J Hum Genet (2000) 45:207–211

ORIGINAL ARTICLE

Anjana Saha · Ramesh Bamezai

Detection of genetic variation in Indian population groups using a novel minisatellite probe and finding relationships through tree construction

Received: December 27, 1999 / Accepted: March 13, 2000

Abstract Genetic variation in *Hae*III-digested genomic DNA samples from different individuals belonging to population groups from Bengal, Uttar Pradesh (UP), Punjab, and South India was assessed at hypervariable loci, using a minisatellite probe, pBA1.2 (accession number, AF 157691), the repeat unit of which was 24 mer long and rich in G-bases. Comparison of DNA profiles between individuals showed a very low probability of band sharing, which ranged from 0.18 to 0.24. A dendrogram, based on Nei's genetic distance, constructed by the neighbor-joining method, showed the formation of separate clusters by both South Indian and non-Indian samples, whereas the construction of a dendrogram based on the Unweighted pair group method arithmetic average (UPGMA) method with Jaccard's similarity coefficient at the individual level led to the formation of several small clusters which were interleaved; also, the subgroups for each of the populations were intermingled with the subgroups for the other populations. A separate analysis was carried out to check the consistency of the proximity between different individuals forming a cluster and between those individuals who were in the vicinity of two clusters. The dendrograms thus obtained did not change the relationship between the individuals from all the populations studied. Despite the distinct clustering observed in the population group comparison, a probable admixture was reflected in the finding that some individuals belonging to one population group were dispersed or embedded within a cluster generated by the individuals of another population group, when a minute dissection of the data for generating a tree at the individual level was carried out.

Key words Minisatellite · Polymorphism · Dendrogram · India

A. Saha · R. Bamezai (✉)
Human Genetics Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi-110067, India
Tel. +91-11-6107676; Fax +91-11-6187338; 6165886; 6169962
e.mail: bamezai@jnuniv.ernet.in/bamezai@hotmail.com

Introduction

In the past decade, variation in the DNA sequence, mostly in the satellite region, has been exploited to carry out genome individualization (Jeffreys et al. 1985; Epplen 1988), which, in turn, has proved to be useful for establishing individuality in forensic (Gill and Werrett 1987; Helminen et al. 1988; Kasai et al. 1990) and immigration cases (Jeffreys et al. 1985; Jeffreys et al. 1986), for resolving paternity disputes (Ito et al. 1985), and for the analysis of pedigree and genetic disorders. Highly polymorphic markers, such as short tandem repeats (STRs) or microsatellites, have further proved to be useful for this purpose, as they are distributed at many loci and are amenable to automation (Budowle and Moretti 1999). A comparison of the frequency of these highly polymorphic markers among different populations contributes to the understanding of genetic relationships (Katsuyama et al. 1998) and the evolution of human populations. In a recent study, based on 100 restriction fragment length polymorphism (RFLP) markers, a tree inferred at the individual level was found to be relatively more consistent, showing the level of consistency between the tree and population affiliation to be relatively high, compared with the tree obtained with data from the mitochondrial genome (Mountain and Cavalli-Sforza 1997).

Here we present a comparative analysis of a multilocus DNA band profile generated by the use of a novel minisatellite probe, pBA1.2, in *Hae*III-digested genomic DNA samples from 100 individuals (95 Indians belonging to four regional population groups and 5 non-Indians). Our goal, to be inferred from the individual-specific DNA pattern, was to find the extent of the pattern shared between the four regional population subgroups within India and some of the representative non-Indian samples.

Subjects and methods

Population samples

A total of 100 samples were considered for the present study, which included 23 from Bengal, 28 from Uttar Pradesh (UP), 23 from Punjab, and 21 from South India. One sample from Ireland and 2 each from Africa and Korea, as easily available representative samples outside India, were also included in this study. Samples from these volunteers were collected after obtaining their consent.

