insertion alleles identical by descent. So when using a number of *Alus* together a high forensic discrimination and phylogenetic analyses can be achieved (Mamedov at al., 2010) which is advantageous when analysing DNA for both ethnicity studies and forensic analyses.

India exhibits a large amount of genetic diversity for a country. Since the dispersal of humans from Africa, India has experienced large amounts of migration from many different populations resulting in a melange of ethnicities and cultural practices. In turn, this has led to both genetic divergences and convergences between sub populations within India ((Majumder et al. 1999; Basu et al. 2003; Sahoo et al. 2006; Reich et al. 2009; Mastana, 2014; ArunKumar et al 2015).

While there are many reports on *Alu* polymorphisms in different populations of the world (Batzer et al. 1996), studies from the Indian subcontinent are limited to relatively small number of loci and populations (Majumder et al. 1999; Tripathi et al. 2008; Kshatriya et al. 2012; Saini et al 2012). In order to extend and document genetic variation of *Alu* polymorphisms, we analysed 8 new *Alu* insertion polymorphisms together with 4 extensively studied *Alus* (ACE, TPA, APO, PV92) (Majumder et al. 1999; Mamedov et al. 2010) in three distinct populations from India (1 caste population from Punjab (Lobana Sikhs) and 2 tribal populations (Chenchu and Koya) from Andhra Pradesh (now Telangana). The eight new *Alu* loci were selected from published paper (Mamedov et al. 2010) based on the observed heterozygosity levels (medium to high) and their adaptability to a multiplex reaction.

Methods

Subjects and study protocol

Genomic DNA samples were collected from three populations of India, Lobana Sikh caste population from Punjab region of Northern India, and two tribal populations (Chenchu and Koya) from Andhra Pradesh in South India (Mastana et al 1991; Mastana et al 2013, Papiha et al. 1997). The three populations represent groups who have both cultural and geographical barriers to intermarriage (Papiha et al. 1997). Brief details of the study populations are given below.

Chenchu tribe; a scheduled tribe, are also known as Chenchuvara inhabit three areas of Andhra Pradesh (AP); Mahabub Nagar, Kurnool and Guntur (Singh 1997; Papiha et al.1997). They traditionally live in patriarchal and patrilineal nuclear families and have their own Chenchu language, which is part of the Dravidian family of languages (Bhasin et al 1994, Bhasin and Walter 2000, Singh 1997). Primarily described as "shy hunter- gatherers" in 1694, they work on the land and in forests collecting produce and more recently have become agricultural labourers. In 1971, Chenchus numbered around 18,000 but only a minority still live in the traditional subsistence lifestyle (Singh 1997).

Koya tribe; also known as Konda Rajulu are inhabitants of Andhra Pradesh and have a larger population size (359,799, as of the 1981 census) (Singh 1997). They speak Gondi which has links with Dravidian languages common in Southern India (Singh 1997). Koya people are primarily farmers but exist in several endogamous groups which have their own occupations, for example blacksmiths brass workers and basket makers (Papiha et al.1997). They live in villages in exogamous patrilocal clans and consanguinity between cousins along with uncles and nieces is common (Singh 1997).

Lobana Sikh is an endogamous now settled agricultural population of Punjab. In the recent past, they were a semi-nomadic tribe which traded in different commodities across India and abroad. Their population size is large (above 1 million) and they now profess to a range of occupations (Mastana 1989; Mastana et al. 1991; Mastana et al. 2013). Marriage patterns are strict and caste endogamy and clan exogamy is commonly practiced.

The Koya were sampled from Khamman District, and the Chenchu samples were collected from the Guntur and Hyderabad districts of Andhra Pradesh (Papiha et al 1997). Lobana samples were collected from Patiala and Kapurthala districts of Punjab (Mastana et al. 1991).

All participants were healthy unrelated individuals who participated in our ongoing genetic research among Indian populations. The research was approved by the Loughborough University Ethical Advisory Committee and all participants gave full written consent. The DNA was extracted using the organic method (Sambrook et al. 1989).

Genotyping method

Alu polymorphisms were analysed by PCR amplification using specific primers as detailed by Mamedov et al., (2010). Individual allele genotypes were determined based on the respective DNA amplicon sizes visualised after gel electrophoresis (Mamedov et al. 2010; refer also S-Table 1 and S-Table 2).

Statistical analysis

Allele frequencies were calculated by the gene counting method and were tested for Hardy-Weinberg equilibrium. Chi-square test was used to compare frequencies between populations using PowerMarker Version 3.25 (http://statgen.ncsu.edu/powermarker/).