Isolation of genomic DNA and fractionation

Total genomic DNA was obtained from peripheral blood cells by a method described elsewhere (Kunkel et al. 1977) in which phenol: chloroform: isoamylalcohol extractions following sodium dodecylsulfate (SDS) lysis and proteinase K treatment was carried out. Genomic DNA (8–10μg) was restrict digested with *Hae*III (as per the manufacturer's recommendations; Bangalore Genei, Bangalore, India), and the fractionated DNA fragments were run in 0.8% agarose gel in 0.5× TBE (80mM Tris, 40mM boric acid, and 2mM ethylenediaminetetraacetic acid [EDTA], pH 8.3) followed by their transfer to a positively charged nylon membrane (Boehringer Mannheim, GmbH Manheim, Germany) by capillary transfer for 16–24h (Southern 1975).

Generation of the probe and its nature

Unidirectional deletion in the cloned insert of pBA18 (obtained from pCMM86; Genome database [GDB], 168382, D17S74; Nakamura et al. 1988) for different time points was carried out, using Exonuclease III/S1 nuclease digestion. Aliquots, taken out at regular time intervals, were then incubated with S1 nuclease to allow polishing of the deleted ends, followed by self-ligations at 22°C overnight. Transformation was carried out with competent XL-1 Blue cells. Random selection of colonies led to the obtaining of a clone with an insert of 1.2kb.

A dideoxy method (Sanger 1981) of sequencing, using the Taq dideoxy terminator cycle sequencing kit (Perkin-Elmer) and an ABI 377 version 3.0 DNA sequencer (Applied Biosystems), was carried out to sequence part of pBA1.2. The partial sequence obtained was subjected to University of Wisconsin Genetics Computer Group (UWGCG) programs: REPEAT and FINDPATTERN analysis (Wisconsin package version 10, 1999; Madison, USA) to determine the presence of any repeat motif, which unraveled the presence of a tandem repeat of a 24-mer-long unit in an imperfect fashion, the core unit of which was found to be highly G-rich (GTGGG TGTGTTGGAGGGGGTGAGG); submitted to the National Center for Biotechnology Information (NCBI) databank under accession number AF157691.

Hybridzation, RFLP, and scoring of bands

A 1.2-kb insert of pBA1.2, obtained after double digestion of the plasmid with *Eco*RV and *Sac*I, was gel-purified and labeled with random hexamers and digoxigenin-2′-deoxy-uridine-5′-triphosphate (dig-dUTP) (Boehringer) (Sachdeva et al. 1995), to be used as a probe to carry out hybridization at 37°C for 16–20h in hybridization buffer containing labeled probe in a solution of 50% formamide, 6× standard saline citrate (SSC), 0.02% SDS, 0.1% N-laurylsarcosine (NLS), and 5% blocking reagent (Boehringer Mannheim). After hybridization, the blots were washed thrice with 2× SSC and 0.1% SDS at room temperature and thrice with 1× SSC and 0.1% SDS at 54°C. Finally, chemiluminescent detection (Sachdeva et al. 1997) was carried out with alkaline phosphatase enzyme conjugated with antibody; the substrate used was Disodium 3-(4 methoxyspiro{1,2-dioxetane-3,2′-(5′-[3.3.1.13,7]decan}-4-yl}phenylphosphate (CSPD).

The sizes of the bands generated with pBA1.2 were determined by measuring their mobility with respect to the DNA fragments of a known molecular size marker (lambda phage DNA predigested with *Hind*III) included in each run. After the deduction of band sizes from lumograms, a matrix consisting of binary codes was formed, where "1" designated the presence of a band, and "0" designated the absence of the same. These codes were subsequently used to run the GW BASIC program to calculate the values for the mean probability of band sharing (x), which gave the probability of finding the same DNA band in two different individuals (Bhat et al. 1995). The mean probability of identity ($\overline{X}$) was calculated by using the formula of x^f where "f" was the mean number of bands present.

Construction of dendrogram based on Nei's genetic distance at the population level and Jaccard's similarity coefficient at the individual level

The neighbor-joining method (Saitou and Nei 1987) was used to construct a dendrogram based on Nei's genetic distance (Nei 1978), calculated from the mean frequency of different DNA fragments, to analyze the relationship between the different population groups studied.