There is a lack of published studies on the 8 *Alu* loci employed in this study among Global, European and Indian populations, only a few studies are available from

Russia therefore comparative allele frequencies representing a European population (Russian) were collected from previously published literature (Mamedov et al. 2010 and Litvinov et al. 2008; Refer also S-Table 3). This population was considered as an outgroup for Indian populations.

Results

The sample size, insertion allele frequency, HWE p value, observed and expected heterozygosity and inbreeding coefficient (f) are given in Table 1. All *Alu* loci were polymorphic in three populations studied. Hardy-Weinberg equilibrium (HWE) deviation was observed for 8 populations/loci combinations (APOA1 in Chenchu, 2b, 8b, 12 and ACE in Punjabi, 9a, 8b and 12 in Koya). Pairwise genotype comparisons using chi-square showed that the Chenchu differed significantly from Koya at 2b, 9a,11, TPA and APO loci (all p values <0.05); and from Punjabis at 2a, 2b, 4a, 9a,11 and ACE loci, while Koya differed significantly from Punjabis at 2b, 9a,and ACE loci. Overall population heterogeneity was observed at 7 out of 12 loci (2a, 4a, 2b, 9a, 8b, 11 and ACE), indicating significant differences in genotype frequencies in the different populations. Punjabi population had the lowest observed heterozygosity (0.294) compared to both tribal groups (0.329 for Chenchu and 0.357 for Koya) for this set of *Alu* loci. Marked heterozygosity differences were observed for, 2b, ACE and APO different populations. Inbreeding co-efficient (f) values also showed wide variation among the different populations (Table 1). The f values indicate that Punjabi population has substantial deficit of observed heterozygosity among different loci; 6 loci have heterozygosity below the average (0.294). Average G_{ST} value 0.033$\pm$ 0.032) suggest wide variation among loci and populations. Individual locus GST values varied from -0.004(*Alu* 12) to 0.107 (*Alu* 2b). A negative value of the GST was due to high similarity in Alu insertion allele frequency; while a high positive value for *Alu* 2b (0.107) and ACE (0.072) were due to substantial differences in allele frequencies of the study populations.

[INSERT TABLE 1 HERE]

Eight loci in this study are new and have not been studied in European, global and Indian populations previously so it was not possible to compare these to infer genetic diversity. However, a Russian population data set (sample size 90-252 for different loci) was available (Mamedov et al. 2010 and Litvinov et al. 2008) and this was used for genetic distance calculations and correspondence analysis. We calculated Nei's unbiased DA distance based on 12 loci and reduced the multidimensional matrix into a dendrogram by UPGMA method (Figure 1 and S-Table 3). Bootstrap values (out of 100) for this dendrogram are high giving confidence in use of *Alu* polymorphisms in understanding of population relationships.

[INSERT FIGURE 1 HERE]

In this figure, it is clear that there are distinct differences between the populations, Chenchu and Koya tribes belonging to South India form a close cluster which is joined by Lobana Sikhs and finally by Russians. Similar results were also obtained with correspondence analysis (not shown).

On developing the data set further, we assessed the use of these 12 loci for forensic and paternity purposes. The power of discrimination (for forensic calculation) was

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

high for all three populations analysed (>99.99%). Power of exclusion calculation for paternity testing showed that using these 12 loci one could exclude 42%(Lobana), 31% (Chenchu) and 24% (Koya) of the alleged fathers in a paternity investigation, which is typical for any two allele systems like SNPs and *Alus*.

Discussion

The analysis of *Alu* polymorphisms from three Indian populations highlights a level of genetic variation previously unreported with possible implications for the understanding of the evolution of genetic diversity among Indian populations. One should be cautious in interpreting the level and extent of genetic variation observed in these populations as the sample sizes are relatively small, although the sampling is homogeneous and well characterised. The results show that 8 new *Alu* loci are polymorphic in all populations with variable level of heterozygosity. There is paucity of data on new *Alu* polymorphisms to carry out comprehensive comparative analyses. Overall results of well-studied *Alu* loci like ACE, TPA, APO and PV92 are comparable to many studies from Indian subcontinent (Majumder et al. 1999; Veerraju et al. 2008; Tripathi et al. 2008; Kshatriya et al. 2012; Saini et al 2012).

Alu elements provide useful amounts of variation in evolutionary heritage. Koya and Chenchu share the similar environmental and geographical location but differ significantly at a number of *Alu* loci. This genetic diversity between the two groups could be attributed to their different ethnic origins (Koya: Gond tribe and Chenchu: proto-Australoid). This inference is also supported by other genetic studies on the same South Indian populations (Papiha et al 1997). Papiha et al. (1997) further studied both tribal populations using a battery of blood groups, red cell enzymes and serum proteins. They found that the two tribes were significantly genetically different even though they have close geographical proximity. Veerraju et al. (2008) studied five *Alu* insertions (ACE, TPA, APO, PV92 and D1) in Chenchu and Koya tribes and found significant differences between these tribes. In their extended analysis, they also showed Chenchus clustered with caste populations of Uttar Pradesh and not with their geographical neighbours from Southern India. The Lobana Sikh population, as expected, shows significant differences from the tribal populations due to different geographical location and different genetic origins.

Analysis of four previously studied loci (ACE, TPA, APO, PV92) using DA genetic distances showed the distinct positions of tribal groups (Chenchu and Koya). In the dendrogram (not shown) based on these 4 loci only, Chenchu and Koya were closest to each other while Lobanas and Russian formed a separate cluster. One needs to be cautious about these relationships as these are based on very small number of loci and small sample size. Analysis of 12 loci, though better predictor of population relationships (based on bootstrap values) among the study populations, should also be interpreted with caution.

Overall our results demonstrate that these Indian samples, although displaying significant genetic variation and differences among themselves, still form an Indian cluster distinct from the European sample set (Russian) (Figure 1).These conclusions are strongly supported by other studies on STR, *Alus*, mtDNA, Y chromosome and other expansive genome-wide evaluations of Indian populations (Basu et al. 2003; Sahoo et al 2006; Reich et al. 2009; Chaubey et al 2011; Mastana, 2014; ArunKumar et al 2015). Further populations studies are required on these *Alus* to deconvolute the level and extent of genetic variation and population origins.

Conclusion

Overall, results suggest that these Indian populations show significant allelic and genotypic variation which is patterned on geographical and ethnic origins.

Author Contribution

SL, SM, and EA analysed the data and drafted the manuscript. EA, SM, and SL designed the study, and directed implementation and data collection. SL, EA, and NC collected the data, and PPS provided necessary sample support. All authors read and agreed to this manuscript.

References

ArunKumar G, Tatarinova TV, Duty J, Rollo D, Syama A, et al. (2015) Genome-wide signatures of male-mediated migration shaping the Indian gene pool. J Hum Genet. 60: 493-499.

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

Batzer MA, Arcot SS, Phinney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, Gill P, Hochmeister M, Ioannou PA, Herrera RJ, Boudreau DA, Scheer WD, Keats BJ, Deininger PL, Stoneking M.1996. Genetic variation of recent Alu insertions in human populations. J Mol Evol 42:22-29.

Batzer MA, Deininger PL.2002. Alu repeats and human genomic diversity. Nat Rev Genet 3:370-379.

Bhasin MK, Walter H, Danker-Hopfe. 1994. People of India: An investigation of biological variability in ecological, ethno-economic and linguistic groups. Delhi, Kamla-Raj Enterprises. p41-46

Bhasin MK, Walter H. 2000. Genetics of castes and tribes of India. Delhi, Kamla-Raj Enterprises. p65-73.

Chaubey G, Metspalu M, Choi Y, Magi R, Romero IG, et al. (2011) Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol 28: 1013-1024.

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, and others. 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72:313-32.

Kshatriya GK, Aggarwal A, Khurana P, Italia YM. 2011.Genomic congruence of Indo-European speaking tribes of western India with Dravidian-speaking populations of southern India: A study of 20 autosomal DNA markers. Ann Hum Biol. 38:583-591.

Litvinov S, Kutuev I, Yunusbayev B, Khusainova R, Valiev R, Khusnutdinova E. 2008. Alu insertion polymorphisms in populations of the South Caucasus. Balkan Journal of Medical Genetics 11/2: 25-30

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

Mamedov IZ, Shagina IA, Kurnikova MA, Novozhilov SN, Shagin DA, Lebedev YB. 2010. A new set of markers for human identification based on 32 polymorphic Alu insertions. Eur J Hum Genet. 18:808-814

Mastana SS. 1989. Age at menarche among the Lobanas of Punjab (India). Acta Med. Auxol. 21:63-65

Mastana SS. 2014.Unity in diversity: an overview of the genomic anthropology of India. Ann Hum Biol. 41:287-299.

Mastana SS, Garlick JP, Papiha SS. 1991, Serogenetic studies of the Lobanas in North-West India. Anthrop. Anz. 49:177-186

Mastana SS, Kaur A, Hale R, Lindley MR. 2013. Influence of glutathione S-transferase polymorphisms (GSTT1, GSTM1, GSTP1) on type-2 diabetes mellitus (T2D) risk in an endogamous population from north India. Mol Biol Rep. 40:7103-7110.