Phylogenetic analysis of the data was carried out using the sequential, agglomerative, hierarchical, and nested (SAHN) clustering method in the Numerical taxonomy and multivariate analysis system (NTSYS-pc) version 1.70 program Applied Biostatistics, (Rholf and Slice 1992) New York. From the matrix of binary codes of each individual, on the basis of the presence and absence of each and every DNA fragment, Jaccard's similarity matrix was constructed, and subsequently a dendrogram based on the UPGMA method (Sneath and Sokal 1973) of clustering was generated, depicting the formation of clusters at the individual level.

Results and discussion

RFLP analysis with pBA1.2

Analysis of a multilocus RFLP profile, at the individual level, of unrelated individuals belonging to four regions of India revealed polymorphic bands in the size range of 27.5 to 2.5 kb, with an average presence of 21.80 ± 0.34 bands per individual. The band-sharing probability among Indians ranged from 0.18 ± 0.0040 to 0.24 ± 0.0054, and the mean probability of identity ranged from 2.31×10^{-14} to 4.19×10^{-16} (Table 1). These results overlapped with the values for the probability of band sharing and identity obtained with another 624-bp-GC rich probe obtained from the pCMM86 clone (accession no., AF079321; unpublished observation). The values thus obtained were considered suitable for the carrying out of individualization studies. The absence of unaccountable band(s) in the offspring of a few families studied suggested that the loci recognized by this probe were not hypermutable in nature.

Hybridization carried out with the *Hae*III-digested genomic DNA of five non-Indian samples in a preliminary investigation also resulted in a ladder of bands in a size range similar to that observed in the Indian samples. DNA profile comparisons pointed towards the sharing of three isomorphic bands, of 29.0, 2.4, and 2.1 kb, in all Indian and non-Indian individuals, suggesting a probable positive selection pressure in retaining these DNA sequences in *Homo sapiens* in general, despite geographical isolation to a large extent. However, this observation needs further corroboration from more work in future, probably by comparing the sequences represented in these DNA bands. We have further observed that the 29.0-kb band observed in all human DNA samples is absent in both plant and animal species (data not shown). An overlap of DNA fragments of 10.5, 5.8, 4.0, 3.1, and 2.8 kb was observed between the non-Indian and a few of the Indian samples. This probably indicates the existence of remnants of common descendants, which needs to be confirmed on the basis of an identifiable form of more or less similar repeat patterns of the reported minisatellite, reflected in these RFLPs.

Table 1. Comparison of DNA fingerprint profiles in different individuals from four different population groups of India, with the minisatellite probe, pBA1.2

Region	*n*	Mean band no. (f) ± SE	Mean band sharing prob. (x) ± SE	Mean prob. of identity ($\overline{X}$) ± SE
Bengal	23	21 ± 0.747	0.18 ± 0.00406	2.29×10^{-16}
Punjab	23	22 ± 0.646	0.24 ± 0.00542	2.31×10^{-14}
South Ind.	21	21 ± 0.380	0.20 ± 0.00482	2.09×10^{-15}
UP	28	22 ± 0.814	0.20 ± 0.00429	4.19×10^{-16}

Ind, Indian; UP, Uttar Pradesh; prob., probability

Genetic similarity between Indian and non-Indian individuals and within Indian samples

A tree based on Nei's genetic distance method, using the mean frequency of the presence of bands for five population groups separately (Fig. 1), showed the genetic distance between South Indians and the remaining three Indian population groups to be high, as compared with the values obtained within and between the population groups of Bengal, Punjab, and Uttar Pradesh (Table 2). The type of placement of the South Indian individuals in the tree obtained through Nei's genetic distance could have been caused by the migration of earlier settlers in India, both southwards and eastwards, during prehistoric and historic times (Singh et al. 1994). The non-Indian samples also showed a high genetic distance when compared with the Indian population in general. However, minute dissection of the data for generating a tree at the individual level showed the presence of many subclusters comprising subgroups of various Indian populations. At this level, genetic similarity, based on Jaccard's similarity coefficient, was calculated for each pair of 100 individuals. A dendrogram, constructed based on the RFLP profile of individuals from the four population groups of India, led to the generation of six major clusters (Fig. 2), each comprising five or more individuals. Individuals B1, B2, B5, B3, B4, and B7 from Bengal formed a distinct cluster, as did individuals from UP (U13, U18, U19, U15, U16, and U17; U2, U5, U12, U9, and U10; and U8, U24, U25, U26, U28, and U27) and Punjab

Table 2. Genetic distance matrix between different populations

	Non-Indian	Punjabi	Bengali	South Indian	UPite
Non-Indian	0				
Punjabi	0.5171	0			
Bengali	0.3658	0.1571	0		
South Indian	0.2911	0.2626	0.1690	0	
UPite	0.3214	0.1934	0.1408	0.2000	0

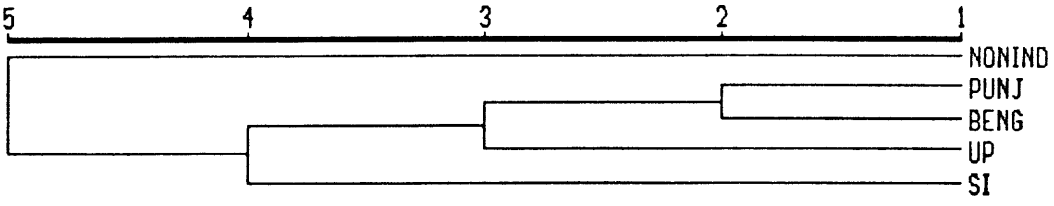
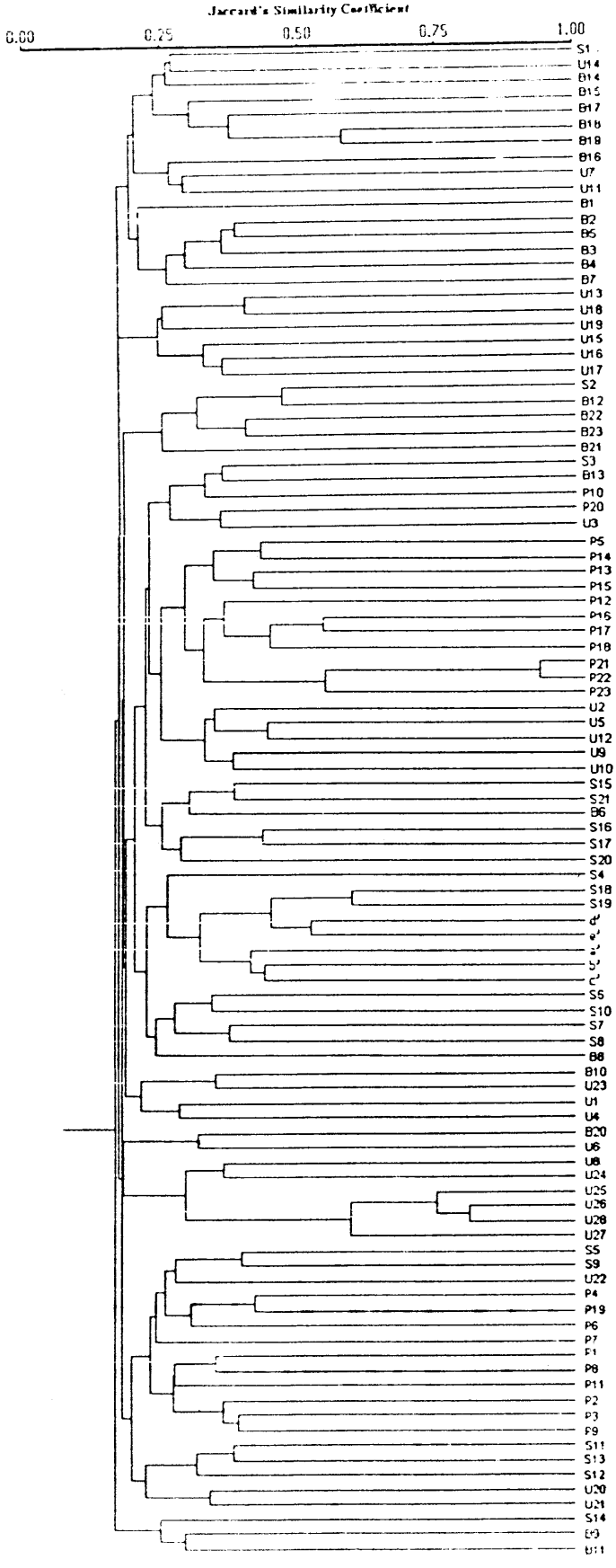