Papiha SS, Singh BN, Lanchbury JS, Mastana SS, Rao YS. 1997.Genetic study of the tribal populations of Andhra Pradesh, south India. Hum Biol. 69:171-199.

PowerMarker Version 3.25 (http://statgen.ncsu.edu/powermarker/).

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461(7263):489-494

Saini JS, Kumar A, Matharoo K, Sokhi J, Badaruddoza, Bhanwer AJ. 2012. Genomic diversity and affinities in population groups of North West India: an analysis of Alu insertion and a single nucleotide polymorphism. Gene. 511:293-299.

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

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, pp. 9.14–9.23.

Singh K. 1997. The scheduled tribes. In: People of India. Oxford: Oxford University Press. p213.

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HY, Lee WP, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT; 1000 Genomes Project. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 7(8):e1002236. doi: 10.1371/journal.pgen.1002236.

Tripathi M, Tripathi P, Chauhan UK, Herrera RJ, Agrawal S. 2008. Alu polymorphic insertions reveal genetic structure of north Indian populations. Hum Biol 80:483-499.

Veerraju P, Demarchi DA, Lakshmi N, Rao TV. 2008. Insertion/deletion polymorphisms in Indian tribal populations. International journal of human genetics; Int.J.Hum.Genet. 8:75-83

Supporting Information

Supporting Information - Methods

[INSERT S-TABLE 1 HERE]

[INSERT S-TABLE 2 HERE]

[INSERT S-TABLE 3 HERE]

Table 1: Sample Size, Allele frequency, HWE, Heterozygosities, inbreeding coefficients and G_{ST} statistics for 12 loci among 3 Indian populations

Marker	Sample size			fAlu			HWE p			H_o			H_E			f			G_{ST}
	Chenchu	Koya	Lobana Sikh	Chenchu	Koya	Lobana Sikh	Chenchu	Koya	Lobana Sikh	Chenchu	Koya	Lobana Sikh	Chenchu	Koya	Lobana Sikh	Chenchu	Koya	Lobana Sikh	
2a	64	56	62	0.695	0.527	0.468	0.225	0.187	0.080	0.359	0.411	0.387	0.424	0.499	0.498	0.160	0.185	0.230	0.036
4a	64	56	63	0.102	0.152	0.254	0.066	0.461	0.198	0.141	0.232	0.317	0.182	0.257	0.379	0.237	0.107	0.170	0.030
2b	65	56	60	0.423	0.607	0.208	0.747	0.841	0.000	0.508	0.464	0.150	0.488	0.477	0.330	-0.032	0.036	0.551	0.107
9a	65	55	53	0.169	0.091	0.330	0.447	0.012	0.168	0.308	0.109	0.358	0.281	0.165	0.442	-0.087	0.348	0.199	0.051
8b	65	54	53	0.677	0.722	0.481	0.903	0.055	0.002	0.431	0.296	0.283	0.437	0.401	0.499	0.023	0.270	0.441	0.043
10	65	56	60	0.077	0.134	0.117	0.502	0.022	0.138	0.154	0.161	0.167	0.142	0.232	0.206	-0.076	0.315	0.199	-0.006
11	65	53	52	0.808	0.679	0.606	0.634	0.360	0.090	0.292	0.491	0.365	0.311	0.436	0.478	0.067	-0.116	0.244	0.017
12	65	54	54	0.138	0.167	0.157	0.798	0.014	0.000	0.246	0.185	0.130	0.239	0.278	0.265	-0.024	0.342	0.518	-0.004
ACE	65	56	63	0.731	0.625	0.413	0.417	0.285	0.001	0.354	0.536	0.286	0.393	0.469	0.485	0.108	-0.134	0.417	0.072
TPA	65	56	63	0.431	0.589	0.476	0.591	0.425	0.170	0.523	0.536	0.413	0.490	0.484	0.499	-0.059	-0.098	0.181	0.018
APO	59	55	54	0.805	0.682	0.778	0.002	0.788	0.066	0.186	0.418	0.259	0.314	0.434	0.346	0.413	0.045	0.259	0.012
PV92	59	52	61	0.398	0.548	0.418	0.728	0.440	0.218	0.458	0.442	0.410	0.479	0.495	0.487	0.054	0.117	0.166	0.016
Average										**0.330**	**0.357**	**0.294**	**0.348**	**0.386**	**0.409**	**0.065**	**0.118**	**0.298**	**0.033**

fAlu, Insertion allele frequency

p, Pearson chi square p value

H_o, Observed heterozygosity

H_E, Expected heterozygosity

f, Inbreeding coefficient

Figure 1: UPGMA dendrogram of 4 populations based on 12 *Alu* loci