Fig. 1. Dendrogram, constructed by the neighbor-joining method, showing the relationships among the four Indian population groups and non-Indian (*NONIND*) samples. *PUNJ*, Punjab; *BENG*, Bengal; *UP*, Uttar Pradesh; *SI*, South India



(P5, P14, P13, P15, P12, P16, P17, P18, P21, P22, and P23 and P1, P8, P11, P2, P3, and P9) (Fig. 2). The formation of clusters indicated the close proximity of these individuals belonging to a population group. All the non-Indian samples formed a distinct separate cluster, represented by d', e', a', b', and c'. The b' and c' individuals (African samples) showed more genetic similarity, with a higher Jaccard's similarity coefficient, with the only Irish sample studied, whereas d' and e' (Korean samples) formed a cluster together, also showing proximity with few South Indian samples S18, S19 (Fig. 2). The high Jaccard's coefficient value obtained in the one Irish and two African samples, as compared with other individuals from different subgroups of India (Fig. 2), appears to contradict the nature of divergence seen between non-Indian — South Indian and other Indian population groups (Fig. 1). However, it would be improper to assess the past evolutionary relationships, because the parentage of the only randomly selected Irish sample could not be established. Moreover, before any conclusion is drawn the inclusion of a greater number of samples from the specified regions for the study is required, and these are not easily available at the moment. Nevertheless, the proximity between some South Indians and Africans and Koreans for the minisatellite sequence studied is suggestive that the South Indian population in India is probably the original Indian population.

A further advantage in the individual analysis in comparison with the population level analysis was observed in the resolution of more information about the genetic proximity within and between population groups when a probable admixture was reflected in our findings that some individuals belonging to one population group were dispersed or embedded within a cluster generated by the individuals of another population group. This is exemplified by the clustering seen with individuals P5, P14, P13, P15, P12, P16, P17, P18, P21, P22, and P23 of Punjab with individuals U2, U5, U12, U9, and U10 from UP. South Indians formed a separate cluster of individuals, S15 to S19, and S6 to S8, with only a very few South Indian individuals showing clustering with individuals from the rest of the three population groups (e.g., individuals S11, S13, and S12 of South India with individuals U20 and U21 from UP and individuals S5 and S9 with U22, which, in turn, show proximity with the Punjabi samples, P4, P19, P6, P7, P1, P8, P11, P2, P3, and P9). The sub-tree containing the South Indian subgroups S4, S18, and S19 and S6, S10, S7, and S8 also contained all five members of the out-group clustered within a subgroup, thus suggesting proximity of the South Indian with the non-Indian samples. While there are, indeed, small clusters, they are interleaved, and the subgroups for each of the populations are intermingled with the subgroups for the other populations. A separate analysis was carried out to check

Fig. 2. Unweighted pair group method arithmetic average (UPGMA)-based dendrogram of the normal individuals studied on the basis of Jaccard's similarity coefficient values. The *U* Indicates individuals from UP; *S* indicates individuals from South India; *B* indicates individuals from Bengal; and *P* indicates individuals from Punjab; *a' to e'* are non-Indian samples

the consistency of the proximity between different individuals forming a cluster and the consistency of the proximity between those individuals who fall in the vicinity of two clusters. The dendrograms thus obtained did not change the relationship between the individuals from all the populations studied.

Future studies involving a greater number of samples and markers will provide more insight into the nature and origin of population diversity into India.

Acknowledgments We thank Dr. K.V. Bhat, NBPGR, New Delhi for rendering his help in the construction of the dendrograms. The author A.S. also acknowledges University Grants Commission for providing fellowship in the form of Senior Research Fellow.

References

Bhat KV, Bhat SR, Chandel KPS, Lakhanpaul S, Ali S (1995) DNA fingerprinting of *Musa* cultivars with oligodeoxyribonucleotide probes specific for simple repeat motifs. Genetic Analysis: Biomolecular Engineering 12:45–51

Budowle B, Moretti TR (1999) DNA profiling and DNA fingerprinting, 1st edn. Birkhauser Verlag, Basel Switzerland, pp 101–116

Epplen JT (1988) On simple repeated GATCA sequences in animal genomes: a critical reappraisal. J Hered 79:409–417

Gill P, Werrett DJ (1987) Exclusion of a man charged with murder by DNA fingerprinting. Forensic Sci Int 35:145–148

Helminen P, Ehnholm C, Lokki ML, Jeffrey AJ, Peltonen L (1988) Application of DNA "fingerprints" to paternity determinations. Lancet I:574–576

Ito H, Yasuda N, Matsumoto H (1985) The probability of parentage exclusion based on restriction fragment length polymorphism. Jpn J Hum Genet 30:261–269

Jeffreys AJ, Wilson V, Thein SL (1985) Individual specific "fingerprints" of human DNA. Nature 316:76–79

Jeffreys AJ, Wilson V, Thein SL, Weatherall DJ, Ponder BAJ (1986) DNA "fingerprints" and segregation analysis of multiple markers in human pedigrees. Am J Hum Genet 39:11–24

Kasai K, Nakamura Y, White R (1990) Amplification of a VNTR locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. J Forensic Sci 35:1196–1200

Katsuyama Y, Inoko H, Imanishi T, Mizuki N, Gojobori T, Ota M (1998) Genetic relationships among Japanese, Northern Han, Uygur, Kazakh, Greek, Saudi Arabian, and Italian populations based on allelic frequencies at four VNTR (D1S80, D4S43, COL2A1, D17S5) and one STR (ACTBP2) loci. Hum Hered 48:126–137

Kunkel LM, Smith KD, Boyer SH, Bargaonkar DS (1977) Analysis of human Y chromosome specific reiterated DNA in chromosome variants. Proc Natl Acad Sci USA 74:1245–1249

Mountain JL, Cavalli-Sforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61:705–718

Nakamura Y, Martin C, Myers R, Ballard L, Leppert M, O'Connell P, Lathrop GM, Lalouel JM, White R (1988) Isolation and mapping of a polymorphic DNA sequence (pCMM86) on chromosome 17q. Nucleic Acids Res 11:5223

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

Rohlf FJ, Slice DE (1992) NTSYS-pc Numerical taxonomy and multivariate analysis system: Version 1.70 Applied Biostatistics, New York

Sachdeva G, Kaur G, Bamezai R (1995) Noise-free chemiluminescent detection of human T-cell receptor and interleukin-2 receptor genes after optimization of digoxigenin labelled probe concentration. Ind J Exp Biol 33:173–176

Sachdeva G, Kaur G, Bhutani LK, Bamezai R (1997) Genetic variations at the T-cell receptor gamma locus in circulating peripheral blood mononuclear cells of clinically categorized leprosy patients. Hum Genet 100:30–34

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sanger F (1981) Determination of nucleotide sequences in DNA. Science 214:1205–1210

Singh KS, Bhalla V, Kaul V (1994) The biological variations in Indian populations. People of India; national series volume X. Indian Oxford University Press, Oxford, pp 1–8

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517